The Journal of Supercomputing (2020) 76:3623–3645
https://doi.org/10.1007/s11227-018-2597-x

Robust continuous piecewise linear regression model with multiple change points

Shurong Shi[1] · **Yi Li**[1] · **Chuang Wan**[2]

Published online: 7 September 2018

Abstract
This paper considers a robust piecewise linear regression model with an unknown number of change points. Our estimation framework mainly contains two steps: First, we combine the linearization technique with rank-based estimators to estimate the regression coefficients and the location of thresholds simultaneously, given a large number of change points. The associated inferences for all the parameters are easily derived. Second, we use the LARS algorithm via generalized BIC to refine the candidate threshold estimates and obtain the ultimate estimators. The rank-based regression guarantees that our estimators are less sensitive to outliers and heavy-tailed data, and therefore achieves robustness. Simulation studies and an empirical example on BMI and age relationship illustrate the proposed method.

Keywords Piecewise linear · Change points · Rank-based estimators

1 Introduction

Threshold regression models are commonly used for modeling certain kinds of nonlinear relationships between the response and explanatory variables by introducing one or more threshold parameters, also known as change points. Comparing with spline models, threshold regression models provide a simple but interpretable way to address nonlinear problem and give specific threshold estimates and inferences. In this paper, we mainly focus on a special case of threshold model, the segmented linear regression, also referred to as piecewise linear regression model. The distinct feature of this kind of regression model is that the threshold covariate has different regression coefficients in different regions but continuous at the change point.

✉ Yi Li
liyihnu@126.com

1 School of Finance and Statistics, Hunan University, Changsha 410082, China

2 Gregory and Paula Chow Center for Economics Research, Xiamen University, Xiamen 361005, China

This regression model has many applications. For instance, body mass index (BMI) increases stably with age when people are young. However, BMIs tend to stabilize after adulthood and then decrease when one becomes older. Similar situations are widely encountered in biostatistics, finance, economics, and so on. Many authors have employed segmented regression models to capture this threshold effect. Li and He [1] introduced a special segmented quantile regression model with a single change point, called bent line quantile regression. Lee et al. [2] considered a sup-likelihood-ratio-based method to test for threshold effects in the regression model. Zhang et al. [3] developed a sup-score-type statistic to test for change points due to a covariate threshold. Zhang et al. [4] proposed a composite change point estimation for bent line quantile regression and derived the estimators' consistency and asymptotic properties. Zhang and Li [5] proposed a continuous threshold expectile model, which is also a bent line model under expectile regression. Yan et al. [6], based on the linearization technique, developed a new estimating method for the bent line quantile regression model proposed by Li and He [1]. All these works employ standard segmented regressions, with only one change point in the responses.

The above techniques may work quite well when dealing with a single change point. In many applications, however, a single threshold parameter may not be capable of capturing multiple structural changes, which are quite common in many research fields. Motivated by this limitation, we propose a flexible segmented regression model with multiple change points, where their number needs to be estimated. The proposed model can not only capture a threshold effect but also accommodate situations with multiple change points. Besides, datasets in empirical application often have outliers or heavy tails, which influence the accuracy of parameter estimations and disturb the detection of change points. For instance, not a few young children have obesity, causing their BMIs to deviate far away from the normal level in that age. These participants can substantially influence the fitting of the model and reduce the credibility of the estimation results. In such circumstances, a robust estimation method is desirable. Rank-based regression is a mature tool to achieve robustness with no assumption on the distribution of the error term; see Abebe et al. [7], Hettmansperger and Mckean [8] for a comprehensive review.

In this paper, we develop a continuous rank-based segmented linear model with multiple thresholds. The main contribution of this article is to combine the multiple segmented model with rank-based regression. To estimate a continuous multiple threshold model, we determine the location of change points and regression parameters simultaneously by drawing on the merits of Muggeo's linear reparameterization technique [9]. We then use the LASSO method to identify the number of thresholds. Rank-based regression can ensure that our estimation result is robust and less influenced by outliers or heavy-tailed data, which can maintain high efficiency. The inference for all parameters is directly derived by using existing theories and can be readily implemented in mainstream statistical software.

The rest of presented paper is organized as follows: In Sect. 2, we present the continuous rank-based piecewise linear regression model and describe the estimating procedures. In Sect. 3, simulation experiments are conducted to assess the accuracy of the number of change points detected and investigate the finite sample perfor-

mance of the proposed method. In Sect. 4, we apply the proposed model and method to real-world regression results between BMI and age. Section 5 concludes the paper.

2 Methodology

2.1 Model

Assume that Y is the response variable, Z is a q-dimensional vector of covariates with constant slopes, and X is a univariate threshold scalar variable that shows a piecewise linear relationship with Zhang and Li [10] who presented a typical continuous robust bent line regression model with a single change point.

$$Y = \gamma^T Z + \alpha X + \beta(X - \lambda)_+ + e \tag{1}$$

where $(X-\lambda)_+ = (X-\lambda)I(X > \lambda)$, $I(\cdot)$ is the indicator function and β is commonly assumed to be a nonzero value for identifying threshold λ. Obviously, the robust bent line regression model is continuous on X at λ, but has different slopes on either segment of λ. To be specific, the slope relating Y to X is α when $X \leq \lambda$ and becomes $\alpha + \beta$ when $X \geq \lambda$. Threshold parameter λ is the so-called change point.

Although the model above works well for a single change point, it may not be applicable for multi-segment piecewise linear relationships. Therefore, we extend model 1 for multiple change points and propose a robust continuous piecewise linear regression (RCPLR) model.

$$Y = \gamma^T Z + \alpha X + \sum_{j=1}^{K} \beta_j (X - \lambda_j)_+ + e \tag{2}$$

Where e is an identical and independent random error term with no distributional assumption, K is the unknown number of change points, and $\lambda_1, \ldots, \lambda_K$ are the locations of the K-bounded change points. γ is a $q \times 1$ vector of the linear regression coefficients for Z. The scalar α is the slope of X before the appearance of a change point, and β_k is the difference in slopes between kth and $(k+1)$th change points; for example, $\alpha + \sum_{j=1}^{k} \beta_j$ is the slope of X when $\lambda_k \leq X < \lambda_{k+1}$.

It is typically assumed that $E(e) = 0$ and $Var(e) < +\infty$ in model 2, as in Muggeo and Adelfio [11] and Giada Adelfio [12], or $e \sim N(0, \sigma^2)$. However, when the sample is extremely heavy-tailed or includes many outliers that are far away from the mean level, parameter estimation is markedly influenced and lacks robustness under the above assumptions, which more or less complicates change point detection. In such circumstances, robust regression is desirable.

The basic statistical problems of interest with the RCPLR model are twofold: identification of the number of change points and the estimation and associated inference of the change point locations ($\lambda_j s$) and the regression coefficients (γ^T, α, $\beta_1, \ldots, \beta_K)^T$ in the robust regression framework. We will describe the process.

2.2 Rank-based estimator

Rank-based regression is a commonly used tool to achieve robustness. The rank-based estimator was first introduced by Jureckova [13] and Jaeckel [14]. Zhang and Li [10] have discussed this type of estimator in the context of a bent line model with a change point. Here, we alternatively apply it to a continuous piecewise linear model. Let the observed data $\left\{(y_i, x_i, z_i^T)^T, i = 1, \ldots, n\right\}$ be independent and identical samples from population (Y, X, Z), $e = (e_1, \ldots, e_n)$ be residuals. For ease of notation, let $\lambda = (\lambda_1, \ldots, \lambda_K)^T, \theta = (\gamma^T, \alpha, \beta_1, \ldots, \beta_K)^T$.

Recall that the ordinary least squares estimator is the minimizer of Euclidean distance in an objective function, $\|e\|_2^2 = \sum_{i=1}^n e_i^2$. However, in rank-based regression, the Euclidean distance is replaced by a different measure of distance based on Jaeckel's dispersion function; see Hettmansperger and Mckean [8] for details. To obtain the rank-based estimator for model 2, the objective function to be minimized is given by

$$D_n(\theta, \lambda) = \|e\|_\varphi \tag{3}$$

where $\|\cdot\|_\varphi$ is pseudo-norm defined as

$$\|e\|_\varphi = \sum_{i=1}^{n} \alpha(R(e_i))e_i \tag{4}$$

where $e_i = y_i - \gamma^T z_i - \alpha x_i - \sum_{j=1}^K \beta_j (x_i - \lambda_j)_+$ is the ith residual, $R(e_i)$ denotes the rank of e_i, $\alpha(i) = \varphi(i/(n+1))$. $\varphi(u)$ is a nondecreasing score function defined on the interval (0, 1) and standardized such that $\int \varphi(u)\mathrm{d}u = 0$ and $\in \int \varphi(u)^2 \mathrm{d}u = 1$.

As pointed out by Hettmansperger and Mckean [8], the rank-based fit and their statistical analysis can be optimized by a prudent choice of score function, as it determines Jaeckel's dispersion function. Thus, the robustness of the estimator depends on the choice of score function. Theoretically, the specific score function is specified if the distributional form of the error is known. In most statistical software, the default score function for robust regression is the Wilcoxon linear score function as it has been shown highly efficient and robust in symmetric and slightly heavy-tailed situations. Hence, we choose the Wilcoxon score function, $a(t) = \sqrt{12}(t - 0.5)$, as the rank score function here.

2.3 Iterative estimating procedures for RCPLR

Before identifying the number of change points, we first introduce the estimation method for the parameters in RCPLR, assuming that K is known or fixed. Here, we aim to estimate the locations of the change points λ and regression coefficients θ and make the associated statistical inferences. However, it should be noticed that the objective function 3 is neither differentiable concerning λ nor linear due to the existence of the indicator function $I(X > \lambda)$. To overcome this problem, we employ a linear reparameterization technique proposed by Muggeo [9].

Take the kth threshold estimation as an example. Specifically, we can approximate the non-differentiable term $\beta_k(X-\lambda_k)_+$ around the initial value $\lambda_k^{(0)}$ using first-order Taylor expansion wherein $\lambda_k^{(0)}$ should be close to the true value as much as possible:

$$(X-\lambda_k)_+ \approx (X-\lambda_k^{(0)})_+ + (-1)I(X>\lambda_k^{(0)})(\lambda_k-\lambda_k^{(0)}) \tag{5}$$

Then, the above linearization structure is applied to all change points and the model 2 can be approximated by

$$Y = \gamma^T Z + \alpha X + \sum_{j=1}^{K}\beta_j\tilde{U}_j + \sum_{j=1}^{K}\delta_j\tilde{V}_j + e, \tag{6}$$

where $\delta_j = \beta_j(\lambda_j-\lambda_j^{(0)})$, $\tilde{U}_j = (X-\lambda_j^{(0)})_+$ and $\tilde{V}_j = -I(X>\lambda_j^{(0)})$ are two new covariates with coefficients β_j and δ_j. In this way, model 5 can be viewed as a fully linear robust regression model for which the estimations and associated inferences are quite standard. Thus, the new objective function for estimating the regression coefficients of model 5 is as follows:

$$D_n(\theta,\delta) = \sum_{i=1}^{n}\sqrt{12}\left(\frac{R_i^{(0)}}{n+1}-0.5\right)e_i^{(0)} \tag{7}$$

where $e_i^{(0)} = y_i - \gamma^T z_i - \alpha x_i - \sum_{j=1}^{K}\beta_j\tilde{U}_{ij} - \sum_{j=1}^{K}\delta_j\tilde{V}_{ij}$, $R_i^{(0)}$ is the corresponding rank, $\delta = (\delta_1, \ldots, \delta_K)^T$. As the thresholds are contained in δ, that is, $\delta_j = \beta_j(\lambda_j - \lambda_j^{(0)})$ for jth threshold, a natural iterated estimate for λ_j may be updated by $\hat{\lambda}_j^{(1)} = \hat{\lambda}_j^{(0)} + \frac{\hat{\delta}_j}{\hat{\beta}_j}$, where Muggeo and Adelfio [11] called the $\hat{\delta}_j$ as "working" coefficient, since it measures the gap between the two fitted lines (before and after $\hat{\delta}_j$).

By virtue of what we have discussed previously, the main procedure for the estimation of parameters, including thresholds, can be summarized as:

1. Initialize parameters $(\hat{\theta}^{(0)}, \hat{\delta}^{(0)})$, in which $\hat{\delta}$ is set to be a small value vector;
2. At the tth step, fix vector $\hat{\lambda}^{(t)}$ and update parameters $(\hat{\theta}^{(t+1)}, \hat{\delta}^{(t+1)})$ by fitting a rank-based regression.

$$(\hat{\theta}^{(t+1)}, \hat{\delta}^{(t+1)}) = \arg\min_{\theta,\delta} i = \sum_{i=1}^{n}\sqrt{12}\left(\frac{R_i^{(t)}}{n+1}-0.5\right)e_i^{(t)} \tag{8}$$

$R_i^{(t)}$ denotes the rank of the ith residual $e_i^{(t)}$ defines as

$$e_i^{(t)} = y_i - z_i^T\hat{\gamma}^{(t)} - \hat{\alpha}^{(t)}x_i - \sum_{j=1}^{K}\hat{\beta}_j^{(t)}\tilde{U}_{ij}^{(t)} - \sum_{j=1}^{K}\hat{\delta}_j^{(t)}\tilde{V}_{ij}^{(t)} \tag{9}$$

where $\tilde{U}_{ij}^{(t)} = (x_i - \hat{\lambda}_j^{(t-1)})_+$, $\tilde{V}_{ij}^{(t)} = -I(x_i - \hat{\lambda}_j^{(t-1)})$.

3. Update the jth change point estimate $\hat{\lambda}_j^{(t+1)}$ by $\hat{\lambda}^{(t+1)}=\hat{\lambda}_j^{(t)}+\frac{\hat{\delta}_j^{(t+1)}}{\hat{\beta}_j^{(t+1)}}$. The other change points estimates are identified similarly.
4. Repeat steps (2)–(3) until a possible convergence criterion holds, for example, when the "working" coefficients vector $\hat{\delta}$ approximates to zero, that is, $\max_j \left|\hat{\delta}\right| < 10^{-4}$.

Remark 1 The above algorithm draws on the core idea of the linearization technique in Muggeo [9] to overcome the non-smoothness problem caused by change points through Taylor expansions. Thus, optimizing the objective function 3 is equivalent to iteratively minimizing the loss function of the standard rank-based linear regression model 5. Moreover, the algorithm has been shown computationally efficient for bent line models with a single change point under different regression structures such as rank-based regression [10], quantile regression [6], or expectile regression [4]. We will demonstrate the efficiency of the RCPLR model for multiple change points by simulations.

Remark 2 Denote $\lambda = (\lambda_1, \ldots, \lambda_K)^T$, $\beta = (\beta_1, \ldots, \beta_K)^T$. Let $(\hat{\theta}, \hat{\delta})$, and $\hat{\lambda}$ be the ultimate estimators of $(\hat{\theta}^{(t)}, \hat{\delta}^{(t)})$, and $\hat{\lambda}^{(t)}$, respectively. Clearly, the statistical inference for $(\hat{\theta}, \hat{\delta})$ is quite standard from the theory of rank-based linear regression. Hettmansperger and Mckean [8] derived the properties of linear rank-based estimators that have an asymptotically normal distribution and achieve $\sqrt{n}$-consistency. However, the estimates for change points take an indirect pattern and the asymptotic covariance matrix of λ can be obtained by using the delta method:

$$\mathrm{SE}(\hat{\lambda}) = \frac{\left[\mathrm{Var}(\hat{\delta}) + \mathrm{Var}(\hat{\delta})(\hat{\delta}/\hat{\beta})^2 - 2(\hat{\delta}/\hat{\beta})\mathrm{Cov}(\hat{\delta}/\hat{\beta})\right]^{1/2}}{\left|\hat{\beta}\right|} \tag{10}$$

In particular, since δ measures the difference between two fitted hyperplanes, $\hat{\delta}$ is expected to approach the zero vector when the algorithm converges. Then, 7 can be simplified as $\mathrm{SE}(\hat{\lambda}) = \mathrm{Var}(\hat{\delta})/|\hat{\beta}|$. Thus, the important $1-\alpha$ Wald-type confidence intervals are also given by

$$\left[\hat{\lambda} - z_{\alpha/2}\mathrm{SE}(\hat{\lambda}), \hat{\lambda} + z_{\alpha/2}\mathrm{SE}(\hat{\lambda})\right] \tag{11}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution.

Remark 3 From the viewpoint of technique, the linear rank-based estimators as well as other statistics such as standard errors, confidence intervals can be readily implemented by the R package Rfit. It is important to note that all the above estimation procedures are conditional on the fact that the number of change points is known or fixed. Regarding the selection procedure for K change points, we propose a new algorithm in the subsequent text, but the iterative estimation procedures described here form the basis of the next algorithm.

2.4 Estimating the number of change points

The aforementioned "segmented" algorithm allows us to get parameter estimates efficiently provided the number of change points is known. However, the true number of change points is usually unknown and needs to be estimated. To this end, we propose a new algorithm to solve this issue as follows.

The algorithm starts with K_0 (K_0 should be large enough) candidate change points. We do this mainly to avoid missing potential change points. The request for initial candidate threshold values should be uniformly distributed on the domain and usually set as the quantiles of threshold variable X. Then, the segmented model 5 is iteratively fitted.

In each iteration, the inadmissible change point as well as the corresponding covariates, $\tilde{U}_k$ and $\tilde{V}_k$ are discarded. We say the change point is inadmissible when

- the estimate $\hat{\lambda}_k$ does not belong to the allowed range. If $\hat{\lambda}_k$ is not a change point, $\hat{\beta}_k$ tends to be zero, causing the ratio $\hat{\lambda}_k/\hat{\beta}_k$ to go to infinity. Therefore, the updated estimate $\hat{\lambda}_k$ will fall outside the domain of the definition of X.
- the estimate $\hat{\lambda}_k$ is much close to another change point estimate $\hat{\lambda}_{k'}$, that is $\hat{\lambda}_k \approx \hat{\lambda}_{k'}$. This may happen especially when K_0 is very large, because the change point is more likely to be estimated repeatedly. Consequently, the redundancy can prevent the estimation for model 5. In this case, only one change point should be retained.

At the convergence step, only the $K_1 (\leq K_0)$ likeliest change points remain. The fitted model is produced by

$$\hat{Y}_i = \hat{\gamma}^T Z_i + \hat{\alpha} X_i + \hat{\beta}_1 \tilde{U}_{i1} + \cdots + \hat{\beta}_{K1} \tilde{U}_{iK_1} \tag{12}$$

However, this is a rough estimation of the number of change points. For more accuracy, it is important that we assess whether the K_1 change points are really indispensable. That is to say, we should retain the really significant change points by removing the spurious ones. Determining that $\hat{\lambda}_k$ is a real change point is equivalent to examining whether or not U_k is a noise variable. If $\tilde{U}_k$ is a change point, then $\hat{\beta}_k \approx 0$. Therefore, selecting K_1 change points actually reduces to identifying the significant variables among $(\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_{K_1})$. Clearly, the issue of estimating the number of change points is transformed into a variable selection problem. Many modern techniques are used for selecting variables. For simplicity, we use the LASSO algorithm proposed by Efron et al. [15] to solve this problem. When the LASSO algorithm is employed, the commonly used penalized goodness-of-fit criteria for the model dimension are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Minimum Description Length (MDL). Following Muggeo and Adelfio [11], we choose the generalized BIC proposed by Wang et al. [16] as the penalized criterion. The expression of the generalized BIC is

$$\mathrm{BIC}_{C_n} = \log(\hat{\sigma}^2) + \mathrm{edf}\frac{\log(n)}{n} C_n \tag{13}$$

where $\hat{\sigma}^2$ is the residual variance estimate, edf represents the degrees of freedom for the model, quantified by the number of estimated parameters, and C_n is a given constant

related to n. In summary, the proposed algorithm for RCPLR can be expressed as follows.

1. Fix a set comprising a large number of change points, K_0, and initialize the corresponding coefficient values;
2. Compute the variables $\tilde{U}_{ij}$ and $\tilde{V}_{ij}$ for $j = 1, 2, \ldots, K$, and fit the transformed linear model 5 via a rank-based estimator;
3. Update the change points by extracting those change points whose estimate $\hat{\lambda}_j$ is not inadmissible, and delete the corresponding covariates, $(\tilde{U}_{ij}, \tilde{V}_{ij})$. Refit the piecewise linear model for the updated response by rank-based regression;
4. Repeat steps 2–3 until the number of change points is stabilized or remains unchanged; we denote this number as K_1;
5. Refine the change points through the LASSO algorithm, selecting the necessary variables from among $(\tilde{U}_{i1}, \ldots, \tilde{U}_{iK_1})$. The aim of this step is to discard the noise variables, $\tilde{U}_{ij}$ and retain the significant change points, numbering K_2.
6. Fit the segmented linear model with the number of change points, $K = K_2$ by iteratively estimating the above algorithm. The ultimate regression coefficients, the number and locations of the change points can be obtained now.

Remark 1 When the LASSO algorithm is used to select the significant change points, the edf in the generalized BIC expression 8 is actually the degrees of freedom of the model 5. Note that edf changes if one or more change points are deleted. Thus, edf is a changing value and can be calculated by $\text{edf} = 1 + 2 \cdot K' + q$,where K' denotes the current number of change points in the present iteration process.

Remark 2 As discussed by Wang et al. [16], the usage of $C_n > 1$ plays an important role in finding the optimal model when the number of parameters is not fixed, but it diverges as $n \to +\infty$. Supported by simulation studies, Muggeo and Adelfio [11] show that $C_n = \log\log n$ appears to be the most suitable value for a piecewise constant model based on Gaussian-distributed IID errors. In this paper, we choose $C_n = \log\log n$ because it works reasonably well when the error structure is typically unknown in practice.

Remark 3 As the change points are selected, we also need to test for their existence in the statistics. However, this involves some additional work. Hahn et al. [17] suggests a potential method to do so, but it is beyond the scope of this paper.

3 Numerical studies

Two simulations are presented in this section to assess the performance of the proposed approach. The first consists in identifying the number of change points. The second is to evaluate the finite sample properties of the proposed rank-based estimators. All codes are written and executed in the R language with some additional packages.

3.1 Simulation 1

To evaluate the performance of the robust regression in detecting the number of change points, we consider the following two scenarios:

(I) Independent and identically distributed (IID): $Y = \alpha_0 + \alpha_1 x + \gamma^Z + \sum_{j=1}^{K} \beta_j (x - \lambda_j)_+ + e$

(II) Heteroscedasticity: $Y = \alpha_0 + \alpha_1 x + \gamma^Z + \sum_{j=1}^{K} \beta_j (x - \lambda_j)_+ + (1 + 0.2z)e$

where x is generated from a uniform distribution, $U(-10, 6)$ and z is generated from a binomial distribution, $B(1, 0.5)$. For each scenario, we consider four different error structures: (1) $e \sim N(0, 1)$, (2) $e \sim t_4$, (3) $e \sim 0.9N(0, 1) + 0.1t_4$, and (4) $e \sim 0.9N(0, 1) + 0.1Cauchy(0, 1)$ (t_4 is a Student t distribution with four degrees of freedom). To study the capability of the proposed method, we set the true number of change points as $K = 0, 2, 4, 6$ for each scenario. We consider $K = 0$ to check whether the proposed method could detect the existence of change points effectively. The regression coefficients and change point parameters are set up as follows:

1. when $K = 0$, $(\alpha_0, \alpha_1, \gamma)^T = (1, 3, -2)^T$;
2. when $K = 2$, $(\alpha_0, \alpha_1, \gamma, \beta_1, \beta_2)^T = (1, 3, -2, 4, -3)^T$, $(\lambda_1, \lambda_2)^T = (-6, 2)^T$;
3. when $K = 4$, $(\alpha_0, \alpha_1, \gamma, \beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 3, -2, 4, -3, 2, 4)^T$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T = (-8, -4.5, -0.5, 4)^T$;
4. when $K = 6$, $(\alpha_0, \alpha_1, \gamma, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T = (1, 3, -2, 4, -5, 4, -3, 5, -6)^T$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6)^T = (-8.5, -6, -3, -0.5, 2, 4)^T$.

For each simulation setup, we conduct 1000 repetitions with sample sizes $n = 100$, 200, 500, and 1000.

We also compare the proposed method with the cumSeg method proposed by Muggeo and Adelfio [11]. The cumSeg method can be implemented with the existing R package cumSeg developed by Muggeo. As Muggeo has not provided the estimation procedure for the covariates without threshold effect, we upgrade the cumSeg package and incorporate the additional non-threshold variables into the model-fitting framework. The comparison results for different sample sizes, error types, and numbers of change points are summarized as follows: (see Tables 1, 2).

The mean numbers of change points detected by our method are closer to the true preset values in almost every setup. When $K = 0$, our method attains an accuracy rate of at least 42.9%. By contrast, the cumSeg method has an accuracy rate of less than 1%, indicating that our method can detect the existence of change points effectively. Moreover, our method performs well as K increase, even with sample sizes of less than $n = 100$, 200. Specifically, results detected by the proposed method have sometimes exceeded 90%. On the contrary, the cumSeg rarely achieves nonzero accuracy rates (and only with large sample sizes) that are nevertheless lower than those of the proposed method. Indeed, the cumSeg method yields 0% accuracy in many cases. Clearly, our method outperforms the cumSeg method in identifying the number of change points, regardless of the error structure, variance type, or sample size.

When the error terms follow the standard Gaussian distribution, the proposed method works quite favorably, as does Muggeo's method too though not better than the proposed method. However, when it comes to heavy-tailed or contaminated distributions, the proposed linearization technique outperforms the cumSeg method to a

Table 1 Performance comparison between the proposed method and cumSeg method for IID scenario

True	Error		Proposed				*cumSeg*			
			100	200	500	1000	100	200	500	1000
$K=0$	1	m	0.577	0.541	0.497	0.465	5.865	5.382	4.911	4.364
		s	0.977	0.804	0.793	0.545	0.983	0.848	0.820	0.580
		a	0.429	0.461	0.506	0.537	0.028	0.030	0.040	0.041
	2	m	0.570	0.553	0.518	0.498	9.362	8.245	6.226	5.010
		s	9.782	0.788	0.912	0.659	1.018	0.880	0.965	0.729
		a	0.432	0.454	0.487	0.510	0.000	0.000	0.000	0.000
	3	m	0.582	0.562	0.519	0.499	9.606	8.852	6.948	5.635
		s	0.926	0.834	0.736	0.495	0.979	0.865	0.796	0.553
		a	0.422	0.443	0.486	0.511	0.000	0.000	0.000	0.000
	4	m	0.567	0.510	0.482	0.476	9.475	8.741	6.861	5.587
		s	1.322	1.780	2.336	2.257	1.619	1.990	2.629	2.864
		a	0.438	0.492	0.522	0.529	0.000	0.000	0.000	0.000
$K=2$	1	m	2.508	2.462	2.422	2.342	3.435	3.399	2.617	1.958
		s	0.919	1.006	0.912	0.897	1.010	1.070	1.122	0.979
		a	0.633	0.664	0.676	0.728	0.089	0.086	0.132	0.223
	2	m	2.468	2.475	2.416	2.345	7.902	6.031	4.084	2.552
		s	1.049	1.033	0.967	0.892	1.157	1.055	1.057	0.921
		a	0.623	0.661	0.679	0.718	0.000	0.000	0.047	0.250
	3	m	2.474	2.488	2.409	2.361	8.513	6.588	4.800	3.381
		s	0.945	0.932	0.924	0.918	1.042	1.056	1.024	0.995
		a	0.651	0.642	0.680	0.693	0.000	0.000	0.014	0.138
	4	m	2.156	2.164	2.101	2.066	8.334	6.572	4.691	3.209
		s	2.335	2.217	2.099	1.865	2.635	2.310	2.191	1.959
		a	0.515	0.529	0.549	0.550	0.000	0.000	0.009	0.174
$K=4$	1	m	2.919	4.128	4.129	4.077	1.072	1.742	5.402	4.656
		s	1.457	1.323	1.089	0.936	1.694	1.214	1.113	0.954
		a	0.534	0.821	0.880	0.926	0.015	0.121	0.475	0.512
	2	m	1.927	3.925	4.116	4.092	5.921	5.306	5.743	5.071
		s	1.509	1.245	0.823	0.587	1.517	1.349	0.848	0.617
		a	0.334	0.795	0.873	0.917	0.004	0.047	0.395	0.484
	3	m	3.150	4.118	4.122	4.086	1.090	1.903	5.493	4.824
		s	1.537	1.233	1.056	0.897	1.732	1.319	1.073	0.924
		a	0.577	0.820	0.879	0.920	0.021	0.139	0.150	0.506
	4	m	1.581	2.256	2.525	2.438	0.740	0.946	1.670	1.909
		s	1.649	1.522	1.289	0.939	1.949	1.602	1.389	0.968
		a	0.268	0.426	0.528	0.527	0.005	0.031	0.160	0.230
$K=6$	1	m	4.687	5.931	6.032	6.023	1.211	1.757	4.889	5.660
		s	0.961	0.793	0.724	0.682	1.224	1.093	0.924	0.882
		a	0.446	0.804	0.914	0.950	0.000	0.024	0.431	0.475

Table 1 continued

True	Error		Proposed				*cumSeg*			
			100	200	500	1000	100	200	500	1000
	2	m	3.773	5.710	6.021	6.027	1.169	1.366	4.106	5.336
		s	1.484	0.983	0.877	0.847	1.784	1.481	1.665	0.972
		a	0.287	0.747	0.893	0.941	0.000	0.000	0.276	0.452
	3	m	4.867	5.927	6.016	6.021	1.209	1.814	4.997	5.695
		s	0.876	0.645	0.518	0.492	1.529	1.262	1.139	1.011
		a	0.498	0.806	0.900	0.933	0.000	0.013	0.413	0.438
	4	m	3.107	4.131	4.128	4.128	1.059	1.174	1.826	2.288
		s	0.989	0.889	0.833	0.875	1.923	1.785	1.685	1.388
		a	0.230	0.477	0.541	0.582	0.000	0.001	0.060	0.141

m and s: the mean and standard deviation of detected change points number in simulation studies, respectively; a: the accuracy rate, that is, the proportion of detected change points number equaling to the real number

Table 2 Performance comparison between the proposed method and cumSeg method for heteroscedasticity scenario

True	Error		Proposed				*cumSeg*			
			100	200	500	1000	100	200	500	1000
$K=0$	1	m	0.506	0.468	0.438	0.389	9.404	8.332	6.564	4.237
		s	0.527	0.515	0.514	0.496	0.580	0.541	0.530	0.521
		a	0.508	0.540	0.571	0.615	0.024	0.033	0.045	0.056
	2	m	0.490	0.454	0.434	0.408	9.876	7.546	5.885	4.034
		s	0.516	0.504	0.514	0.517	0.552	0.546	0.579	0.602
		a	0.518	0.549	0.575	0.605	0.000	0.000	0.000	0.000
	3	m	0.495	0.467	0.417	0.427	8.707	7.850	6.540	5.893
		s	0.518	0.523	0.511	0.509	0.631	0.557	0.546	0.503
		a	0.514	0.545	0.591	0.580	0.000	0.000	0.000	0.000
	4	m	0.508	0.433	0.401	0.400	5.200	5.169	5.086	4.890
		s	0.510	0.504	0.498	0.494	0.795	0.627	0.535	0.565
		a	0.497	0.570	0.603	0.602	0.000	0.000	0.000	0.000
$K=2$	1	m	2.505	2.524	2.416	2.361	8.221	6.441	4.609	3.098
		s	1.023	0.977	0.956	0.942	1.023	0.989	1.141	1.002
		a	0.620	0.629	0.686	0.709	0.081	0.086	0.137	0.285
	2	m	2.455	2.459	2.421	2.343	7.843	5.858	3.819	2.337
		s	1.033	0.945	0.985	0.927	1.064	1.043	1.083	0.977
		a	0.611	0.662	0.693	0.718	0.000	0.000	0.064	0.219
	3	m	2.485	2.480	2.427	2.327	8.349	6.536	4.801	3.257
		s	1.004	0.986	0.985	0.926	1.066	1.044	1.006	0.964
		a	0.620	0.661	0.662	0.721	0.000	0.000	0.009	0.185
	4	m	2.150	2.134	2.054	1.975	4.322	4.740	3.362	3.190

Table 2 continued

True	Error		Proposed				*cumSeg*			
			100	200	500	1000	100	200	500	1000
		s	1.312	1.239	1.233	1.227	2.312	2.554	1.839	1.445
		a	0.477	0.520	0.524	0.562	0.000	0.000	0.009	0.198
$K=4$	1	m	2.798	4.161	4.122	4.073	1.054	1.746	5.386	4.693
		s	0.878	0.573	0.520	0.449	1.676	1.237	1.099	0.971
		a	0.518	0.810	0.871	0.925	0.016	0.123	0.375	0.503
	2	m	1.885	3.861	4.142	4.067	0.891	1.360	4.681	4.115
		s	1.074	0.833	0.805	0.778	1.574	1.210	1.099	0.928
		a	0.296	0.784	0.870	0.938	0.004	0.053	0.173	0.474
	3	m	3.126	4.155	4.116	4.064	1.052	1.840	4.490	5.770
		s	0.588	0.852	0.833	0.795	0.688	1.285	1.087	0.959
		a	0.576	0.806	0.890	0.935	0.016	0.140	0.146	0.488
	4	m	1.359	2.303	2.487	2.507	0.744	0.925	1.635	1.944
		s	0.462	0.766	0.835	0.827	0.666	0.836	1.660	1.885
		a	0.223	0.449	0.527	0.557	0.005	0.024	0.161	0.214
$K=6$	1	m	4.429	5.903	6.012	6.006	1.207	1.605	4.844	5.643
		s	0.335	0.895	0.784	0.778	0.535	1.096	1.288	0.970
		a	0.399	0.774	0.906	0.921	0.000	0.016	0.428	0.476
	2	m	3.264	5.533	5.992	6.008	1.124	1.324	3.715	5.336
		s	0.533	0.666	0.772	0.754	0.420	0.760	1.762	0.959
		a	0.160	0.691	0.872	0.928	0.000	0.001	0.220	0.455
	3	m	4.592	5.924	6.017	6.018	1.251	1.686	4.917	5.675
		s	0.735	0.798	0.754	0.757	0.620	1.175	1.245	1.004
		a	0.430	0.800	0.909	0.942	0.000	0.010	0.219	0.438
	4	m	2.759	3.926	3.985	4.135	0.994	1.082	1.585	2.200
		s	0.533	0.854	0.832	0.848	0.510	0.754	1.542	2.044
		a	0.173	0.457	0.512	0.563	0.000	0.002	0.050	0.133

m and s: the mean and standard deviation of detected change points number in simulation studies, respectively; a: the accuracy rate, that is, the proportion of detected change points number equaling to the real number

greater degree, with the latter performing less effectively than in the Gaussian distribution case. In the heteroscedasticity scenario, too, we draw the same conclusions from the accuracy rates and mean values. This is not surprising, since the cumSeg method is based on ordinary least squares optimization techniques. By comparison, our rank-based estimators are robust against heteroscedasticity and heavy-tailed errors. This conclusion is also confirmed in the subsequent simulations works.

As expected, with the sample size increasing, the performances of both methods behave more outstanding than the small sample sizes situations. The evidence is apparent. In almost all the simulation setups, the accuracy rates detected by larger sample sizes exceed those detected by less sample sizes and their averages figures of detecting

change points are more close to the true values. This phenomenon behaves even more significant in the proposed method results, supporting that the sample size can exert a substantial effect on the capacity of detecting change points.

3.2 Simulation 2

We evaluate the finite sample performance of the proposed robust estimators in Sect. 2.4. This simulation setting is similar as above, but here, we only consider $K = 2; 4$ with sample size $n = 1000$ for saving space. Meanwhile, we also compare the performance with Muggeo's method [18], which was easily implemented in R package segmented. Fortunately, this package allows for the existence of covariates that are not subjected to thresholding and therefore does not need some additional adjustment. The brief results of the two estimators are collected and exhibited in Tables 3 and 4.

From these tables, both the two estimators are consistent, as their biases ("bias") are small. In addition, the estimated standard errors ("ESE") are close to the empirical standard deviations ("SD"). The coverage probabilities ("CP") for both threshold parameters and regression coefficients approach the nominal level 95%. However, the mean squares errors ("MSE") and the average widths ("AW") of our estimators are slightly lager than those of Muggeo's when the error terms follow a standard normal distribution. This phenomenon also appears in Zhang and Li [10]. It is reasonable because the rank-based estimators can achieve 95% relative efficiency faster than least square estimators as demonstrated in Hettmansperger and Mckean [8].

When the error terms follow heavy-tailed (t_4 distribution) or contaminated distributions, the proposed method works more favorably for both homoscedasticity and heteroscedasticity scenarios. It is clear that, the SDs and MSEs of our method are relatively small, and the average confidence intervals are shorter than those of Muggeo's for most estimators. Besides, the empirical coverage probabilities of our estimator approach the nominal level more closely than those estimators for comparison. The above findings once again demonstrate convincingly that the proposed rank-based estimators have advantages in dealing with outliers and heavy-tailed errors.

It should be noted that the thresholds that are close to either edges of the domain of threshold variable perform not as well as those who are located in the middle. The phenomenon is more obvious for more change points. This may be because there are not enough observed data in one side of threshold, and thus consequently effects the threshold parameters estimation results.

In short, the proposed method has desirable finite sample performance, especially when dealing with non-standard error structures.

4 An empirical illustration

It is well known that obesity has become a common threat to human health, especially epidemic in childhood. Obesity not only brings long-term inconvenience in daily life, but also may lead to high blood pressure, fatty liver, diabetes and many other health problems. Body mass index, or BMI, calculated by dividing weight (in kg) by height

Table 3 Simulation results based on 1000 simulated samples of 1000 observations under four different error distributions for model with $K = 2$ change points for both IID and heteroscedasticity cases

Error		α_0	α_1	γ	β_1	β_2	λ_1	λ_2
IID								
1	Bias	−0.001	0.000	0.000	−0.001	0.001	−0.001	0.000
	SD	0.452	0.066	0.056	0.059	0.058	0.040	0.057
	ESE	0.457	0.065	0.056	0.060	0.060	0.040	0.053
	MSE	0.204	0.004	0.003	0.004	0.003	0.002	0.003
	CP	0.956	0.943	0.958	0.951	0.959	0.946	0.932
	AW	1.791	0.255	0.221	0.234	0.234	0.156	0.208
2	Bias	−0.025	0.001	−0.003	0.003	0.001	−0.002	−0.003
	SD	0.545	0.073	0.067	0.070	0.073	0.049	0.063
	ESE	0.534	0.076	0.066	0.070	0.070	0.046	0.062
	MSE	0.297	0.005	0.004	0.005	0.005	0.002	0.004
	CP	0.944	0.964	0.949	0.938	0.942	0.942	0.950
	AW	2.093	0.297	0.258	0.274	0.274	0.182	0.243
3	Bias	0.008	0.001	0.001	0.000	0.001	0.002	−0.003
	SD	0.414	0.060	0.051	0.053	0.055	0.037	0.050
	ESE	0.418	0.059	0.052	0.055	0.055	0.036	0.049
	MSE	0.172	0.004	0.003	0.003	0.003	0.001	0.003
	CP	0.957	0.941	0.954	0.955	0.941	0.944	0.941
	AW	1.639	0.233	0.202	0.214	0.214	0.143	0.190
4	Bias	−0.031	−0.002	−0.004	0.004	0.001	−0.002	0.001
	SD	0.482	0.067	0.059	0.062	0.062	0.042	0.053
	ESE	0.465	0.066	0.057	0.061	0.061	0.040	0.054
	MSE	0.234	0.004	0.004	0.004	0.004	0.002	0.003
	CP	0.951	0.950	0.951	0.949	0.946	0.942	0.949
	AW	1.821	0.259	0.225	0.238	0.238	0.158	0.211
Heteroscedasticity								
1	Bias	−0.004	−0.004	0.000	0.000	0.002	0.000	0.001
	SD	0.495	0.070	0.061	0.065	0.069	0.044	0.062
	ESE	0.502	0.071	0.062	0.066	0.066	0.044	0.058
	MSE	0.245	0.005	0.004	0.004	0.005	0.002	0.004
	CP	0.947	0.952	0.953	0.950	0.939	0.945	0.927
	AW	1.967	0.279	0.242	0.257	0.257	0.171	0.229
2	Bias	0.021	0.004	0.002	−0.001	0.000	0.003	−0.004
	SD	0.600	0.083	0.074	0.078	0.079	0.053	0.069
	ESE	0.585	0.083	0.072	0.077	0.077	0.051	0.068
	MSE	0.360	0.007	0.006	0.006	0.006	0.003	0.005
	CP	0.944	0.945	0.945	0.942	0.940	0.944	0.941
	AW	2.294	0.327	0.283	0.300	0.300	0.200	0.267
3	Bias	−0.011	−0.001	−0.002	0.002	−0.003	0.000	0.001
	SD	0.472	0.062	0.058	0.061	0.062	0.041	0.054

Table 3 continued

Error		α_0	α_1	γ	β_1	β_2	λ_1	λ_2
	ESE	0.458	0.065	0.056	0.06	0.060	0.040	0.053
	MSE	0.222	0.004	0.003	0.004	0.004	0.002	0.003
	CP	0.945	0.966	0.949	0.951	0.950	0.936	0.949
	AW	1.794	0.255	0.221	0.235	0.235	0.156	0.209
4	Bias	−0.016	0.005	−0.002	0.004	−0.002	0.001	0.002
	SD	0.517	0.075	0.064	0.066	0.068	0.046	0.061
	ESE	0.509	0.072	0.063	0.067	0.067	0.044	0.059
	MSE	0.267	0.006	0.004	0.004	0.005	0.002	0.004
	CP	0.939	0.935	0.942	0.943	0.949	0.942	0.938
	AW	1.995	0.284	0.246	0.261	0.261	0.174	0.232

Bias: the empirical bias; SD: the empirical standard deviations; ESE: the average of estimated standard errors; MSE: the average of the squares of estimated errors; CP: 95% coverage probabilities; AW: the average width of asymptotic intervals

(in m), is a commonly used tool for assessing and monitoring human being's health condition. Without doubt, human body functions are in a dynamic change with the increase in ages. One interest in nutriology research is to capture the association of age with BMI. For instance, Krobot et al. [19] studied the interrelationship among some potential factors including age with BMI for Germany, and showed that there was a strong correlation between age and BMI. Thinggaard et al. [20] found that the relation between BMI and mortality is changing with advancing age, indicating that there may exist a threshold effect. All these literatures focused on mean regression, neglecting the influence of outliers and heteroscedasticity. Although Chen [21] adopted smooth quantile regression technique to model for the two variables, it cannot provide any information on the locations of thresholds, which is often of great research interest. In this section, we analyze a BMI and age data from the National Health and Nutrition Examination Survey (NHAENES) by using the proposed robust continuous thresholds model, RCPLR model.

The NHAENES program, began in the early 1960s, was carried out to monitor the health condition and nutritional status of adults and children in USA. Here, we only focus on a subset collected in year 2005–2006, including 3191 males and 2993 females. To eliminate the effects of gender, we analyze the data for females and males separately. Table 5 reports the descriptive statistics. From the table, both the two skewness indexes are greater than 0, implying that the two groups of data are positively skewed and therefore asymmetric. Skewness indexes indicate that the sample distributions are leptokurtic and fat tailed. These truths are also verified in Fig. 1. As displayed in Fig. 1, the empirical kernal probability density (the solid line) makes a significant difference with the practical normal density (the dotted line), which means that the BMI data for both male and female does not follow normal distribution.

Figures 2 and 3 show that the relationships between BMI and age for both male and female subjects are clearly piecewise linear and may exist multiple continuous

Table 4 Simulation results based on 1000 simulated samples of 1000 observations under four different error distributions for model with $K = 4$ change points for both IID and heteroscedasticity cases

Error		α_0	α_1	γ	β_1	β_2	β_3	β_4	λ_1	λ_2	λ_3	λ_4
IID												
1	Bias	0.099	0.099	0.011	−0.014	0.007	0.012	0.006	0.014	0.007	0.008	−0.010
	SD	1.872	0.063	0.209	0.324	0.247	0.281	0.247	0.196	0.187	0.224	0.188
	ESE	1.838	0.065	0.159	0.294	0.239	0.274	0.166	0.188	0.164	0.189	0.155
	MSE	3.508	0.004	0.044	0.105	0.061	0.079	0.061	0.038	0.035	0.050	0.035
	CP	0.936	0.964	0.932	0.944	0.946	0.952	0.944	0.938	0.92	0.934	0.936
	AW	1.637	0.255	0.625	0.682	0.351	0.290	0.651	0.227	0.250	0.350	0.216
2	Bias	0.028	0.002	0.004	−0.014	0.009	−0.013	0.011	0.017	0.022	0.010	0.017
	SD	1.698	0.077	0.271	0.437	0.341	0.386	0.321	0.274	0.262	0.301	0.264
	ESE	1.691	0.076	0.187	0.404	0.325	0.387	0.295	0.268	0.274	0.304	0.265
	MSE	2.737	0.006	0.073	0.191	0.116	0.149	0.103	0.075	0.069	0.091	0.070
	CP	0.938	0.960	0.940	0.938	0.950	0.946	0.948	0.914	0.904	0.922	0.898
	AW	1.627	0.298	0.734	0.801	0.410	0.340	0.763	0.266	0.291	0.407	0.253
3	Bias	0.080	−0.003	0.009	−0.019	0.019	−0.031	0.026	0.024	0.019	0.030	−0.027
	SD	1.475	0.061	0.281	0.494	0.392	0.472	0.349	0.315	0.321	0.359	0.313
	ESE	1.317	0.06	0.246	0.459	0.382	0.468	0.352	0.353	0.358	0.332	0.351
	MSE	1.622	0.064	0.079	0.244	0.154	0.224	0.122	0.100	0.103	0.130	0.099
	CP	0.934	0.936	0.932	0.936	0.940	0.946	0.920	0.914	0.940	0.926	0.938
	AW	2.161	0.234	0.572	0.625	0.32	0.267	0.595	0.208	0.228	0.32	0.198
4	Bias	0.009	−0.001	0.001	−0.001	0.001	−0.006	0.004	0.011	0.009	0.012	−0.009
	SD	1.415	0.067	0.215	0.324	0.240	0.275	0.246	0.189	0.196	0.222	0.189
	ESE	1.466	0.066	0.192	0.297	0.231	0.275	0.227	0.179	0.165	0.191	0.186
	MSE	1.660	0.046	0.046	0.105	0.058	0.076	0.061	0.036	0.039	0.049	0.036

Table 4 continued

Error		α_0	α_1	γ	β_1	β_2	β_3	β_4	λ_1	λ_2	λ_3	λ_4
	CP	0.938	0.948	0.94	0.942	0.926	0.946	0.952	0.940	0.954	0.924	0.930
	AW	1.745	0.259	0.637	0.694	0.356	0.295	0.665	0.231	0.254	0.356	0.221
Heteroscedasticity												
1	Bias	−0.031	0.006	−0.003	−0.001	−0.003	0.007	−0.016	−0.002	0.007	−0.003	0.002
	SD	1.706	0.073	0.189	0.198	0.099	0.077	0.195	0.069	0.073	0.105	0.067
	ESE	1.599	0.071	0.177	0.193	0.098	0.081	0.184	0.064	0.070	0.098	0.061
	MSE	2.904	0.005	0.036	0.039	0.010	0.006	0.038	0.005	0.005	0.011	0.005
	CP	0.926	0.944	0.934	0.942	0.942	0.958	0.936	0.938	0.934	0.930	0.914
	AW	1.968	0.280	0.695	0.756	0.383	0.318	0.722	0.250	0.273	0.384	0.238
2	Bias	0.032	−0.009	0.036	−0.080	0.075	−0.099	0.062	0.074	0.076	0.085	−0.075
	SD	1.871	0.079	0.464	0.839	0.675	0.806	0.578	0.542	0.544	0.615	0.542
	ESE	1.840	0.084	0.404	0.823	0.615	0.797	0.611	0.575	0.582	0.614	0.507
	MSE	3.502	0.006	0.216	0.709	0.460	0.657	0.337	0.298	0.300	0.385	0.298
	CP	0.925	0.963	0.915	0.918	0.918	0.945	0.936	0.881	0.902	0.912	0.912
	AW	2.015	0.328	0.800	0.876	0.450	0.382	0.828	0.294	0.322	0.446	0.274
3	Bias	0.021	0.001	−0.022	−0.049	0.039 −	−0.045	0.016	0.045	0.044	0.049	−0.042
	SD	1.830	0.084	0.372	0.649	0.52	0.602	0.460	0.413	0.404	0.463	0.405
	ESE	1.830	0.083	0.363	0.622	0.514	0.596	0.412	0.395	0.381	0.414	0.407
	MSE	1.691	0.007	0.139	0.423	0.272	0.363	0.211	0.173	0.165	0.216	0.166
	CP	0.918	0.952	0.926	0.932	0.934	0.932	0.926	0.920	0.918	0.908	0.920
	AW	2.172	0.327	0.795	0.869	0.448	0.376	0.831	0.292	0.319	0.445	0.274
4	Bias	−0.069	0.000	−0.007	0.005	0.000	−0.004	0.003	0.006	0.008	0.002	−0.009
	SD	1.977	0.071	0.220	0.326	0.238	0.267	0.258	0.187	0.194	0.223	0.190
	ESE	1.914	0.073	0.219	0.295	0.231	0.283	0.286	0.165	0.171	0.221	0.162

Table 4 continued

Error		α_0	α_1	γ	β_1	β_2	β_3	β_4	λ_1	λ_2	λ_3	λ_4
	MSE	3.907	0.005	0.048	0.106	0.056	0.071	0.066	0.035	0.038	0.050	0.036
	CP	0.948	0.956	0.950	0.944	0.950	0.952	0.952	0.944	0.928	0.924	0.924
	AW	1.826	0.284	0.701	0.763	0.390	0.324	0.729	0.253	0.277	0.391	0.241

Bias: the empirical bias; SD: the empirical standard deviations; ESE: the average of estimated standard errors; MSE: the average of the squares of estimated errors; CP: 95% coverage probabilities; AW: the average width of asymptotic intervals

Table 5 Descriptive statistics analysis for BMI of males and females

	Mean	Min	Max	Median	SD	Skewness	Kurtosis
Male	26.95	12.41	57.41	25.78	7.11	0.89	1.03
Female	26.34	13.43	60.85	25.86	6.28	0.78	1.37

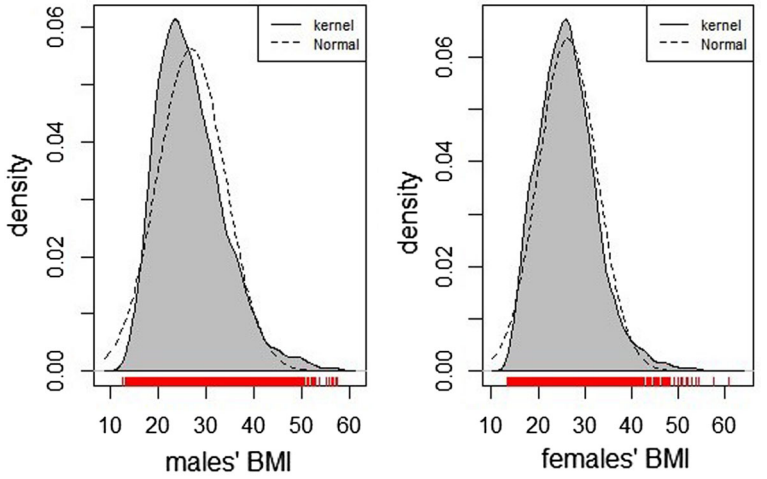

Fig. 1 The kernal probability density of BMI of males and females

change points. To capture this phenomenon, we use the continuous piecewise linear model to analyze the two datasets,

$$y_{i,j} = \alpha_0 + \alpha_1 x_{i,j} + \sum_{k=1}^{K_j} \beta_{k,j}(x_{i,j} - \lambda_{k,j}) + e_{i,j}, i = 1, \ldots, n_j, j = 1, 2 \qquad (14)$$

where y is BMI, x is the age, $\lambda_{k,j}$ is the kth location of threshold for jth dataset, and e without any distributional assumption. Here unknown multiple change points indicate that the age has different linear forms in different regions with BMI. We aim to estimate the change points number K_j, the regression coefficient ($\alpha_{0,j}$, $\alpha_{1,j}$, $\beta_{1,j}$, …, $\beta_{K_1,j}$) and the location of thresholds ($\lambda_{1,j}$, …, $\lambda_{K_j,j}$) for $j = 1, 2$.

The estimation results of Muggeo's cumSeg method [11] and our method for both male and female subjects are tabulated in Table 6. As shown in the Table, our method has detected three change points for both male and female subjects' datasets, while Muggeo's method only detects two change points for males and one change point for female. The estimation results of Muggeo's cumSeg method [11] and our method for both male and female subjects are tabulated in Table 6. As shown in the Table, our method has detected three change points for both male and female subjects' datasets, while Muggeo's method only detects two change points for males and one change point for female. This may be because the cumSeg method is based on ordinary least square

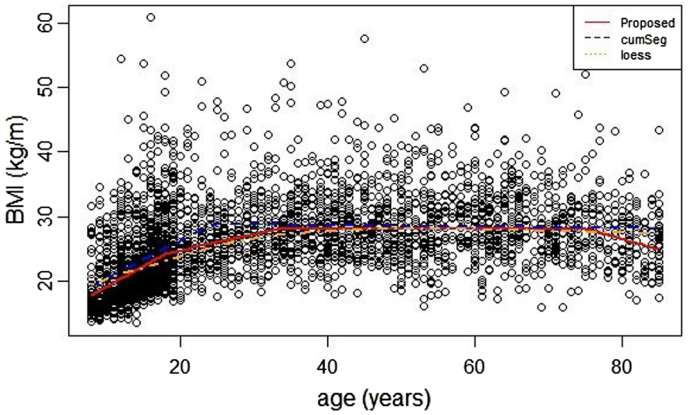

Fig. 2 Fitted curves for age versus BMI for female subjects

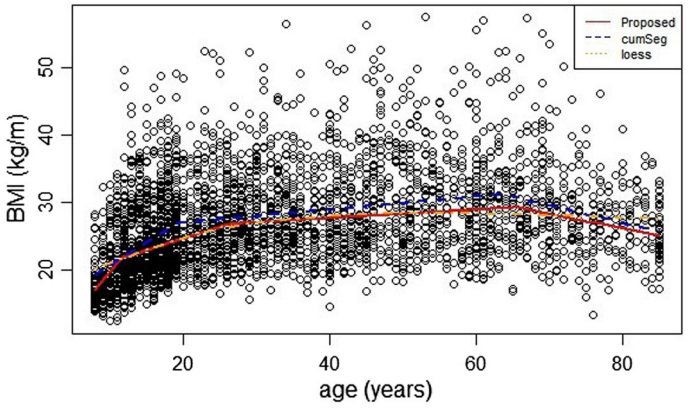

Fig. 3 Fitted curves for age versus BMI for male subjects

estimators, and thus more sensitive to outliers. By visual inspection of the fitted curves, we can find that the Muggeo's method is heavily influenced by many outliers above the overall point cluster region. In contrast, our estimates are more robust and could alleviate the influence of outliers. In addition, the change points estimates for male subjects are 11.638, 26.235, 65.654, which have prominent difference with female groups whose change points are 18.150, 23.669 and 75.229. This tells us that gender has an effect on the nonlinear relationship between age and BMI.

For comparison, we also take local polynomial regression (loess) into consideration. The loess method is flexible in fitting nonlinear trend, which is also suitable for continuous segmented linear situation. All the fitted curves are displayed in Figs. 2 and 3. From these figures, our method's fitted curves are closer to the loess curves, indicating that our method is more precious in depicting the segmented linear relationship than Muggeo's method. The curves of Muggeo's method are a little higher than ours, implying that the Muggeo's method is influenced by the outliers above the point cluster belt easily. Besides, the proposed method can provide direct information

Table 6 The estimates and standard errors ("se") for all parameters in age and BMI data

Subjects	Methods	α_0	α_1	β_1	β_2	β_3	λ_1	λ_2	λ_3
Male	Proposed	6.664 (3.040)	1.302 (0.314)	−0.958 (0.317)	−0.282 (0.042)	−0.279 (0.048)	11.638 (0.816)	26.235 (1.698)	65.654 (2.242)
	cumSeg	13.918 (0.837)	0.676 (0.058)	−0.581 (0.059)	−0.352 (0.0477)	–	19.325 (0.791)	63.627 (1.887)	–
Female	Proposed	12.652 (0.702)	0.640 (0.050)	−0.384 (0.068)	−0.253 (0.048)	−0.337 (0.111)	18.150 (1.167)	32.669 (1.965)	75.229 (2.205)
	SumSeg	14.532 (0.612)	0.584 (0.039)	−0.595 (0.040)	–	–	24.471 (0.785)	–	–

Figures in the parentheses denote the standard errors; "–" means that there are no detected change points

on the locations of the thresholds in contrast to no specified information provided by loess. These empirical findings clearly illustrate the availability of our method.

As inspired by Zhang and Li [10], to evaluate the performance of the model fitting, we also compare the prediction errors for both methods by using M-fold cross-validation. To be specific, we divide each dataset into M equal-scaled subgroups and denote D_m as mth subgroups. For each subgroup $m = 1, \ldots, M$, the corresponding prediction error is

$$\mathrm{PE}_m = \sum_{i \in D_m} \left[Y_i - \hat{Y}_i^{(-m)} \right]^2 \tag{15}$$

where $\hat{Y}_i^{(-m)}$ are the fitted values from the model estimated by using the data based on all the subgroups without D_m. Therefore, the total prediction error ("PE") is calculated by $\mathrm{PE} = \sum_{m=1}^{M} \mathrm{PE}_m$. Here, we set $K = 5$ for both male and female subjects' datasets. The total prediction errors of our method are 15,646.43, 12,256.49 for males and females, respectively. Both of them are less than that in Muggeo's method (16,017.34 for males and 12,811.73 for females), indicating that our method has higher predictive ability.

5 Discussion

Compared to the robust bent line regression model proposed by Zhang and Li [10], this paper extend their model into multiple change points situation and proposes a continuous piecewise linear regression model with more than one threshold, called RCPLR model. A complete rank-based estimation framework is also developed for the proposed model. The framework is based on a quite efficient method, the linearization technique, to estimate the locations of change points and the LARS algorithm to refine the number of change points via the generalized BIC.

All the estimating procedures are trained under the rank-based optimization routine to achieve robustness against outliers and heavy-tailed errors. Therefore, our method is computationally more efficient and robust. However, a formal test procedure to test the existence of multiple thresholds when the number is known is warranted. A possible idea concerning this investigation may follow the further research of Muggeo, see Hahn et al. [18].

References

1. Li C, He X (2011) Bent line quantile regression with application to an algometric study of land mammals' speed and mass. Biometrics 67(1):242–249
2. Lee S, Seo MH, Shin Y (2010) Testing for threshold effects in regression models. Publ Am Stat Assoc 106(493):220–231
3. Zhang L, Wang HJ, Zhu Z (2014) Testing for change points due to a covariate threshold in quantile regression. Stat Sin 24(4):1859–1877
4. Zhang L, Wang HJ, Zhu Z (2017) Composite change point estimation for bent line quantile regression. Ann Inst Stat Math 69(1):145–168

5. Zhang F, Li Q (2017) A continuous threshold expectile model. Comput Stat Data Anal 116:49–66
6. Yan Y, Zhang F, Zhou X (2017) A note on estimating the bent line quantile regression model. Kluwer, Dordrecht
7. Abebe A, Crimin K, Mckean JW, Haas JV, Vidmar TJ (2001) Rank-based procedures for linear models: applications to pharmaceutical science data. Ther Innov Regul Sci 35(35):947–971
8. Hettmansperger TP, Mckean JW (2011) Robust nonparametric statistical methods. Technometrics 16(3):477–478
9. Muggeo VMR (2003) Estimating regression models with unknown break-points. Stat Med 22(19):3055–3071
10. Zhang F, Li Q (2017) Robust bent line regression. J Stat Plan Inference 185:41–55
11. Muggeo VMR, Adelfio G (2011) Efficient change point detection for genomic sequences of continuous measurements. Oxford University Press, Oxford
12. Adelfio Giada (2012) Change-point detection for variance piecewise constant models. Commun Stat Simul Comput 41(4):437–448
13. Jureckova J (1971) Nonparametric estimate of regression coefficients. Ann Math Stat 42(4):1328–1338
14. Jaeckel LA (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Stat 43(5):1449–1458
15. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–451
16. Wang H, Li B, Leng C (2008) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc 32(2):928–961
17. Hahn G, Banergjee M, Sen B (2017) Parameter estimation and inference in a continuous piecewise linear regression model. Manuscript, Department of Statistics, Columbia University (December 2016), vol 32, no 2, pp 407–451
18. Muggeo VMR (2008). Segmented: an r package to fit regression models with broken-line relationships. In R News, pp 20–25
19. Krobot K, Hense HW, Cremer P, Eberle E, Keil U (1992) Determinants of plasma fibrinogen: relation to body weight, waist-to-hip ratio, smoking, alcohol, age, and sex. Results from the second monica augsburg survey 1989–1990. Arterioscler Thromb J Vasc Biol 12(7):780
20. Thinggaard M, Jeune JB, Martinussen T, Christensen K (2010) Is the relationship between bmi and mortality increasingly u-shaped with advancing age? a 10-year follow-up of persons aged 70–95 years. J Gerontol 65(5):526
21. Chen C (2005) Growth charts of body mass index (bmi) with quantile regression. In: International Conference on Algorithmic Mathematics and Computer Science, AMCS 2005, Las Vegas, Nevada, USA, pp 114–120