Ann Oper Res
https://doi.org/10.1007/s10479-018-2875-2

S.I.: COMPUTATIONAL BIOMEDICINE

Nonparametric additive beta regression for fractional response with application to body fat data

Kuangnan Fang[1,3] · **Xinyan Fan**[1] · **Wei Lan**[2] · **Bingquan Wang**[1]

Abstract Fractional data that are restricted in the standard unit interval (0, 1) with a highly skewed distribution are commonly encountered. Such data arise in various areas, such as economics, finance, and medicine, among others. One natural idea to model such data is to use the beta family due to its flexibility to accommodate various density shapes. In this paper, we propose a nonparametric additive beta regression model along with a variable selection procedure, where the mean response is related to covariates through the combination of unknown functions of covariates, which can be approximated on a B-spline basis. By using this approximation method, we transform the problem of variable selection into the problem of selecting the groups of coefficients in the expansion. Based on the penalized likelihood method for group variable selection, we successfully select the significant covariates. Moreover, the estimation and selection consistencies and the properties of the penalized estimators are established. The simulation studies demonstrate that the performance of our proposed method is quite good. Finally, we apply the proposed method to body fat data, and we obtain several important findings with satisfactory selection and prediction performance.

Keywords Nonparametric additive beta regression · Fractional data · Variable selection · Group SCAD

1 Introduction

Fractional (or proportional) data are commonly encountered in many areas, such as medicine, economics, and finance. One typical example of such data is body fat data, in which the

✉ Kuangnan Fang
xmufkn@xmu.edu.cn

1 Department of statistics, School of Economics, Xiamen University, Xiamen, China

2 Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics, Xiamen, China

3 Data Mining Research Center, Xiamen University, Xiamen, China

Published online: 21 July 2018

percentage of body fat is within the unit interval (0, 1), and its distribution is highly skewed (see Fig. 1). Other examples of fractional data can be found in the forms of firm dividend yields, test pass rates, company market shares, television rates and so on. As we can observe, such fractional data generally cannot be negative, the distribution is asymmetric and highly skewed, and in particular, the errors are likely to be heteroscedastic and responses to covariates nonlinear. All of these attributes make properly modeling fractional data difficult and present us with several interesting statistical challenges.

Over the past two decades, various approaches have been proposed to model fractional data. As we know, the most straightforward approach to model such data is to use linear regression. However, one can easily argue that the linear regression model cannot always guarantee that the fitted or predicted values will fall into the unit interval (0, 1), which makes the results questionable and difficult to be interpreted. Moreover, linear regression may also incur the heteroscedasticity problem (Fang and Ma 2013; Ferrari and Cribari-Neto 2004). To overcome such issues, one possible approach is to first transform the response such that it can take values within $(-\infty, +\infty)$ and then apply a regression model to the transformed response. Unfortunately, such an approach still has shortcomings; one shortcoming is that the coefficients cannot easily be interpreted in terms of the original response, and another shortcoming is that the fractional response is generally asymmetric and highly skewed. Consequently, the inference based on the normality assumption might be misleading. Another approach that appears to be more appealing is to use beta regression, which was originally studied by Ferrari and Cribari-Neto (2004). In beta regression, the response variable is assumed to follow a beta distribution within the unit interval (0, 1). Note that the density of the beta distribution has different shapes depending on the values of the two parameters that index the distribution (Johnson et al. 1995; Ferrari and Cribari-Neto 2004); therefore, the beta distribution provides a very flexible approach for modeling fractional data. Moreover, beta regression has an additional benefit in that it has the same interpretation as logistic regression, and its coefficients can be estimated by the maximum likelihood estimation method.

Let $(y_i, X_i^\top)^\top$, $i = 1, \ldots, n$, be vectors that are independent and identically distributed as (y, X), where y is a response variable that is restricted to the unit interval (0, 1) and $X_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is the ith observation of the p covariates, which are assumed to be fixed and known. To obtain a regression structure for the mean of the response $y \in (0, 1)$, Ferrari and Cribari-Neto (2004) reparameterized the beta density as follows:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{1.1}$$

where $\mu \in (0, 1)$ is the mean of y, $\phi > 0$ is a precision parameter, and $\Gamma(\cdot)$ is the gamma function. The variance of y is $\mathrm{var}(y) = \mu(1-\mu)/(1+\phi)$. To work with this distribution in a regression model related to some covariates, Ferrari and Cribari-Neto (2004) proposed the following linear beta regression model:

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \tag{1.2}$$

where $g(\cdot)$ is a strictly monotonic and twice-differentiable link function that maps (0, 1) into $\mathbb{R}$; x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, is the ith observation of the jth covariate; and β_j is the coefficient of the jth covariate. The modeling and statistical inference procedures are similar to those for generalized linear models, except that the distribution of the response variable is not required to be a member of the exponential family. Within this framework, the mean response is related to a linear combination of covariates through a known link function $g(\cdot)$,

and this model setting has commonly been used in practice due to its simplicity. However, when the relationship between covariates and transformed response is not linear, the above model setting may become too restrictive and could also be misleading. As a direct solution to this issue, it may be better to assume that the mean response is related to covariates through the combination of some unknown functions of covariates, which means that the covariates are nonparametric additive.

Meanwhile, the measurements are obtained based on a large number of potential covariates to avoid missing any important link between a predictive factor and the response. It is well known that the classical maximum likelihood estimator method suffers from large variances in this case and may lead to overfitting. Moreover, it is possible that the number of variables may be larger than the sample size. To address such cases with high-dimensional data, an approach called *variable selection* has become increasingly popular and received considerable attention in diverse fields of scientific research. Based on the gamboostLSS algorithm proposed by Rigby and Stasinopoulos (2005), Schmid et al. (2013) proposed a boosted beta regression. However, no consistency theory was established to show that the variable selection procedure of their proposed boosted beta regression is consistent. Zhao et al. (2014) proposed variable selection methods for parametric linear beta regression models using the penalized likelihood method, which has been successfully developed over the past decade to address high dimensionality while simultaneously selecting important variables and estimating their effects in high-dimensional statistical inference (Fan and Lv 2010). To the best of our knowledge, the nonparametric beta regression has not been investigated, nor has variable selection for nonparametric beta regression, which is the main focus of this paper.

As we know, many penalized methods have been proposed to select the significant nonzero components for nonparametric regression. Among these studies, Zhang et al. (2004) and Lin and Zhang (2006) investigated the use of penalized methods in smoothing spline ANOVA with a fixed number of covariates. Xue (2009) proposed a penalized polynomial spline method for simultaneous variable selection and model estimation in additive models by using the SCAD penalty. Huang et al. (2010) proposed the adaptive group LASSO to select nonzero components in nonparametric additive models. Meier et al. (2008) extended the group LASSO to logistic regression. Although many methods have been proposed for variable selection in both parametric and nonparametric regression models, our literature review suggests that no research has been performed for the nonparametric beta regression model.

In this paper, we aim to extend the linear beta regression model to the nonparametric additive beta regression model, which enables us to model possible misspecification in a more flexible and robust manner. We further employ the SCAD penalty for group variable selection based on the B-spline approximations to the nonlinear functions. By using the B-spline approximation method, each nonlinear function can be represented by a linear combination of spline basis functions. Consequently, the problem of nonlinear function selection becomes the problem of selecting the groups of coefficients in the linear combinations. To this end, we propose a penalized group variable selection method that simultaneously selects the significant functions and estimates the group of coefficients. In addition, the estimation and selection consistencies of the proposed penalized estimators are established. The numerical study demonstrates that our proposed method could outperform other competing alternative methods.

The remainder of this paper is organized as follows. Section 2 describes the nonparametric beta regression and its variable selection procedure. The asymptotic properties of the proposed method are presented in Sect. 3. Section 4 presents the results of simulation studies to evaluate the finite sample performance of the proposed method. An illustrative application to body

fat data is provided in Sect. 5. Section 6 includes concluding remarks. Proofs of the results presented in Sect. 3 are provided in the "Appendix".

2 Models and methodology

In this section, we propose the following nonparametric additive beta regression:

$$g(\mu_i) = \sum_{j=1}^{p} f_j(x_{ij}), \tag{2.1}$$

where $\mu_i = E(y_i)$, $i = 1, \ldots, n$, is the mean of the response. $g(\cdot)$ is a strictly monotonic and twice- differentiable link function that maps $(0, 1)$ into $\mathbb{R}$, and we use the logit link function $g(\mu) = \log\{\mu/(1-\mu)\}$ in this study. Therefore, the results of a beta regression have essentially the same interpretation as those of a logistic regression. Note that other link functions may be used in place of the logit function. f_js are unknown smooth functions to be estimated, and suppose that some of them are zero. Note that the variance of y_i is a function of μ_i and consequently of the covariate values. Hence, nonconstant response variances are naturally accommodated in the model.

In the spirit of Huang et al. (2010), we use B-splines to approximate each of the unknown functions f_js. Suppose that each x_j takes values in a finite interval $[a, b]$ for $a < b$. To ensure unique identification of the f_js, we assume that $Ef_j(x_{ij}) = 0, 1 \le j \le p$. Moreover, let $a = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = b$ be a partition of $[a, b]$ into K subintervals with the kth interval $I_{kt} = [\xi_t, \xi_{t+1})$ for $k = 0, \ldots, K-1$, and $I_{KK} = [\xi_K, \xi_{K+1}]$, where $K \equiv K_n = n^\eta$ with $0 < \eta < 0.5$ is a positive integer such that $\max_{1\le k\le K+1} |\xi_k - \xi_{k-1}|$. Let S_n be the space of polynomial splines of degree $h \ge 1$ consisting of functions s satisfying (i) the restriction of s to I_{kt} is a polynomial of degree h for $1 \le k \le K$; (ii) for $h \ge 2$ and $0 \le h^* \le h-2$, s is h^* times continuously differentiable on $[a, b]$. According to the results in Schumaker (1981) and Huang et al. (2010), for any $f_{nj} \in S_n$, there exists a normalized B-spline basis $\{\Phi_k, 1 \le k \le m_n\}$ in S_n with $m_n \equiv K_n + h$ such that

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\Phi_k(x), \quad 1 \le j \le p. \tag{2.2}$$

Under suitable smoothness assumptions, the f_js can be well approximated by functions in S_n. For example, suppose there are two knots, $a = \xi_0 < \xi_1 < \xi_2 < \xi_3 = b$, and that the degree of polynomial h is 3; then, the specific representations of B-spline functions are $\Phi_1(x) = 1, \Phi_2(x) = x, \Phi_3(x) = x^2, \Phi_4(x) = x^3, \Phi_5(x) = (x-\xi_1)_+^3, \Phi_6(x) = (x-\xi_2)_+^3$, where t_+ denotes the positive part. More generally, an order h spline with knots $\xi_k, k = 1, \ldots, K$ is a piecewise-polynomial of order h and has continuous derivatives up to order $h-2$. The general form for the truncated-power basis set would be $\Phi_i(x) = x^{l-1}, l = 1, \ldots, h, \Phi_{h+j}(x) = (x-\xi_j)_+^{h-1}, j = 1, \ldots, K$, as noted by Hastie et al. (2009).

For each function f_j, $j = 1, \ldots, p$, we can approximate it using the functions in S_n. Therefore, we can have the following approximation

$$f_j(x) \approx f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\Phi_k(x)$$

with m_n coefficients $\beta_{jk}, k = 1, \ldots, m_n$. For simplicity, we assume that all f_js have the same value m_n. As we know, the convergence rate of $f_{nj}(x) - f_j(x)$ largely depends on

m_n, and we will discuss the selection of m_n later. The first basic function of each f_{nj} is $\Phi_1(x) = 1$, which is simply the intercept. Consequently, we do not need to include the first basic function in every expansion, and we only need to add the intercept in front of the combination of f_{nj}s and approximate every f_{nj} with the B-spline basis without the first basic function. For notational convenience, we still notate the index from 1 to k, that is, Φ_j is the original Φ_{j+1}. Subsequently, $g(u_i)$ defined in Eq. (2.1) can be approximated as

$$g(\mu_i) \approx \beta_0 + \sum_{j=1}^{p} \sum_{k=1}^{m_n} \beta_{jk} \Phi_k(x_{ij}). \tag{2.3}$$

Therefore, the problem of distinguishing the nonzero components from the zero components and estimating the nonzero components in Eq. (2.1) is equivalent to the problem of distinguishing the nonzero group coefficients from the zero group coefficients in Eq. (2.3). Consequently, if $f_j \equiv 0$, then we expect that all β_{jk}s are also equal to 0. Note that the coefficients of the zero function are all zero as a group, which motivates us to use the penalized likelihood estimator for group variable selection to select out the zero functions.

Let $\beta_j = (\beta_{j1}, \ldots, \beta_{jm_n})^\top$, and let $\beta = (\beta_1^\top, \ldots, \beta_p^\top)^\top$. Let $||b||_2 = (\sum_{k=1}^{m_n} |b_k|^2)^{1/2}$ denote the l_2 norm of any vector $b \in R^{m_n}$. The penalized likelihood estimator for group variable selection is obtained by maximizing the following objective function:

$$L(\beta, \phi) = l(\beta, \phi) - n \sum_{j=1}^{p} w_j(m_n) P_\lambda(||\beta_j||_2), \tag{2.4}$$

where $l(\beta, \phi) = \sum_{i=1}^{n} l_i(\mu_i, \phi)$ is the log-likelihood function; $l_i(\mu_i, \phi) = \ln \Gamma(\phi) - \ln \Gamma(\mu_i \phi) - \ln \Gamma((1-\mu_i)\phi) + (\mu_i \phi - 1) \ln y_i + \{(1-\mu_i)\phi - 1\} \ln(1-y_i)$; $P_\lambda(\cdot)$ is a penalty function, which conducts regularized estimation and, more importantly, selection of important covariates; and $w_j(\cdot)$ is used to rescale the penalty with respect to the dimensionality of the parameter vector β_j. Since m_n are the same for all covariates, we use $w_j(m_n) = 1$ in the remainder of this paper for simplicity. The penalized likelihood estimators are then defined as $(\tilde{\beta}_n, \tilde{\phi}) = \text{argmax}_{\beta,\phi} L(\beta, \phi)$.

There are a series of group penalties, such as group LASSO, group adaptive LASSO, group SCAD, and group MCP. In this paper, we use the group SCAD in the simulation studies of Sect. 5 and the real data analysis in Sect. 6. The penalty function of SCAD is

$$P_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta, & \theta \le \lambda \\ \frac{\gamma\lambda\theta - 0.5(\theta^2+\lambda^2)}{\gamma-1}, & \lambda < \theta \le \lambda\gamma \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \theta > \lambda\gamma \end{cases} \tag{2.5}$$

where $\lambda > 0$ and $\gamma > 2$ are tuning parameters. Note that group SCAD can be replaced by any other penalty, such as group MCP.

The penalized likelihood estimator $\tilde{\beta}_n = (\tilde{\beta}_{n1}^\top, \ldots, \tilde{\beta}_{np}^\top)^\top \in \mathbb{R}^p$ is obtained by maximizing the penalized objective function using some nonlinear optimization algorithm based on the Newton algorithm (see Nocedal and Wright 1999), such as the new unified algorithm proposed by Fan and Li (2001). The optimization algorithms require specification of the initial values to be used in the iterative scheme. Let $X = (X_1, \ldots, X_n)^T = (x_{ij})_{n \times p}$ and $Z = (g(y_1), \ldots, g(y_n))^\top$. According to Ferrari and Cribari-Neto (2004), the initial value for β can be obtained by the ordinary least squares estimator based on linear regression of the transformed response Z on $\Phi(X)$, where $\Phi(X) = (\mathbf{1}, \Phi_1(x_1), \ldots, \Phi_{m_n}(x_1), \ldots, \Phi_1(x_p), \ldots, \Phi_{m_n}(x_p))$ with $\Phi_k(x_j) =$

$(\Phi_k(x_{1,j}), \ldots, \Phi_k(x_{n,j})^\top$ and $\mathbf{1}$ is an n-dimensional vector with all elements equal to 1. That is, $\beta_{initial} = (\Phi(X)^\top \Phi(X))^{-1}\Phi(X)^\top Z$. We also need an initial guess for the value of ϕ. As shown, $\text{var}(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}$ implies that $\phi = \frac{\mu_i(1-\mu_i)}{\text{var}(y_i)} - 1$. Since $\text{var}(g(y_i)) \approx \text{var}\{g(\mu_i) + (y_i - \mu_i)g'(\mu_i)\} = \text{var}(y_i)\{g'(\mu_i)\}^2$, we take $\phi = \frac{1}{n}\sum_{i=1}^{n} \frac{\check{\mu}_i(1-\check{\mu}_i)}{\check{\sigma}_i^2} - 1$, where $\check{\mu}_i = g^{-1}(\Phi(X)_i\beta_{initial}))$ and $\check{\sigma}_i^2 = \check{e}^\top\check{e}/[(n-(m_n p+1))\{g'(\check{\mu}_i)\}^2]$ with $\check{e} = Z - X\beta_{initial}$, where g' is the first derivative of g and $\Phi(X)_i$ is the ith row of $\Phi(X)$.

Overall, the corresponding computational algorithm is as follows.

Algorithm 1:

1. Calculate $\Phi(X)$ and convert $y_1, \ldots, y_n$ to Z.
2. Determine the initial values of β and ϕ as $\beta_{initial}$ and $\phi_{initial}$ as follows.
   (a) $\beta_{initial} = (\Phi(X)^\top\Phi(X))^{-1}\Phi(X)^\top Z$
   (b) Calculate $\check{\mu}_i = g^{-1}(\Phi(X)_i\beta_{initial}))$, $\check{e} = Z - X\beta_{initial}$, $\check{\sigma}_i^2 = \check{e}^\top\check{e}/[(n-(m_n p+1))\{g'(\check{\mu}_i)\}^2]$
   (c) $\phi = \frac{1}{n}\sum_{i=1}^{n} \frac{\check{\mu}_i(1-\check{\mu}_i)}{\check{\sigma}_i^2} - 1$
3. Maximize the objective function with the Newton algorithm.

Note that there are two parameters in Eq. (2.4), in which the tuning parameter λ can be selected using fivefold cross-validation, whereas for the value of γ in SCAD, Fan and Li (2001) suggested examining a small number of values. In our numerical study, we observed that the results are not very sensitive to $\gamma = 3.7$; therefore, we set $\gamma = 3.7$.

3 Asymptotic properties

We present the theoretical results of the proposed method in this section. Let k be a nonnegative integer, and let $\mathcal{F}$ be the class of functions f on $[a, b]$, whose kth derivative $f^{(k)}$ exists and satisfies the following Lipschitz condition of order α:

$$|f^{(k)}(s) - f^{(k)}(t)| \le C|s-t|^\alpha \text{ for } s, t \in [a, b]$$

where $\alpha \in (0, 1]$ satisfy $d = \alpha + k > 0.5$. Without loss of generality, we assume that the first q components of Eq. (2.1) are nonzero, i.e., $f_j(x) \neq 0$ for $1 \le j \le q$, and $f_j(x) = 0$ otherwise. Let $\mathcal{F}_F = \{1, \ldots, p\}$ and $\mathcal{F}_T = \{1, \ldots, q\}$. Define $||f||_2 = [\int_a^b f^2(x)dx]^{1/2}$ for any function f whenever the integral exists. To proceed, we first set the following technique conditions, which are helpful for deriving the theoretical results.

(C1) The number of nonzero components q is fixed, and there is a finite constant $c_f > 0$ such that $\min_{1\le j\le q}||f_j||_2 \ge c_f$.
(C2) We assume that the unknown functions $f_j(x) \in \mathcal{F}$ and $E\{f_j(X_j)\} = 0$ for any $j = 1, \ldots, q$.
(C3) The covariate vector X_j for any $j = 1, \ldots, p$ has a continuous density, and constants C_1 and C_2 exist such that the density function g_j of X_j satisfies $0 < C_1 \le \sup_{a\le x\le b} g_j(x) \le C_2 < \infty$ for every $1 \le j \le p$.
(C4) Define Σ as the asymptotic covariance of the maximum likelihood estimator β^*, i.e., $\Sigma = -\left[n^{-1}E\{l''(\beta^*)\}\right]^{-1}$, where $l''(\cdot)$ represents the corresponding second derivative

of $l(\cdot)$. We assume that finite constants $C_3 > 0$ and $C_4 > 0$ exist such that $0 < C_3 < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < C_4 < \infty$.

Conditions (C1)–(C3) are directly from Huang et al. (2010), condition (C4) ensures that the maximum likelihood estimator of β is well defined, and these conditions are all mild and sensible in practice. By Lemma 1 of Huang et al. (2010), under conditions (C2) and (C3), an f_n satisfying $||f_n - f||_2 = O_p(m_n^{-d} + m_n^{1/2} n^{-1/2})$ exists. Specifically, if we choose $m_n = O(n^{1/(2d+1)})$, then $||f_n - f||_2 = O_p(m_n^{-d}) = O_p(n^{-d/(2d+1)})$. For convenience, we use $m_n = O(n^{1/(2d+1)})$ in the subsequent analysis. Then, we can establish the following theorem.

Theorem 1 (Estimation consistency) *Define* $\tilde{\mathcal{F}}_T = \{j : ||\tilde{\beta}_{nj}||_2 \neq 0, 1 \leq j \leq p\}$, *and let* $|\mathcal{M}|$ *denote the cardinality of any set* $\mathcal{M} \subseteq \{1, \ldots, p\}$. *Under conditions (C1)–(C4), we can obtain the following:*

(i) *With probability converging to 1,* $|\tilde{\mathcal{F}}_T| \leq M_1 |\mathcal{F}_T| = M_1 q$ *for a finite constant* $M_1 > 1$.
(ii) *If* $\left[\max\{P'_\lambda(||\beta_j||_2)\}^2\right] m_n/n^2 \to 0$ *as* $n \to \infty$, *then* $P(\tilde{\mathcal{F}}_T \supset \mathcal{F}_T) \to_p 1$.
(iii) $\sum_{j=1}^{p} ||\tilde{\beta}_{nj} - \beta_j||_2^2 = O_p(m_n^{-2d+1}) + O_p(\frac{4 \max\{P'_\lambda(||\beta_j||_2)\}^2 m_n^2}{n^2})$.

The proof of Theorem 1 is provided in "Appendix A". In this theorem, we generalize the results of Huang et al. (2010), which were established for the nonparametric additive linear regression model, to the nonparametric additive beta regression model. Our theoretical results imply that, as long as $\max\{P'_\lambda(||\beta_j||_2)\}^2 m_n/n^2 \to 0$, the proposed selection procedure can select the nonzero functions with probability approaching 1, and the resulting estimators are all consistent. However, from Theorem 1, we still do not know whether this method can rule out all zero predictors. Therefore, further studies on the properties of selection consistency are needed, which motivates us to propose the following assumptions on the penalty function.

(C5) Assume that

$$\frac{\max\{P'_\lambda(||\beta_j||_2)\} m_n^{1/4}}{n r_n} = o(1) \text{ and } \frac{n}{\max\{P'_\lambda(||\beta_j||_2)\} m_n^{(2d+1)/2}} = o(1),$$

where $r_n \max_{j \in \mathcal{F}_F/\mathcal{F}_T} ||\tilde{\beta}_{nj}||_2 = O_p(1)$. Define $\beta_n = {}_0\beta$ if $\text{sign}_0||\tilde{\beta}_{nj}|| = \text{sign}_0||\beta_j||$ for any $1 \leq j \leq p$, where $\text{sign}_0(|x|) = 1$ if $|x| > 0$ and $= 0$ if $|x| = 0$. Under the above conditions, we can derive the following theorem.

Theorem 2 (Selection consistency) *Under conditions (C1)–(C5), we have:*

(i) $P(\tilde{\beta}_n = {}_0\beta) \to 1$.

The proof is given in "Appendix B". Theorem 2 implies that the signs of the estimated coefficients are the same as the real ones as a group with probability converging to 1.

Based on the above results in Theorems 1 and 2, we can subsequently prove that the nonzero functions involved in Eq. (2.1) can be consistently selected with probability approaching 1. The results are summarized in the following proposition. Since the proofs are similar to that of Theorem 4 in Huang et al. (2010), we refer the reader to the work of Huang et al. (2010) for more details and omit the corresponding proof.

Proposition 1 *Define* $\hat{f}_j(x) = \sum_{k=1}^{m_n} \tilde{\beta}_{jk} \Phi_k(x)$. *Under conditions (C1)–(C5), we have*

(i) $P(||\hat{f}_j||_2 > 0, j \in \mathcal{F}_T$ *and* $||\hat{f}_j||_2 = 0, j \in \mathcal{F}_F/\mathcal{F}_T) \to_p 1$.
(ii) $\sum_{j=1}^{q} ||\hat{f}_j - f_j||_2^2 = O_p\big(m_n^{-2d}\big) + O_p\big(4m_n(\max\{P'_\lambda(||\beta_j||_2)\})^2/n^2\big)$.

4 Simulation studies

We conduct simulation studies to assess the performance of our proposed approach and compare it with alternative methods. The covariates are simulated from a multivariate normal distribution with two types of correlation structures of covariates: (a) no correlation (independence) and (b) autoregressive correlation, where the correlation coefficient between covariates j and k is $0.5^{|j-k|}$. The response data are generated from the model

$$f(y_i;\mu_i,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma(\phi)}y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1}$$
$$g(\mu_i)=\sum_{j=1}^{p}f_j(x_{ij}),\quad 0<y_i<1, i=1,\ldots,n,$$

and the following specific examples are considered.

Example 1 In this example, we set $p=12$ and consider $n=100,300,500$ and $\phi=10,60,120$. We set $f_1(x_1)=0.6\exp(-0.8x_1^2)$, $f_2(x_2)=0.3\ln(x_2^2+x_2+1)$, $f_3(x_3)=-0.5x_3^2$, $f_4(x_4)=0.3x_4$, $f_5(x_5)=0.5\sin(0.2\pi x_5)$, $f_6(x_6)=0.5x_6$, and $f_7(x_7)=\cdots=f_{12}(x_{12})=0$. Thus, in the true model, the number of nonzero functions and the number of zero functions are both 6.

Example 2 In this example, we set $p=30$ and consider $n=300,500$ and $\phi=10,60,120$. We set $f_1(x_1)=0.2\ln(x_1^2+1)$, $f_2(x_2)=0.3\exp(-x_2)$, $f_3(x_3)=0.2(1-x_3)x_3$, $f_4(x_4)=0.5\sin(0.2\pi x_4)$, $f_5(x_5)=0.3x_5$, and $f_6(x_6)=\cdots=f_{30}(x_{30})=0$. Thus, in the true model, the number of nonzero functions is 5, and the number of zero functions is 25.

As we can see, Example 2 has more covariates than Example 1. Moreover, the covariates in Example 2 are more sparse since its ratio of the number of significant variables to the number of total variables is $1/6$, whereas that in Example 1 is $1/2$. To better evaluate the performance of the proposed method (group SCAD), we also consider two direct competitors: parametric linear beta regression with SCAD (pSCAD) and boosted nonlinear beta regression (boost). Parametric linear beta regression does not take the grouping structure in the spline expansions of the components into account. We note that it may not be fair to compare the parametric linear beta regression with the proposed nonparametric additive beta regression and boosted nonlinear beta regression since the generating model is highly nonlinear. Our purpose is to illustrate that it is necessary to use nonlinear models when the underlying model is nonlinear in the case of variable selection with high-dimensional data and that model misspecification could lead to poor selection results. In group SCAD, we use the cubic B-spline with two knots for all functions $f_j(j=1,\ldots,p)$. The locations of the two knots are 1/3 quantile and 2/3 quantile. The boosted beta regression is implemented by the R package "gamboostLSS". To evaluate the estimation performance of the proposed model, the mean square error (MSE) is used, which is computed as $n^{-1}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$. Moreover, we measure the performance of the selection by the true positive number of selected variables (TP) and the false positive number of selected variables (FP), which are generally used for measuring the performance of variable selection (Breheny and Huang 2015; Wu et al. 2014). All simulation results are obtained through 200 replications. The simulation results of Examples 1 and 2 are summarized in Tables 1 and 2, respectively.

Table 1 shows that the proposed nonparametric additive beta regression with group SCAD has a comparable MSE with boosted beta regression, and both of them have a smaller MSE

Table 1 The results of Example 1

ϕ	n	Group SCAD			pSCAD			Boost		
		MSE	TP	FP	MSE	TP	FP	MSE	TP	FP
Independent										
10	100	0.0299	5.6	1.5	0.0447	4.0	1.8	0.0256	6.0	4.0
		(0.0052)	(0.5)	(1.3)	(0.0060)	(0.9)	(1.5)	(0.0033)	(0.0)	(1.1)
	300	0.0225	6.0	1.3	0.0413	4.9	1.6	0.0218	6.0	3.5
		(0.0022)	(0.0)	(0.9)	(0.0031)	(0.8)	(1.2)	(0.0019)	(0.0)	(0.8)
	500	0.0209	6.0	1.1	0.0398	5.0	1.8	0.0207	6.0	2.7
		(0.0016)	(0.0)	(1.1)	(0.0023)	(0.7)	(1.5)	(0.0014)	(0.0)	(1.0)
60	100	0.0064	6.0	1.2	0.0284	4.6	2.7	0.0055	6.0	4.4
		(0.0013)	(0.2)	(1.2)	(0.0053)	(0.6)	(1.3)	(0.0010)	(0.0)	(1.4)
	300	0.0042	6.0	1.2	0.0243	5.2	2.3	0.0041	6.0	4.8
		(0.0005)	(0.0)	(1.1)	(0.0028)	(0.8)	(1.5)	(0.0005)	(0.0)	(1.2)
	500	0.0039	6.0	0.7	0.0244	4.9	1.5	0.0039	6.0	5.3
		(0.0003)	(0.0)	(0.6)	(0.0021)	(0.7)	(0.9)	(0.0002)	(0.0)	(0.8)
120	100	0.0040	6.0	1.3	0.0234	4.8	2.6	0.0033	6.0	4.4
		(0.0014)	(0.0)	(1.2)	(0.0056)	(0.9)	(1.1)	(0.0009)	(0.0)	(1.3)
	300	0.0021	6.0	0.8	0.0221	4.9	2.2	0.0021	6.0	5.2
		(0.0002)	(0.0)	(0.8)	(0.0023)	(0.6)	(1.4)	(0.0002)	(0.0)	(1.0)
	500	0.0021	6.0	0.8	0.0223	4.9	2.0	0.0021	6.0	5.7
		(0.0002)	(0.0)	(0.8)	(0.0022)	(0.6)	(1.2)	(0.0002)	(0.0)	(0.5)
Autoregressive correlation										
10	100	0.0287	5.7	2.1	0.0408	4.1	0.9	0.0260	5.9	3.6
		(0.0054)	(0.7)	(1.8)	(0.0046)	(0.7)	(0.9)	(0.0043)	(0.4)	(1.5)
	300	0.0221	6.0	1.6	0.0372	4.4	1.0	0.0211	6.0	2.9
		(0.0024)	(0.2)	(0.9)	(0.0038)	(0.7)	(0.9)	(0.0016)	(0.0)	(1.1)
	500	0.0203	6.0	1.6	0.0369	4.9	1.0	0.0202	6.0	2.4
		(0.0014)	(0.0)	(1.5)	(0.0026)	(0.9)	(0.6)	(0.0012)	(0.0)	(0.9)
60	100	0.0062	6.0	1.6	0.0216	4.6	2.1	0.0051	6.0	4.7
		(0.0012)	(0.0)	(1.1)	(0.0029)	(0.8)	(1.3)	(0.0010)	(0.0)	(1.2)
	300	0.0041	6.0	1.1	0.0211	4.8	1.7	0.0040	6.0	5.0
		(0.0003)	(0.0)	(1.4)	(0.0021)	(0.7)	(1.2)	(0.0003)	(0.0)	(1.2)
	500	0.0039	6.0	1.4	0.0211	4.9	1.8	0.0038	6.0	5.2
		(0.0004)	(0.0)	(1.1)	(0.0018)	(0.8)	(1.0)	(0.0003)	(0.0)	(1.0)
120	100	0.0032	6.0	1.7	0.0204	4.4	1.9	0.0027	6.0	5.1
		(0.0007)	(0.0)	(1.7)	(0.0053)	(1.0)	(1.3)	(0.0005)	(0.0)	(0.8)
	300	0.0021	6.0	1.3	0.0205	5.0	2.3	0.0022	6.0	5.0
		(0.0002)	(0.0)	(1.2)	(0.0025)	(0.8)	(1.2)	(0.0002)	(0.0)	(1.4)
	500	0.0020	6.0	1.1	0.0198	5.1	1.6	0.0020	6.0	5.5
		(0.0002)	(0.0)	(1.1)	(0.0021)	(0.6)	(1.0)	(0.0002)	(0.0)	(0.6)

Table 2 The results of Example 2

ϕ	n	Group SCAD			pSCAD			boost		
		MSE	TP	FP	MSE	TP	FP	MSE	TP	FP
Independent										
10	300	0.0244	4.6	2.1	0.0277	4.3	6.2	0.0231	4.9	7.9
		(0.0022)	(0.5)	(2.2)	(0.0015)	(0.5)	(2.8)	(0.0013)	(0.3)	(5.3)
	500	0.0223	4.8	1.2	0.0272	4.5	7	0.0217	5.0	8.7
		(0.0021)	(0.4)	(1.0)	(0.0018)	(0.5)	(2.4)	(0.0014)	(0.0)	(3.6)
60	300	0.0047	4.9	0.9	0.0104	4.4	7.9	0.0047	5.0	10.3
		(0.0004)	(0.3)	(1.4)	(0.0007)	(0.5)	(2.8)	(0.0004)	(0.0)	(4.5)
	500	0.0040	5.0	1.3	0.0099	4.4	9.7	0.0043	5.0	8.7
		(0.0004)	(0.0)	(1.6)	(0.0009)	(0.5)	(2.3)	(0.0004)	(0.0)	(5.9)
120	300	0.0023	5.0	0.6	0.0085	4.4	9.1	0.0025	5.0	5.6
		(0.0002)	(0.0)	(0.8)	(0.0009)	(0.5)	(2.7)	(0.0003)	(0.0)	(4.1)
	500	0.0021	5.0	1.1	0.0084	4.1	10.2	0.0023	5.0	5.9
		(0.0002)	(0.0)	(1.0)	(0.0009)	(0.3)	(1.7)	(0.0002)	(0.0)	(4.2)
Autoregressive correlation										
10	300	0.0249	4.6	1.4	0.0280	4.4	7.5	0.0243	4.9	9.6
		(0.0021)	(0.5)	(1.3)	(0.0025)	(0.5)	(2.0)	(0.0020)	(0.3)	(3.4)
	500	0.0227	4.8	1.3	0.0272	4.4	7.7	0.0227	4.9	9.2
		(0.0015)	(0.4)	(1.7)	(0.0018)	(0.5)	(2.7)	(0.0014)	(0.3)	(4.0)
60	300	0.0044	5.0	0.5	0.0094	4.4	9.3	0.0048	5.0	11.4
		(0.0005)	(0.0)	(1.0)	(0.0010)	(0.5)	(3.5)	(0.0004)	(0.0)	(3.6)
	500	0.0041	5.0	0.3	0.0099	4.3	10	0.0048	5.0	8.7
		(0.0002)	(0.0)	(0.6)	(0.0007)	(0.5)	(3.1)	(0.0006)	(0.0)	(3.5)
120	300	0.0023	5.0	0.2	0.0076	4.6	8.8	0.0027	5.0	7.4
		(0.0004)	(0.0)	(0.4)	(0.0007)	(0.5)	(2.6)	(0.0004)	(0.0)	(4.2)
	500	0.0021	5.0	0.6	0.0074	4.3	9.0	0.0025	5.0	8.9
		(0.0002)	(0.0)	(0.8)	(0.0009)	(0.5)	(2.7)	(0.0003)	(0.0)	(5.4)

than that of parametric linear beta regression with SCAD. Group SCAD also has a comparable TP with boosted beta regression, but it has a much smaller FP than that of boost beta regression, which suggests that boosted beta regression tends to identify more covariates than the true number of significant covariates, particularly when the covariate sparsity ratio is high. Moreover, the MSE decreases as the sample size (n) or precision (ϕ) increases. The TP becomes closer to the true number of nonzero functions as the sample size (n) or precision (ϕ) increases, and the FP decreases as the sample size (n) or precision (ϕ) increases. The group SCAD and boosted beta regression perform better than the parametric linear model with SCAD in all of the experiments, which illustrates the importance of taking the nonlinear structure into account. Finally, the proposed method can work well with both independent and dependent covariates.

5 Application to body fat data

We use the body fat data set in Weisberg (1985) to illustrate the utility of the proposed method in nonparametric component selection, which has been analyzed by Hoeting et al. (1999), Leng et al. (2010) and Zhao et al. (2014). The response variable (y) is the percentage of body fat, which is proportional data restricted in the interval (0, 1). Figure 1 shows the histogram of the percentage of body fat. There are 13 covariates: x_1, age (years); x_2, weight (pounds); x_3, height (inches); x_4, neck circumference (cm); x_5, chest circumference (cm); x_6, abdomen circumference (cm); x_7, hip circumference (cm); x_8, thigh circumference (cm); x_9, knee circumference (cm); x_{10}, ankle circumference (cm); x_{11}, extended biceps circumference (cm); x_{12}, forearm circumference (cm); and x_{13}, wrist circumference (cm). The body fat data set has 253 observations. After excluding outliers, 248 observations are retained in the following analysis. We are interested in finding the covariates that are related to the percentage of body fat and measuring their relationship. To evaluate the performance of our proposed method with application to body fat data, we use the proposed nonparametric additive beta regression with group SCAD (group SCAD), boosted beta regression (boost) and parametric linear beta regression with SCAD (pSCAD) to model the relationship between the percentage of body fat and the 13 covariates.

To evaluate the performance of the methods, we use cross-validation to calculate the prediction mean square errors (PEs) and estimate the standard deviation of the PEs. We randomly split the data set into a training data set and a testing data set with size 2:1. We fit the models using the training data set and calculate the PEs using the testing data set. Based on 200 replications, the PEs are 0.0020 (group SCAD), 0.0026 (pSCAD), and 0.0020 (boost), and their corresponding standard deviations are 0.0004, 0.0004 and 0.0004, respectively. Nonparametric additive beta regression with group SCAD and boosted beta regression have better performance than linear regression with SCAD.

Table 3 lists the covariates selected by pSCAD, boost and group SCAD, indicated by check signs based on the entire data set. Nonparametric additive beta regression with group SCAD selects 7 covariates, linear regression with SCAD selects 5 covariates, and boosted beta regression selects 10 covariates. Five covariates (x_1, x_3, x_6, x_8, and x_{13}) are selected by all three methods. It is not surprising that the boosted beta regression identifies the largest number of covariates. The boosted beta regression selects far more covariates than the proposed method, but this does not lead to better prediction performance. Therefore, in this example, the proposed nonparametric additive beta regression with group SCAD provides a more appropriate list of covariates, which can serve better for further investigations. Figure 2 shows plots of the estimated additive components obtained by nonparametric additive beta regression with the group SCAD. All of them are nonlinear, confirming the need for taking nonlinearity into account for analyzing the body fat data.

6 Discussion

Fractional or proportional data are commonly encountered in many areas. When the response data are fractional or proportional data, linear regression is no longer appropriate. An appealing approach is beta regression proposed by Ferrari and Cribari-Neto (2004). In beta regression, the response variable is assumed to follow a beta distribution in the interval (0, 1). However, the classical beta regression assumes that the relationship between the proportional response and covariates is linear, which may make the model setting be too restrictive and mis-

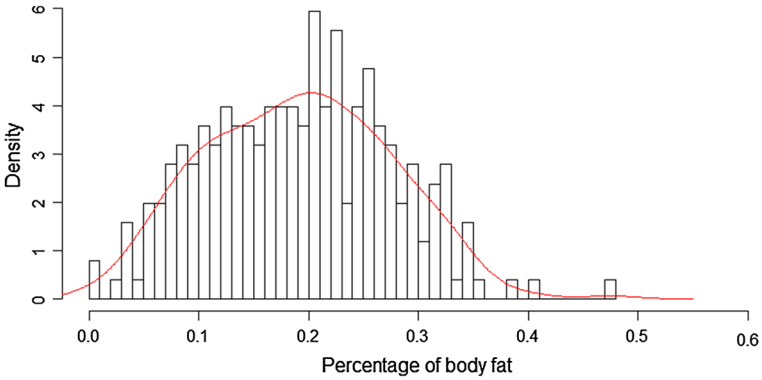

Fig. 1 Histogram of the percentage of body fat

Table 3 Analysis results for the real data

Covariates	pSCAD	Boost	Group SCAD
x_1	√	√	√
x_2	0	0	√
x_3	√	√	√
x_4	0	√	0
x_5	0	0	0
x_6	√	√	√
x_7	0	√	0
x_8	√	√	√
x_9	0	0	0
x_{10}	0	√	√
x_{11}	0	√	0
x_{12}	0	√	0
x_{13}	√	√	√

leading when the relationship between the covariates and transformed response is not linear. In addition, high-dimensional data have become increasingly popular and received considerable attention in diverse fields of scientific research. With these observations, developing a statistical method to model the nonlinear relationship between the proportional response and covariates and its variable selection procedure is of high interest.

In this paper, we extend the parametric linear beta regression to the nonparametric additive beta regression model together with a variable selection procedure. This implies that only a small subset of available predictor variables is included in the final model. Variable selection is of high practical interest in applications. Note that although the boosted beta regression proposed by Schmid et al. (2013) can conduct variable selection, its theoretical properties, particularly the variable selection consistency, are still unknown, which need to be further explored. Under some mild conditions, the penalized estimators are shown to possess estimation and selection consistencies. The results of simulations indicate that our proposed method can simultaneously select out zero components and estimate nonzero components efficiently. Applications using real data sets were presented and discussed. Note that in this work, although we assume that the precision parameter ϕ is fixed in the simulation and esti-

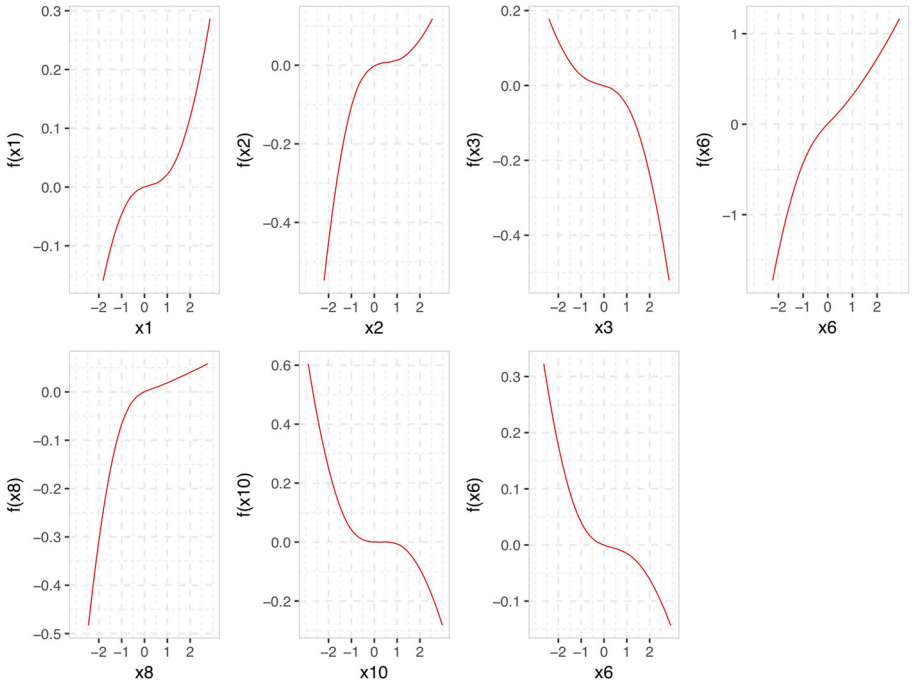

Fig. 2 The estimated additive components obtained with the group SCAD

mated by data in the real example, as with $\sigma^2 = Var(\epsilon)$ in traditional linear regression, it would be interesting and straightforward to further extend the proposed model by regressing the precision parameter ϕ on the covariates in our future work.

Acknowledgements This study has been supported by National Natural Science Foundation of China (71471152) and Fundamental Research Funds for the Central Universities of China (20720181003, 20720171095).

Appendix

Appendix A: The proof of Theorem 1

The proofs of parts (i) and (iii) are the same with the proofs of parts (i) and (iii) of Theorem 1 of Huang et al. (2010), to save space, we only present the proof of part (ii) here. By the definition of $\tilde{\beta}_n = (\tilde{\beta}_{n1}^\top, \ldots, \tilde{\beta}_{np}^\top)^\top$, we have

$$-n^{-1}l_n(\tilde{\beta}_n) + \sum_{j=1}^{p} P_\lambda(||\tilde{\beta}_{nj}||_2) \leq -n^{-1}l_n(\beta) + \sum_{j=1}^{p} P_\lambda(||\beta_j||_2).$$

Define $\mathcal{F}_2 = \{j : ||\beta_j||_2 \neq 0 \text{ or } j : ||\tilde{\beta}_{nj}||_2 \neq 0\}$. As a result,

$$-n^{-1}l_n(\tilde{\beta}_{\mathcal{F}_2}) + \sum_{j\in\mathcal{F}_2} P_\lambda(||\tilde{\beta}_{nj}||_2) \leq -n^{-1}l_n(\beta_{\mathcal{F}_2}) + \sum_{j\in\mathcal{F}_2} P_\lambda(||\beta_j||_2).$$

Let β^* be the maximum likelihood estimator of β. By the results of Horowitz and Mammen (2004), we have $\beta^* - \beta = O_p(m_n^{1/2}/n^{1/2} + m_n^{-2})$, namely, β^* is a consistent estimator of β. Subsequently, employing the Taylor expansion as in Wang and Leng (2007), we have

$$n^{-1}l_n(\beta) = n^{-1}l_n(\beta^*) + n^{-1}l_n'(\beta^*)(\beta - \beta^*) + \frac{1}{2}n^{-1}(\beta - \beta^*)^\top\{l_n''(\beta^*)\}(\beta - \beta^*)\big\{1 + o_p(1)\big\}.$$

Note that $l_n'(\beta^*) = 0$ and $n^{-1}l_n''(\beta^*) = -\Sigma^{-1}\big\{1 + o_p(1)\big\}$. Since $\beta^* - \beta = O_p(m_n^{1/2}/n^{1/2} + m_n^{-2})$, we have

$$\begin{aligned}&(\tilde{\beta}_{\mathcal{F}_2} - \beta^*_{\mathcal{F}_2})^\top \Sigma^{-1}_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta^*_{\mathcal{F}_2}) + \sum_{j\in\mathcal{F}_2} P_\lambda(||\tilde{\beta}_{nj}||_2)\\ &\le (\beta_{\mathcal{F}_2} - \beta^*_{\mathcal{F}_2})^\top \Sigma^{-1}_{\mathcal{F}_2}(\beta_{\mathcal{F}_2} - \beta^*_{\mathcal{F}_2}) + \sum_{j\in\mathcal{F}_2} P_\lambda(||\beta_j||_2).\end{aligned}$$

Let Z be the design matrix generated by the basis functions of X, simple calculation implies that $\Sigma^{-1} = \phi Z^\top W Z$, where $W = \mathrm{diag}\{w_1, \ldots w_n\}$ and $w_i = \phi\{\psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi)\}\{g'(\mu_i)\}^{-2}$ (Ferrari and Cribari-Neto 2004). Since ϕW is positive definite, thus we have $(\tilde{\beta} - \beta^*)^\top \phi Z^\top W Z(\tilde{\beta} - \beta^*) = (\tilde{\beta} - \beta^*)^\top[(\phi W)^{1/2}Z]^\top[(\phi W)^{1/2}Z](\tilde{\beta} - \beta^*)$. Define $(\phi W)^{1/2}Z$ as Z^* and $Z^*\beta^* = Y^*$, we then have $(\tilde{\beta} - \beta^*)^\top\Sigma^{-1}(\tilde{\beta} - \beta^*) = (Z^*\tilde{\beta} - Y^*)^\top(Z^*\tilde{\beta} - Y^*)$, which immediately leads to

$$\begin{aligned}&(Z^*_{\mathcal{F}_2}\tilde{\beta}_{\mathcal{F}_2} - Y^*_{\mathcal{F}_2})^\top(Z^*_{\mathcal{F}_2}\tilde{\beta}_{\mathcal{F}_2} - Y^*_{\mathcal{F}_2}) - (Z^*_{\mathcal{F}_2}\beta_{\mathcal{F}_2} - Y^*_{\mathcal{F}_2})^\top(Z^*_{\mathcal{F}_2}\beta_{\mathcal{F}_2} - Y^*_{\mathcal{F}_2})\\ &\le \sum_{j\in\mathcal{F}_2} P_\lambda(||\beta_j||_2) - \sum_{j\in\mathcal{F}_2} P_\lambda(||\tilde{\beta}_{nj}||_2).\end{aligned} \tag{A.1}$$

We then consider the two parts in (A.1) separately. Let $\eta_n = Y^*_{\mathcal{F}_2} - Z^*_{\mathcal{F}_2}\beta_{\mathcal{F}_2}$, and $v_n = Z^*_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2})$. Write $Y^*_{\mathcal{F}_2} - Z^*_{\mathcal{F}_2}\tilde{\beta}_{\mathcal{F}_2} = Y^*_{\mathcal{F}_2} - Z^*_{\mathcal{F}_2}\beta_{\mathcal{F}_2} - Z^*_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}) = \eta_n - Z^*_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2})$. We have $||Y^*_{\mathcal{F}_2} - Z^*_{\mathcal{F}_2}\tilde{\beta}_{\mathcal{F}_2}||_2^2 = ||Z^*_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2})||_2^2 - 2\eta_n^\top Z^*_{\mathcal{F}_2}(\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}) + \eta_n^\top\eta = ||v_n||_2^2 - 2\eta_n^\top v_n + \eta_n^\top\eta_n$. As a result, we can rewrite the left term of (A.1) as $||v_n||_2^2 - 2\eta_n^\top v_n$.

We next consider the right term of (A.1), employing the Taylor's expansion and Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|\sum_{j\in\mathcal{F}_2} P_\lambda(||\tilde{\beta}_{nj}||_2) - \sum_{j\in\mathcal{F}_2} P_\lambda(||\beta_j||_2)\right| = \left|\sum_{j\in\mathcal{F}_2}[P_\lambda(||\tilde{\beta}_{nj}||_2) - P_\lambda(||\beta_j||_2)]\right|\\ &= \left|\sum_{j\in\mathcal{F}_2} P'_\lambda\left(||\beta_j||_2\right)\left(||\tilde{\beta}_{nj}||_2 - ||\beta_j||_2\right)\right|\big\{1 + o_p(1)\big\}\\ &\le \max\left\{P'_\lambda\left(||\beta_j||_2\right)\right\}\left|\sum_{j\in\mathcal{F}_2}\left[\left(||\tilde{\beta}_{nj}||_2 - ||\beta_j||_2\right)\right]\right|\\ &\le \max\left\{P'_\lambda\left(||\beta_j||_2\right)\right\}\sqrt{|\mathcal{F}_2|}||\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}||_2,\end{aligned}$$

where $|\mathcal{F}_2|$ is the number of elements of $\mathcal{F}_2$.

Then, by (A.1), we have

$$\left(||v_n||_2^2 - 2\eta_n^\top v_n\right) \le \max\{P'_\lambda(||\beta_j||_2)\}\sqrt{|\mathcal{F}_2|}||\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}||_2.$$

Let η_n^* be the projection of η_n to the span of $Z^*_{\mathcal{F}_2}$, that is, $\eta_n^* = Z^*_{\mathcal{F}_2}(Z^{*\top}_{\mathcal{F}_2} Z^*_{\mathcal{F}_2})^{-1} Z^{*\top}_{\mathcal{F}_2} \eta_n$. By the Cauchy-Schwarz inequality,

$$2\left|\eta_n^\top v_n\right| \le 2||\eta_n^*||_2||v_n||_2 \le 2||\eta_n^*||_2^2 + \frac{1}{2}||v_n||_2^2.$$

As a result, we have

$$\begin{aligned}||v_n||_2^2 &\le 2\eta_n^\top v_n + \max\{P'_\lambda(||\beta_j||_2)\}\sqrt{|\mathcal{F}_2|}||\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}||_2 \\ &\le 2||\eta_n^*||_2^2 + \frac{1}{2}||v_n||_2^2 + \max\{P'_\lambda(||\beta_j||_2)\}\sqrt{|\mathcal{F}_2|}||\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}||_2.\end{aligned}$$

Subsequently,

$$||v_n||_2^2 \le 4||\eta_n^*||_2^2 + 2\max\{P'_\lambda(||\beta_j||_2)\}\sqrt{|\mathcal{F}_2|}||\tilde{\beta}_{\mathcal{F}_2} - \beta_{\mathcal{F}_2}||_2.$$

Let c_{n*} be the smallest eigenvalues of $Z^{*\top}_{\mathcal{F}_2} Z^*_{\mathcal{F}_2}/n$. By the results of Huang et al. (2010), we have $c_{n*} = O_p(m_n^{-1})$. Moreover, since $||v_n||_2^2 \ge nc_{n*}||\beta_{\mathcal{F}_2} - \tilde{\beta}_{\mathcal{F}_2}||_2^2$ and $2ab \le a^2 + b^2$, we have

$$\begin{aligned}nc_{n*}||\beta_{\mathcal{F}_2} - \tilde{\beta}_{\mathcal{F}_2}||_2^2 &\le 4||\eta_n^*||_2^2 \\ &+ \frac{\left(2\max\{P'_\lambda(||\beta_j||_2)\}\sqrt{|\mathcal{F}_2|}\right)^2}{2nc_{n*}} + \frac{1}{2}nc_{n*}||\beta_{\mathcal{F}_2} - \tilde{\beta}_{\mathcal{F}_2}||_2^2.\end{aligned}$$

It follows that

$$||\beta_{\mathcal{F}_2} - \tilde{\beta}_{\mathcal{F}_2}||_2^2 \le \frac{8||\eta_n^*||_2^2}{nc_{n*}} + \frac{4\left(\max\{P'_\lambda(||\beta_j||_2)\}\right)^2|\mathcal{F}_2|}{n^2c_{n*}^2}.$$

Note that $\beta^*_{nj} - \beta_j = O_p(m_n^{1/2}/n^{1/2} + m_n^{-2})$, which is the convergence rate of conventional nonparametric estimators (Horowitz and Mammen 2004). If we choose $m_n = O(n^{1/(2d+1)})$, then $\beta^*_{nj} - \beta_j = O_p(m_n^{-d})$, and by condition (C4), it follows that $||\eta_n^*||_2^2 = O_p(n|\mathcal{F}_2|m_n^{-2d})$. By the result of part (i), we have $|\tilde{\mathcal{F}}_T| \le M_1|\mathcal{F}_T| = M_1 q$, then $|\mathcal{F}_2| = |\tilde{\mathcal{F}}_T \bigcup \mathcal{F}_T| \le M_1|\mathcal{F}_T| + |\mathcal{F}_T|$, thus we have

$$\begin{aligned}||\beta_{\mathcal{F}_2} - \tilde{\beta}_{\mathcal{F}_2}||_2^2 &\le \frac{8||\eta_n^*||_2^2}{nc_{n*}} + \frac{4\left(\max\left\{P'_\lambda\left(||\beta_j||_2\right)\right\}\right)^2|\mathcal{F}_2|}{n^2c_{n*}^2} \\ &= O_p\left(\frac{|\mathcal{F}_2|m_n^{-2d}}{c_{n*}}\right) + O_p\left(\frac{4\left(\max\left\{P'_\lambda\left(||\beta_j||_2\right)\right\}\right)^2|\mathcal{F}_2|}{n^2c_{n*}^2}\right) \\ &= O_p\left(m_n^{-2d+1}\right) + O_p\left(\frac{4\left(\max\left\{P'_\lambda\left(||\beta_j||_2\right)\right\}\right)^2 m_n^2}{n^2}\right), \qquad \text{(A.2)}\end{aligned}$$

which completes the entire proof.

Appendix B: The proof of Theorem 2

By the proof of Theorem 1, the objective function can be approximated as

$$(Z^*\tilde{\beta}_n - Y^*)^\top(Z^*\tilde{\beta}_n - Y^*) + \sum_{j=1}^{p} P_\lambda(||\tilde{\beta}_{nj}||_2).$$

Moreover, employing Taylor's expansion, we know

$$\sum_{j=1}^{p} P_\lambda(||\tilde{\beta}_{nj}||_2) \approx \sum_{j=1}^{p} \left\{ P_\lambda(||\beta_j||_2) + P'_\lambda(||\beta_j||_2)[||\tilde{\beta}_{nj}||_2 - ||\beta_j||_2] \right\}$$
$$= \sum_{j=1}^{p} P'_\lambda(||\beta_j||_2)(||\tilde{\beta}_{nj}||_2) + \sum_{j=1}^{p} \left[P_\lambda(||\beta_j||_2) - P'_\lambda(||\beta_j||_2)||\beta_j||_2 \right].$$

Consequently, the objective function is equal to

$$(Z^*\tilde{\beta}_n - Y^*)^\top (Z^*\tilde{\beta}_n - Y^*) + \sum_{j=1}^{p} P'_\lambda(||\beta_j||_2)(||\tilde{\beta}_{nj}||_2).$$

For the convenience of notation, we define $P'_\lambda(||\tilde{\beta}_j||_2) = \lambda_j$ and $C_{\mathcal{F}_T} = n^{-1} Z^{*\top}_{\mathcal{F}_T} Z^*_{\mathcal{F}_T}$. Let ρ_{n1} and ρ_{n2} be the smallest and largest eigenvalues of $C_{\mathcal{F}_T}$, respectively.

By the KKT, a necessary and sufficient condition for $\tilde{\beta}_n$ is

$$\begin{cases} 2Z_j^{*\top}(Y^* - Z^*\tilde{\beta}_n) = \lambda_j \frac{\tilde{\beta}_{nj}}{||\tilde{\beta}_{nj}||}, & ||\beta_{nj}||_2 \neq 0,\ j \geq 1, \\ 2||Z_j^{*\top}(Y^* - Z^*\tilde{\beta}_n)||_2 \leq \lambda_j, & ||\tilde{\beta}_{nj}|| = 0,\ j \geq 1. \end{cases}$$

Let $u_n = (\frac{\lambda_j \tilde{\beta}_{nj}}{2||\tilde{\beta}_{nj}||},\ j \in \mathcal{F}_T)^\top$ and $\tilde{\beta}_{\mathcal{F}_T} = (Z^{*T}_{\mathcal{F}_T} Z^*_{\mathcal{F}_T})^{-1}(Z^{*T}_{\mathcal{F}_T} Y^* - u_n)$. If $\tilde{\beta}_{\mathcal{F}_T} =_0 \beta_{\mathcal{F}_T}$, then the two equations above hold for $\tilde{\beta}_n \equiv (\tilde{\beta}^\top_{\mathcal{F}_T}, 0^\top)^\top$. Thus, since $Z^*\tilde{\beta}_n = Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T}$ for this $\tilde{\beta}_n$ and $\{Z^*_j,\ j \in \mathcal{F}_T\}$ are linearly independent,

$$\tilde{\beta}_n =_0 \beta \text{ if } \begin{cases} \tilde{\beta}_{\mathcal{F}_T} =_0 \beta_{\mathcal{F}_T}, \\ ||Z_j^{*\top}(Y^* - Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T})||_2 \leq \lambda_j/2, \quad \forall j \notin \mathcal{F}_T. \end{cases}$$

This is true if

$$\tilde{\beta}_n =_0 \beta \text{ if } \begin{cases} ||\beta_j||_2 - ||\tilde{\beta}_{nj}||_2 \leq ||\beta_j||_2, & \forall j \in \mathcal{F}_T, \\ ||Z_j^{*\top}(Y^* - Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T})||_2 \leq \lambda_j/2, & \forall j \notin \mathcal{F}_T. \end{cases}$$

Therefore,

$$P\left(\tilde{\beta}_n \neq_0 \beta\right) \leq P\left(||\tilde{\beta}_{nj} - \beta_j||_2 \geq ||\beta_j||_2, \exists j \in \mathcal{F}_T\right)$$
$$+ P\left(||Z_j^{*\top}\left(Y^* - Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T}\right)||_2 > \lambda_j/2, \exists j \notin \mathcal{F}_T\right).$$

Let $\delta_n = Y^* - Z^*_{\mathcal{F}_T}\beta_{\mathcal{F}_T}$, and $H_n = I_n - Z^*_{\mathcal{F}_T}(Z^{*\top}_{\mathcal{F}_T} Z^*_{\mathcal{F}_T})^{-1} Z^{*\top}_{\mathcal{F}_T}$. By the definition of $\tilde{\beta}_{\mathcal{F}_T}$, $\tilde{\beta}_{\mathcal{F}_T} - \beta_{\mathcal{F}_T} = n^{-1} C^{-1}_{\mathcal{F}_T}(Z^{*\top}_{\mathcal{F}_T}\delta_n - u_n)$ and $Y^* - Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T} = H_n\delta_n + Z^*_{\mathcal{F}_T} C^{-1}_{\mathcal{F}_T} u_n/n$. Based on the above two equations, under condition (C5) Lemma 5 of Huang et al. (2010) shows that

$$P(||\tilde{\beta}_{nj} - \beta_j||_2 \geq ||\beta_j||_2, \exists j \in \mathcal{F}_T) \to 0$$

and Lemma 6 of Huang et al. (2010) shows that

$$P\left(||Z_j^{*\top}(Y^* - Z^*_{\mathcal{F}_T}\tilde{\beta}_{\mathcal{F}_T})||_2 > \lambda_j/2, \exists j \notin \mathcal{F}_T\right) \to 0.$$

These two equations lead to the Theorem 2.

References

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, *25*(2), 173–187.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101–148.

Fang, K., & Ma, S. (2013). Three-part model for fractional response variables with application to Chinese household health insurance coverage. *Journal of Applied Statistics*, *40*(5), 925–940.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistics Science*, *44*(4), 382–417.

Horowitz, J., & Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, *32*(6), 2412–2443.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, *38*(4), 2282–2313.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed.). New York: Wiley.

Leng, C., Tran, M., & Nott, D. (2010). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, *66*(2), 221–244.

Lin, Y., & Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, *34*(5), 2272–2297.

Meier, L., Van De Geer, S., & Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, *70*(1), 53–71.

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. New York: Springer.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Schmid, M., Wickler, F., Maloney, K. O., Mitchell, R., Fenske, N., & Mayr, A. (2013). Boosted beta regression. *Plos One*, *8*(4), e61623.

Schumaker, L. (1981). *Spline functions: Basic theory*. New York: Wiley.

Wang, H., & Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, *102*(479), 1039–1048.

Weisberg, S. (1985). *Applied linear regression*. New York: Wiley.

Wu, C., Cui, Y., & Ma, S. (2014). Integrative analysis of gene environment interactions under a multiresponse partially linear varying coefficient model. *Statistics in Medicine*, *33*(28), 4988–4998.

Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica*, *19*, 1281–1296.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., et al. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American statistical Association*, *99*(467), 659–672.

Zhao, W., Zhang, R., Lv, Y., & Liu, J. (2014). Variable selection for varying dispersion beta regression model. *Journal of Applied Statistics*, *41*(1), 95–108.