Chapter 5

Model Testing Based on Regression Spline

Na Li

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74858

Abstract

Tests based on regression spline are developed in this chapter for testing nonparametric functions in nonparametric, partial linear and varying-coefficient models, respectively. These models are more flexible than linear regression model. However, one important problem is if it is really necessary to use such complex models which contain nonparametric functions. For this purpose, p-values for testing the linearity and constancy of the nonparametric functions are established based on regression spline and fiducial method. In the application of spline-based method, the determination of knots is difficult but plays an important role in inferring regression curve. In order to infer the nonparametric regression at different smoothing levels (scales) and locations, multi-scale smoothing methods based on regression spline are developed to test the structures of the regression curve and compare multiple regression curves. It could sidestep the determination of knots; meanwhile, it could give a more reliable result in using the spline-based method.

Keywords: fiducial method, multi-scale smoothing method, nonparametric regression model, partial linear regression model, regression spline, varying-coefficient regression model

1. Introduction

It is well known that the model which contains nonparametric functions, such as partial linear model and varying-coefficient model, plays an important role in applications due to its flexible structure. However, in practice, investigators often want to know whether it is really necessary to fit the data with such more complex models rather than a simpler model. This amounts to testing the linearity of nonparametric functions in a regression model. In this chapter, we first consider the following three frequently used regression models.

Nonparametric regression model:

$$y = f(x) + \varepsilon. \tag{1}$$

Partial linear regression model:

$$y = Z'b + f(x) + \varepsilon. \tag{2}$$

Varying-coefficient model:

$$y = z_1 f_1(x_1) + \ldots + z_p f_p(x_p) + \varepsilon. \tag{3}$$

In models (1)–(3), y is the response variable, $Z = (z_1, \cdots, z_p)$ is a p-dimensional regressor, x and $x_1, \cdots, x_p$ are covariant taking values in a finite interval, ε is the error, b is a parameter vector, and *f(x)* and $f_j(x_j), j = 1, 2, \ldots, p$ are unknown smooth functions. Usually we suppose that (z, x) and ε are independent and $\varepsilon \tilde{} F(\cdot/\sigma)$, where F is a known cumulate distribution function (cdf) with mean 0 and variance 1; σ is unknown. Without loss of generality, we can suppose that x and $x_1, \cdots, x_p$ take values in [0, 1]. We try to test the linearity of $f(x)$ in models (1) and (2) and the constancy of $f_j(x_j)$ in model (3) for some $j \in (1, 2, \cdots, p)$.

The hypothesis testing in nonparametric regression model was considered in many papers. Härdle and Mammen [1] developed the visible difference between a parametric and a nonparametric curve estimates. Based on smoothing techniques, many tests were constructed for testing the linearity in regression model; see Hart [2], Cox et al. [3], and Cox and Koh [4] for a review. Recently, Fan et al. [5] studied a generalized likelihood ratio statistic, which behaves well in large sample case. Tests based on penalized criterion were developed by Eubank and Hart [6] and Baraud [7].

The linearity of partial linear regression model (2) was studied by Bianco and Boente [8], Liang et al. [9], and Fan and Huang [10]. There are also many other papers concerning such testing problems (see [11–16], among others). The constancy of the functional coefficient $f_j(x_j)$ in varying-coefficient model (3) was studied in Fan and Zhang [17], Cai et al. [18], Fan and Huang [19], You and Zhou [20], and Tang and Cheng [21]. Local polynomials and smoothing spline methods to estimate the coefficients in model (3) can be seen in Hoover et al. [22], Wu et al. [23], and so on.

The critical values of most of the previous tests were obtained by Wilks theorem or bootstrap method. So such tests only behave well in the case of relatively large sample size. This chapter would give some testing procedures based on regression spline and the fiducial method [24] in Section 2. It has a good performance even when the sample size is small.

In using the regression spline, the key problem is the determination of knots used in spline interpolation. As we know that, for smoothing methods such as kernel-based method and smoothing spline, the smoothness is controlled by smoothing parameters. For the well-known kernel estimate, the bandwidth that is extremely big or small might leads to over-smoothing or under-smoothing, respectively. In order to avoid the selection of an optimal smoothing parameter, multi-scale smoothing method was introduced by Chaudhuri and Marron [25, 26] based on kernel estimation for exploring structures in data. This multi-scale method is known as significant zero crossings of derivatives (SiZer) methodology. The basic idea of SiZer is to infer a nonparametric model by using a wide range of smoothing parameter (bandwidth) values rather than only using one "optimal" value in some sense.

There have been many versions of SiZer for various applications, such as the local likelihood version of SiZer in Li and Marron [27], the robust version of SiZer in Hannig and Lee [28], and the quantile version of SiZer in Park et al. [29]. In addition, Marron and deUñaÁlvarez [30] applied SiZer to estimate length biased, censored density, and hazard functions; Kim and Marron [31] utilized SiZer for jump detection and Park and Kang [32] applied SiZer to compare regression curves. The smoothing spline version of SiZer was proposed by Marron and Zhang [33]. It used the tuning parameter (penalty parameter) that controls the size of penalty as the smoothing parameter is.

Comparing with bandwidth for kernel-based method and tuning parameter for smoothing spline, it is more difficult to determine the number of knots and their positions. For this reason a multi-scale smoothing method based on regression spline is proposed in Section 3 to test the structures of nonparametric regression model. The proposed multi-scale method does not involve the determination of the "best" number of knots and can be extended easily to a more general case.

2. Tests for nonparametric function based on regression spline

In this section, the linearity of function $f(x)$ in model (1) is tested based on regression spline and fiducial method. Then, the proposed test procedure for model (1) is extended to test the linearity of model (2) and the constancy of function coefficient in model (3), respectively.

2.1. Test the linearity of nonparametric regression model

Without loss of generality, we suppose that x in model (1) takes values in [0, 1] and the set of knots is $\mathbf{T} = \{0 = t_1 < t_2, \cdots, < t_m = 1\}$. In order to estimate model (1), nonparametric function $f(x)$ is fitted by kth order splines with knots $\mathbf{T}$. This means that

$$f(x) \approx \sum_{j=1}^{m+k-1} \beta_j g_j(x), \tag{4}$$

where β_j is coefficient and $g_j(x)$, $j = 1, 2, \cdots, m + k - 1$, is basis function for order k splines, over the knots $t_1, t_2, \cdots, t_m$.

With n-independent observations $Y = (y_1, y_2, \cdots, y_n) \in \mathbb{R}^n$, the basis matrix $G_{n\times(m+k-1)}$ is defined by $G = \left\{g_j(x_i)\right\}$, x_i is the designed point, $i = 1, 2, \cdots, n; j = 1, 2, \cdots, m + k - 1$. Hence, model (1) can be approximated as $Y \approx G\beta + \varepsilon$. The least squares estimator of coefficients is

$$\widehat{\beta} = \left(G^T G\right)^{-1} G^T\, Y, \tag{5}$$

and the estimator of $f(x_i)$ can be expressed as

$$\widehat{Y} = \left\{\widehat{f}(x_1), \widehat{f}(x_2), \cdots, \widehat{f}(x_n)\right\}^T = G\left(G^T G\right)^{-1} G^T\, Y. \tag{6}$$

For testing the linearity of model (1), linear spline is used to approximate $f(x)$. It means that basis function $g_j(x)$ is a linear function:

$$g_1(x) = \frac{-x - t_2}{(t_2 - t_1)\mathbb{I}_2(t)},$$

$$g_{k-1}(x) = \frac{x - t_{k-2}}{(t_{k-1} - t_{k-2})\mathbb{I}_{k-1}(t)} - \frac{x - t_k}{(t_k - t_{k-1})\mathbb{I}_k(t)}, \qquad 3 \le k \le m, \tag{7}$$

$$g_m(x) = \frac{x - t_{m-1}}{(t_m - t_{m-1})\mathbb{I}_m(t)}.$$

In this case, the approximated function in (4) is a linear interpolation with k =1. The true value is $\beta_j = f\left(t_j\right), j = 1, 2, \cdots, m$. The linearity of function $f(x)$ can be written as

$$\mathrm{H}_0 : \frac{\beta_2 - \beta_1}{t_2 - t_1} = \frac{\beta_3 - \beta_2}{t_3 - t_2} = \cdots = \frac{\beta_m - \beta_{m-1}}{t_m - t_{m-1}}.$$

Null hypothesis H_0 can be expressed in matrix as $L'\beta = 0$,

where

$$L' = \begin{bmatrix} h_2 & -h_1 - h_2 & h_1 & 0\cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0\ \cdots & h_{m-1} & -h_{m-1} - h_{m-2} & h_{m-2} \end{bmatrix},$$

where $h_j = t_{j+1} - t_j, j = 1, 2, \cdots, m - 2$. Null hypothesis H_0 is equivalent to the following one:

$$\mathrm{H}_0^* : \ L'\beta = 0. \tag{8}$$

The p-value for testing hypothesis H_0^* will be derived by the fiducial method in the following context. Assume that matrix G has full rank, and let $\varepsilon\tilde{}\sigma N(0, 1)$. In model $Y = G\beta + \varepsilon$, the sufficient statistic of (β, σ^2) is $\left(\widehat{\beta}, S^2\right)$, where $\widehat{\beta}$ is defined in (5) and

$$S^2 = Y'(I - P_G)Y, \qquad P_G = G(G'G)^{-1}G'.$$

By Dawid and Stone [34], the sufficient statistic can be represented as a functional model:

$$\widehat{\beta} = \beta + \sigma(G'G)^{-\frac{1}{2}}E_1, \qquad S = \sigma E_2, E = (E_1,\ E_2) \sim Q, \tag{9}$$

where Q is the probability measure of $E = (E_1,\ E_2)$ and $E_1 \sim N(0, I_m)$ and, independently, $E_2^2 \sim \chi^2(n-m)$. From linear regression model, the fiducial model of β can be obtained:

$$\widehat{\beta} = \beta + \frac{S}{E_2}(G'G)^{-\frac{1}{2}}E_1, \qquad E = (E_1,\ E_2) \sim Q. \tag{10}$$

Given $\left(\widehat{\beta}, S^2\right)$, the distribution of the right side in fiducial model is the fiducial distribution of β. That is, the fiducial distribution of β is the conditional distribution of $R\left(E; \widehat{\beta}, S^2\right)$ when $\left(\widehat{\beta}, S^2\right)$ is given, where

$$R\left(E; \widehat{\beta}, S^2\right) = \widehat{\beta} - \frac{S}{E_2}(G'G)^{-\frac{1}{2}}E_1. \tag{11}$$

For testing hypothesis H_0^*, the p-value is defined as

$$p\left(\widehat{\beta}, S^2\right) = Q\left(\left\| L'\left[R\left(E; \widehat{\beta}, S^2\right) - \mathrm{E}_Q R\left(E; \widehat{\beta}, S^2\right)\right]\right\|_\Sigma^2 \geq \left\| L' \mathrm{E}_Q R\left(E; \widehat{\beta}, S^2\right)\right\|_\Sigma^2\right), \tag{12}$$

where $Q(\cdot)$ and E_Q express the probability for an event and the expectation of a random variable under Q, respectively, and Σ is the conditional covariance matrix of $L'\mathrm{E}_Q R\left(E; \widehat{\beta}, S^2\right)$ given $\widehat{\beta}$, S^2 and $\|v\|_\Sigma^2 = v'\Sigma^{-1}v$ for a vector v.

According to the definition of generalized pivotal quantity in [35], $R\left(E; \widehat{\beta}, S^2\right)$ is a generalized pivotal quantity and also a fiducial pivotal quantity about β. Naturally, $L'R\left(E; \widehat{\beta}, S^2\right)$ is the fiducial pivotal quantity about $L'\beta$. With the definition of Q in Eq. (10), we have that

$$p\left(\widehat{\beta}, S^2\right) = 1 - F_{m-2,\, n-m}\left(\frac{(n-m)\widehat{\beta}'L\left(L'(G'G)^{-1}L\right)^{-1}L'\widehat{\beta}}{(m-2)S^2}\right), \tag{13}$$

where $F_{m-2,\, n-m}$ is the cdf of F-distribution with degrees of freedom $m-2$ and $n-m$.

Under model (1) and the hypothesis that $f(x)$ is a linear function, null hypothesis H_0^* given in (8) is true. Suppose that the error is normally distributed, then the p-value given in Eq. (12) distributes as uniform distribution on interval (0, 1). On the other hand, under some mild condition, the test procedure based on $p\left(\widehat{\beta}, S^2\right)$ is consistent. Which means that $p\left(\widehat{\beta}, S^2\right)$ tends to be zero in probability 1 if H_0^* is false. The corresponding theoretical proof of the large sample properties and finite sample properties of $p\left(\widehat{\beta}, S^2\right)$ is the same as the proof given in Li et al. [36].

In applications, we need to check some hypotheses as follows:

$$\mathrm{H}_{01}: f(x) = C \Leftrightarrow \beta_1 = \beta_2 = \cdots = \beta_m,$$

$$\mathrm{H}_{02}: f(x) = Cx \Leftrightarrow \frac{\beta_2 - \beta_1}{t_2 - t_1} = \frac{\beta_3 - \beta_2}{t_3 - t_2} = \cdots = \frac{\beta_m - \beta_{m-1}}{t_m - t_{m-1}}, \text{and}, \beta_1 = 0.$$

The p-values for testing H_{01} and H_{02} can be obtained by replacing L in (12) by L_{01} and L_{02}, respectively, where $L_{02} = (e_1, L), e_1 = (1, 0, 0, \cdots, 0)'$ and

$$L_{01} = \begin{bmatrix} h_2 & -h_1 & 0 \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \ \cdots & h_m & -h_{m-1} \end{bmatrix}. \tag{14}$$

2.2. Test the linearity of partial linear model

To test the linearity of model (2), p-value can be established analogously. With n-independent observations $Y = (y_1, y_2, \cdots, y_n) \in \mathbb{R}^n$, model (2) can be represented as.

$$y_i = Z_i' b + f(x_i) + \varepsilon_i, i = 1, 2, \cdots, n,$$

where $Z_i' = (z_{i1}, \cdots, z_{ip})', b = (b_1, \cdots, b_p)', x_i, \quad i = 1, 2, \cdots, n$ are fixed designed points. With the approximation of $f(x)$ given in (4), model (2) can be approximated by $Y \approx X\theta + \varepsilon$, where $X = (\mathbb{Z}, G)_{n \times (p+m+1)}; \quad \mathbb{Z} = \ (z_{ij}); \quad i = 1, 2, \cdots, n; j = 1, 2, \cdots, p; G$ is the same as above; and $\theta = (b', , \beta')'$. Then p-value for testing the linearity of model (2) can be defined by replacing G in (12) by X, β by θ, and L by L_{03}, respectively, $L_{03} = (0_{(m-2) \times p}, L')'$.

The large sample and finite sample properties of the testing procedure for model (2) are the same as the test procedure for model (1).

2.3. Test the constancy of functional coefficient in varying-coefficient model

For model (3), investigators often want to know whether the coefficients are really varying; this means to test the constancy of the coefficient functions, that is, testing hypothesis:

$$\mathrm{H}_{31}: f_j(x) = C_j \text{ for } j = 1, 2, \cdots, p \text{ and some constant } C_j, \tag{15}$$

$$\mathrm{H}_{32}: f_{j0}(x) = C_{j0} \text{ for some } j = j_0 \text{ and some constant } C_{j0}. \tag{16}$$

With the set of knots $\mathbf{T} = \{0 = t_1 < t_2, \cdots, < t_m = 1\}$, coefficient $f_j(x)$ can also be approximated by

$$f_j(x) = \sum_{k=1}^{m} \beta_{js} g_j(x), j = 1, 2, \cdots, p,$$

where the true value of $\beta_{js} = f_j(t_k)$. Basic functions $g_j, j = 1, 2, \cdots, m+1$ were defined in (7). The varying-coefficient model (3) is approximately represented as

$$Y = X\beta + \varepsilon, \tag{17}$$

where $X = (F_1, \cdots, F_p)$ is $n \times mp$ matrix and $F_j = (z_{ji}f_k(x_i)), k = 1, 2, \cdots, m, i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, p$. $\beta = \left(\beta_1', \cdots, \beta_p{}'\right)'$ is mp-dimensional parametric vector, $\beta_j = \left(f_j(t_1), \cdots, f_p(t_m)\right)'$.

It is worth noting that under null hypothesis H_{31} defined in (15), regression model (3) is equivalent to model (17). However, this equivalence does not hold under null hypothesis H_{32} defined in (16). Null hypotheses H_{31} and H_{32} can be expressed in matrix as the following two, respectively:

$$H_{31}^* : L_1'\beta = 0, \tag{18}$$

$$H_{32}^* : L_2'\beta = 0, \tag{19}$$

where L_1' is $p(m-1) \times mp$ matrix.

$$L_1' = \begin{pmatrix} L_{01}{}' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_{01}{}' \end{pmatrix}, L_{01} = \begin{bmatrix} 1 & -1 & 0\cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0\ \cdots & 1 & -1 \end{bmatrix},$$

$$L_2' = \left(0_{(m-1)\times(mj_0-m)}, L', 0_{(m-1)\times(mp-mj_0)}\right)_{(m-1)\times mp}.$$

In the same way as the p-value in (13) is defined, p-value to test hypotheses H_{31}^* and H_{32}^* can be defined as below if the error ε distributes as normal distribution:

$$p_{31}\left(\widehat{\beta}, S^2\right) = 1 - F_{p(m-1),\, n-mp}\left(\frac{(n-mp)\widehat{\beta}'L_1\left(L_1'(X'X)^{-1}L_1\right)^{-1}L_1'\widehat{\beta}}{p(m-1)S^2}\right), \tag{20}$$

$$p_{32}\left(\widehat{\beta}, S^2\right) = 1 - F_{m-1,\, n-mp}\left(\frac{(n-mp)\widehat{\beta}'L_2\left(L_2'(X'X)^{-1}L_2\right)^{-1}L_2'\widehat{\beta}}{(m-1)S^2}\right). \tag{21}$$

According to the above discussion, it can be seen that $p_{31}\left(\widehat{\beta}, S^2\right)$ is uniformly distributed over (0, 1) under hypothesis H_{31}^*. However, under null hypothesis H_{32}^*, varying-coefficient model (2) is not linear. Hence, there is a difference between the distribution function of $p_{32}\left(\widehat{\beta}, S^2\right)$ under H_{32}^* and uniform distribution. This difference has an accurate expression, which can be seen in Li et al. [37] (Theorem 3). On the other hand, $p_{31}\left(\widehat{\beta}, S^2\right)$ and $p_{32}\left(\widehat{\beta}, S^2\right)$ both tend to be zero in probability if null hypotheses are false when sample size tends to be infinity under some mild conditions. The corresponding proof was provided also in Li et al. [37].

3. Multi-scale method based on regression spline

For regression spline, the number of knots controls the smoothness of the estimator. The determination of knots is important and plays a large influence on the inference results. The GCV method is usually used to choose an optimal number of knots. While, but after the number of knots is given, the determination of the optimal positions of knots is difficult. Shi and Li [38] chose knots by placing an additional new knot to reduce the value of GCV, until it could not be reduced by placing any additional knots. Hence, once a knot was selected, it cannot be removed from the knot set. Mao and Zhao [39] determined the locations of knots conditioned on the number of knots m first and chose m later by GCV criterion. In fact, the locations of knots can be considered as parameters which can be estimated from data. This is the free-knot spline; see DiMatteo et al. [40] and Sonderegger and Hannig [41]. However, the estimation of the optimal locations is computationally intractable, and replicate knots might appear in the estimated knot vectors [42].

On the other hand, many statisticians think that the statistical inference based on one smoothing level is not reliable although it is the optimal one. Therefore, multi-scale method is developed to estimate and test nonparametric regression curves. Chaudhuri and Marron [25, 26] proposed a multi-scale method to explore the significant structures (local minima and maxima or global trend) in data, which is known as SiZer. Significant zero crossings of derivatives (SiZer) is a powerful visualization technique for exploratory data analysis. It applies a large range of smoothing parameter values to do statistical inference simultaneously and use a 2D colored map (SiZer map) to summarize all of the results inferred at different smoothing levels (scales) and locations.

In this section, a regression spline version of SiZer is proposed for exploring structures of curve and comparing multiple regression curves, respectively. The proposed SiZer employs the number of knots as smoothing parameter (scales). For the sake of simplicity, linear spline is employed first to construct SiZer, which is denoted as SiZerLS. In addition, another version of SiZer—SiZerSS—is introduced, which is proposed in Marron and Zhang [33]. In SiZerSS, smoothing spline is used to infer the monotonicity of $f(x)$, and the tuning parameter (penalty parameter) that controls the size of penalty is chosen to be as the smoothing parameter. Finally, SiZer-RS, a version of SiZer based on higher-order spline interpolation, is constructed to compare multiple regression curves at different scales and locations simultaneously.

In order to understand SiZerLS clearly, we first present an example in which SiZerLS are simulated. This example is modified from Hannig and Lee [28] with the same regression function:

$$f(x) = 5 + 4.2\left(1 + \frac{|x - 0.3|}{0.03}\right) - 4 + 5.1\left(1 + \frac{|x-0.7|}{0.01}\right) - 4.$$

The observations generated from model (1) with 200 equally spaced design points from (0, 1) and $\sigma \sim N(0, 0.5)$ are plotted in **Figure 1**. Estimator $\widehat{f}_m(x)$ denotes the linear spline smoother obtained from (6) using m equally spaced knots chosen from (0, 1). The curves of $\widehat{f}_m(x)$ with

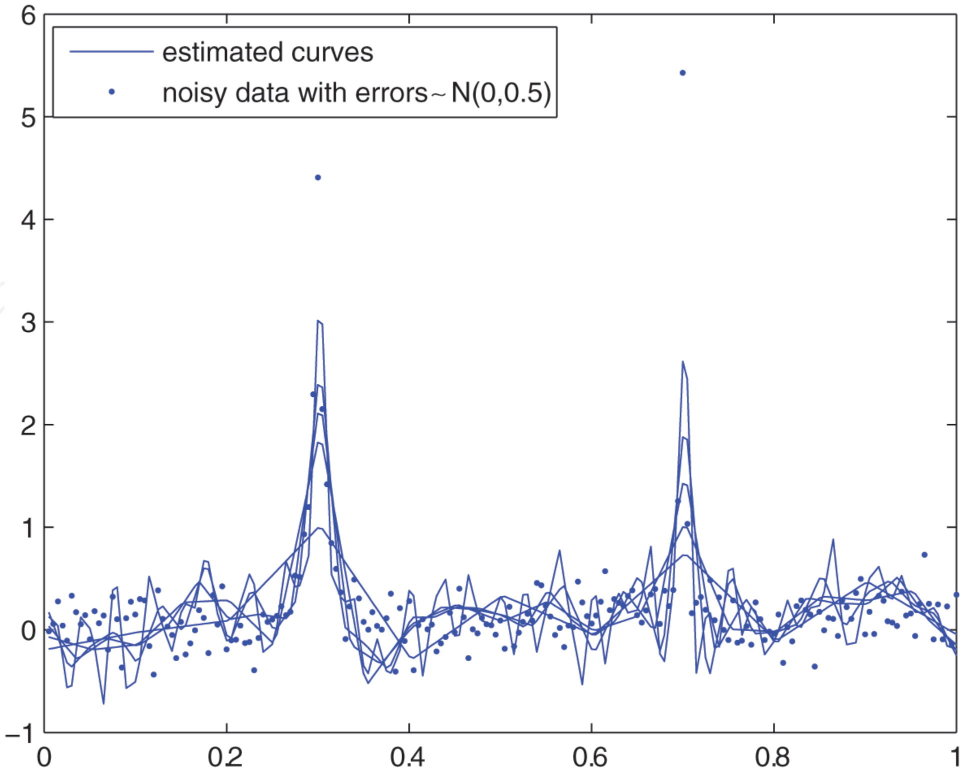

Figure 1. 200 observations and the estimated curves based on different knot sets.

different values of m are plotted in **Figure 1** too. The simulated SiZerLS map and SiZerSS map are shown in **Figure 2**, respectively.

In **Figure 2**, BYP SiZerLS is SiZerLS map based on multiple testing procedures, BYP, where BYP denotes the multiple testing procedure proposed in Benjamini and Yekutieli [43]. SiZerSS is the smoothing spline version of SiZer. The two SiZers are simulated under the same range of scales and nominal level 0.05. There are four colors in SiZer maps: red indicates that the

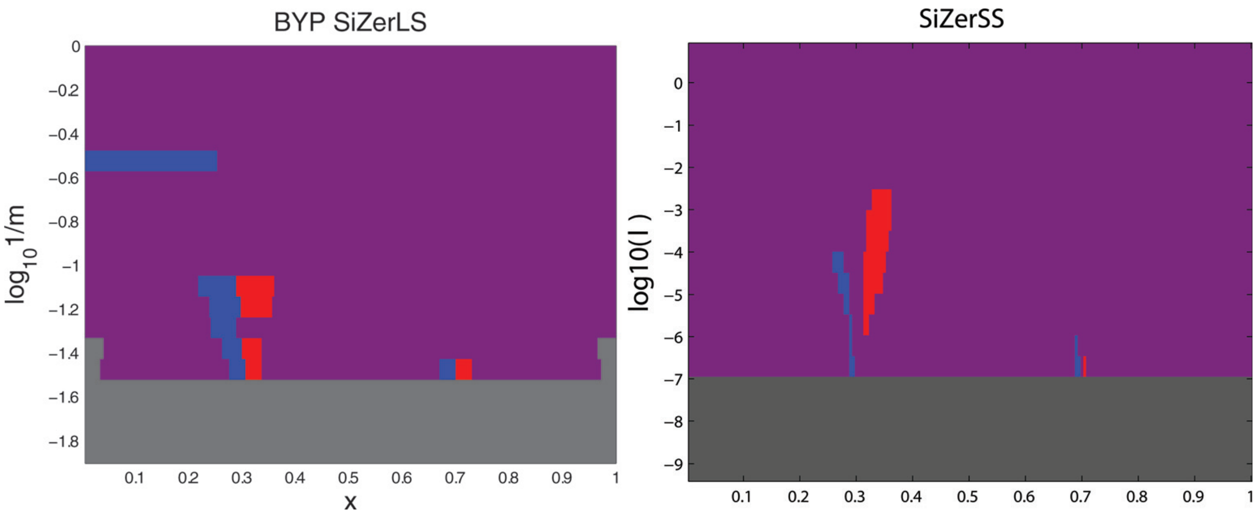

Figure 2. BYP SiZerLS and SiZerSS for detecting peaks of data.

estimated regression curve is significantly decreasing; blue indicates that the estimated regression curve is significantly increasing; purple indicates that the curve is neither significantly increasing nor decreasing; gray shows that there are no sufficient data for conducting reasonable statistical inference. **Figure 1** preliminarily shows that SiZer maps can locate peaks well. The theoretical foundation of SiZerLS and SiZerSS will be discussed in more detail at a later stage.

3.1. Construction of SiZerLS map for exploring features of regression curve

The proposed SiZerLS map will be constructed on the basis of the p-values with multiple testing adjustment. The p-value for testing the monotonicity of the smoothed curve is defined first based on linear spline approximation and fiducial method in the same way as p-values in Section 2. Consequently, multiple testing adjustment is discussed detailedly to control the row-wise false discovery rate (FDR) of SiZerLS.

In the view of SiZer, all of the useful information is included in the smoothed curve, which is defined below. Suppose we have observations $\{x_i, y_i\}_{i=1}^{n}$ from regression model (1). By linear spline estimation, estimator $\widehat{f}_m(x)$ can be obtained:

$$\widehat{f}_m(x) = \boldsymbol{g}(x)'\left(G^T G\right)^{-1} G^T\, Y, \tag{22}$$

where $\boldsymbol{g}(x) = \left(g_1(x), g_2(x), \cdots, g_m(x)\right)'$; $g_j(x), j = 1, \cdots, m$ are the basis functions defined in (7) on the basis of m knots; and G is the matrix defined in Section 2. The smoothed curve at smoothing level m is denoted as.

$$f_m(x) = \mathrm{E}\left(\widehat{f}_m(x)\right) = \boldsymbol{g}(x)'\left(G^T G\right)^{-1} G^T\, \boldsymbol{f},$$

where $\boldsymbol{f} = \{f(x_1), f(x_2), \cdots, f(x_n)\}'$. SiZer focuses on $f_m(x)$. Its monotonicity is determined totally by $\left(G^T G\right)^{-1} G^T\, \boldsymbol{f}$. Hence, it is enough to test the following $m - 1$ pairs of null hypotheses:

$$H_{Ik} = f_m(t_k) = e_k'(G'G)^{-1}G'\boldsymbol{f} \le e_{k+1}'(G'G)^{-1}G'\boldsymbol{f} = f_m(t_{k+1})\,(\text{and})$$

$$H_{Dk} = f_m(t_k) = e_k'(G'G)^{-1}G'\boldsymbol{f} \ge e_{k+1}'(G'G)^{-1}G'\boldsymbol{f} = f_m(t_{k+1}), k = 1, 2, \cdots, m - 1, \tag{23}$$

where e_k is an m-dimensional column vector having 1 in the kth entry and zero elsewhere. Let b denote $(G'G)^{-1}G'\boldsymbol{f}$. Then, H_{Ik} and H_{Dk} can be written as

$$H_{Ik}^* = L_k b \le 0, k = 1, 2, \cdots, m - 1; \quad H_{Dk}^* = L_k b \ge 0, k = 1, 2, \cdots, m - 1, \tag{24}$$

where $L_k \triangleq \left(e_k' - e_{k+1}'\right)$. The p-values to test hypotheses in (24) under linear model $Y = Gb + \varepsilon$ can be defined using pivotal quantity about b. This pivotal quantity is $R\left(E; \widehat{\beta}, S^2\right)$, which is defined in (11). The p-value for testing H_{Ik}^* is the fiducial probability that null hypothesis holds:

$$P^*_{Ik}\left(\widehat{\beta}, S\right) = P\left\{L_k R\left(E; \widehat{\beta}, S\right) \le 0\right\} = P\left\{L_k\widehat{\beta} - \frac{S}{E_2}(G'G)^{-\frac{1}{2}}E_1 \le 0\right\}$$

$$= P\left\{\frac{\sqrt{n-m}L_k(G'G)^{-1}G'E_1}{\left(L_k(G'G)^{-1}L'_k\right)^{\frac{1}{2}}E_2} \ge \frac{\sqrt{n-m}\widehat{\beta}}{S\left(L_k(G'G)^{-1}L'_k\right)^{\frac{1}{2}}}\right\}, \tag{25}$$

where the subscript Ik of P^*_{Ik} represents the interval (t_k, t_{k+1}) in which we test monotonicity and m represents the number of knots used in linear interpolation. In addition, p-value $P^*_{Dk}\left(\widehat{\beta}, S\right)$ for testing H^*_{Dk} satisfies equation $P^*_{Ik}\left(\widehat{\beta}, S\right) + P^*_{Dk}\left(\widehat{\beta}, S\right) = 1$.

It is worth noting that p-value $P^*_{Ik}\left(\widehat{\beta}, S\right)$ is uniformly distributed on (0,1) if all of the hypotheses $H_{Ik}, H_{Dk}, k = 1, 2, \cdots, m-1$ are true (regression function is a constant). In applications, p-value $P^*_{Ik}\left(\widehat{\beta}, S\right)$ for testing H_{Ik} can be approximated as below when $n \to \infty$. This approximation is reasonable (see Theorem 1 in [44]):

$$P_{Ik,m}\left(\widehat{\beta}, S\right) \triangleq 1 - \Phi\left(\frac{\sqrt{n-m}L_k\widehat{\beta}}{S\left(L_k(G'G)^{-1}L_k\right)^{1/2}}\right). \tag{26}$$

The proposed SiZerLS map will be constructed on the basis of the above p-values with multiple testing adjustment. In fact, SiZer is a visual method for exploratory data analysis, and it focuses on exploring features that really exist in data instead of testing whether some assumed features are statistically significant in a strict way. FDR is the expected proportion of the false positives among all discoveries, and FDR can be either permissive or conservative according to the number of hypotheses. Considering that different numbers of hypotheses need to be tested for SiZerLS with respect to various smoothing parameters, the multiple testing adjustment to control FDR would be better if used to improve the exploratory property of SiZer. Hence, the well-known multiple testing procedure which was proposed in Benjamini and Yekutieli [43] (denoted as BYP) is applied to control the row-wise FDR of SiZerLS. The BYP was proved to control FDR under α for any dependent test statistics.

3.1.1. Benjamin-Yekutieli procedure to control FDR (BYP)

Suppose that we have obtained p-values $P_{Ik,m}\left(\widehat{\beta}, S\right)$ for testing hypotheses H_{Ik} in (23), $k = 1, 2, \cdots, m-1$:

1. Order p-values $P^*_{Ik,m}$ and get the ordered p-values $P_{I(1),m}, P_{I(2),m}, \cdots, P_{I(m-1),m}$.
2. For a given p-value α, find the largest i for $k = 1, 2, \cdots, m-1$ for which $P_{I(i),m} \le \frac{k\alpha}{(m-1)\sum_{j=1}^{m-1}\frac{1}{j}}$ and reject all $H_{I(k),m}$ for $k = 1, 2, \cdots, m-1$.

The detailed steps to construct SiZerLS with BYP adjustment are given below:

Step 1. Construct 2D grid map. Without loss of generality, we assume that designed points $x_i, i = 1, 2, \cdots, n$ are chosen from [0, 1]. Then the 2D map is a rectangular area [0, 1; $\log_{10}(1/m_{\max},), \log_{10}(1/m_{\min})$]; see BYP SiZerLS displayed in **Figure 2**. The value of m is determined by the following rule: $m = \text{round}\left(1/10^l\right)$, where function round $(\cdot)$ is the nearest integer function and l takes equally spaced values from interval $\left[\log_{10}(1/m_{\min}), \log_{10}(1/m_{\max})\right]$. For a given m, abscissa x takes values at the corresponding knots $\mathbf{T}_m = \{t_1, t_2, \cdots, t_m\}$. On the basis of different values of m and $\mathbf{T}_m$, the 2D map is divided into many pixels.

Step 2. Calculate p-values for each pixel. Each pixel in the 2D map constructed in step 1 is determined by two adjacent knots and a determined m. For pixel $(t_k, t_{k+1}, m = m_0)$, we calculate p-value P_{Ik,m_0} and P_{Dk,m_0} for testing hypotheses H^*_{Ik,m_0} and H^*_{Dk,m_0}, respectively, with m_0 knots.

Step 3. Multiple testing adjustment. For a given value $m = m_0$, carry out multiple testing procedure BYP using p-values P_{Ik,m_0} (P_{Dk,m_0}), $k = 1, 2, \cdots, m_0$, obtained from step 2 to test the fowling family of hypotheses simultaneously:

$$\left\{H^*_{I1,m_0}, H^*_{I2,m_0}, \cdots, H^*_{Im_0-1,m_0}\right\}\left(H^*_{D1,m_0}, H^*_{D2,m_0}, \cdots, H^*_{Dm_0-1,m_0}\right).$$

Step 4. Color pixels. According to the multiple testing results at smoothing level m_0 if H^*_{Ik} is rejected and H^*_{Dk} is accepted, pixel $(t_k, t_{k+1}, m = m_0)$ is colored red to indicate significant decreasing. On the contrary, if $H^*_{Ik,,m_0}$ is accepted and H^*_{Dk,m_0} rejected, pixel $(t_k, t_{k+1}, m = m_0)$ is colored blue to show significant increasing; purple is used for no significant trend in other cases.

In SiZer map, gray indicates that no sufficient data can be used to test the monotonicity of regression function at point x with m knots. Such sufficiency is quantified as effective sample size (ESS). Noting that the number of nonzero elements in the kth column of G has a demonstrable effect on the inference in interval (t_k, t_{k+1}), and it is determined directly by how many observations are included in (t_k, t_{k+1}), we define $ESS(t_k, m)$ as.

$$(ESS(t_1, m), ESS(t_2, m), \cdots, ESS(t_m, m))' \triangleq G'G(1, 1, \cdots, 1)'.$$

In SiZerLS map, pixel $(t_k, t_{k+1}, m = m_0)$ would be colored gray if.

$$\min(ESS(t_k, m_0), ESS(t_{k+1}, m_0)) < 5.$$

In order to avoid selecting knots, m equally spaced knots or equal x-quantiles are used in interpolation. The smoothing level of regression spline estimate is controlled by m together with the positions of knots. The level of smoothness should be reduced to detect some local fine feature; however, the total number of knots should be limited to avoid excessive under-smoothing in a wide range. In applications of SiZerLS, the range of scales is recommended to include the coarsest smoothing level, $m = 2$, and the finest smoothing level, $\text{avg}_{x \in \mathbf{T}_{m\max}} ESS(x, m_{\max}) < 5$.

3.2. Construction of SiZerSS map for exploring features of regression curve

SiZerSS given in Marron and Zhang [33] employed smoothing spline to construct SiZer map for nonparametric model (1). Given $\{x_i, y_i\}_{i=1}^n$ and a smoothing parameter λ, the smoothing spline estimator is the function $\widehat{f}_\lambda$ that minimizes the regularization criterion over function f:

$$\sum_{i=1}^{n} \omega_i \left[y_i - f(x_i)\right]^2 + \lambda \int \left[f''(x)\right]^2 \mathrm{d}x. \tag{27}$$

By simple calculation, we can get the estimator vector:

$$\widehat{\mathbf{f}}_\lambda = \left(\widehat{f}_\lambda(x_1), \widehat{f}_\lambda(x_2), \cdots, \widehat{f}_\lambda(x_n)\right) = (W + \lambda K)^{-1} W Y = A_\lambda Y, \tag{28}$$

where weight matrix $W = \mathrm{diag}(\omega_1, \omega_2, \cdots, \omega_n)$ and the hat matrix $A_\lambda = (W + \lambda K)^{-1} W$.

In order to construct SiZerSS, the derivative of f at any point x needs to be estimated along with its variance. Let $s_i = x_{i+1} - x_i$ and $n \times (n-1)$ matrix $Q = \left\{q_{ij}\right\}, i = 1, 2, \cdots, n$, $j = 2, \cdots, n-1$, where $q_{j-1,j} = s_{j-1}^{-1}, q_{jj} = -s_{j-1}^{-1} - s_j^{-1}, q_{j+1,j} = s_j^{-1}$, and $q_{i,j} = 0$ for $|i-j| \geq 2$. Let $\left(\gamma_1, \gamma_2, \cdots, \gamma_n\right) = \left(f''(x_1), f''(x_2) \cdots, f''(x_n)\right)$. By the definition of natural cubic spline, $f''(x_1) = f''(x_n) = 0$. Let $\gamma = \left(\gamma_2, \cdots, \gamma_{n-1}\right)'$. According to Theorem 2.1 of Green and Silverman [45], the vectors $\mathbf{f}$ and γ specify a natural cubic spline f if and only if $Q'\mathbf{f} = R\gamma$,

where R is a $(n-2) \times (n-2)$ symmetric matrix with elements $r_{ij}, i = 2, \cdots, n-1$, $j = 2, \cdots, n-1$, which is given by $r_{ii} = \frac{1}{3}(s_{i-1} + s_i), r_{i,i+1} = r_{i+1,i} = \frac{1}{6} s_i$ and $r_{ij} = 0$ for $|i-j| \geq 2$. The estimator $\widehat{\gamma}$ can be obtained from equation $(R + \lambda Q'Q)\gamma = Q'Y$. Then estimator $\widehat{f}(x)$ and $\widehat{f}'(x)$ can be written as a linear combination of $\widehat{\mathbf{f}}$ and $\widehat{\gamma}$. Let $h_i(x) = x - x_i, i = 1, 2, \cdots, n$. When $x < x_1$.

$$\widehat{f}_\lambda(x) = \widehat{f}_\lambda(x_1) + h_1(x) \left\{ \frac{\widehat{f}_\lambda(x_2) - \widehat{f}_\lambda(x_1)}{s_1} - \frac{s_1}{6} \widehat{\gamma}_2 \right\}, \widehat{f}'(x) = \frac{\widehat{f}_\lambda(x_2) - \widehat{f}_\lambda(x_1)}{s_1} - \frac{s_1}{6} \widehat{\gamma}_2.$$

When $x_i \leq x \leq x_{i+1}$, let $\delta_i(x) = \left(1 + \frac{h_i(x)}{s_i}\right) \widehat{\gamma}_{i+1} + \left(1 - \frac{h_{i+1}(x)}{h_i}\right) \widehat{\gamma}_i$ for $i = 1, 2, \cdots, n$,

$$\widehat{f}_\lambda(x) = \frac{h_i(x) \widehat{f}_\lambda(x_{i+1}) - h_{i+1}(x) \widehat{f}_\lambda(x_i)}{s_i} + \frac{h_i(x) h_{i+1}(x) \delta_i(x)}{6},$$

$$\widehat{f}'_\lambda(x) = \frac{\widehat{f}_\lambda(x_{i+1}) - \widehat{f}_\lambda(x_i)}{s_i} + \frac{h_i(x) h_{i+1}(x) (\widehat{\gamma}_{i+1} - \widehat{\gamma}_i)}{6 s_i} + \frac{h_i(x) + h_{i+1}(x)}{6} \delta_i(x).$$

(When) $x > x_n$

$$\widehat{f}_\lambda(x) = \widehat{f}_\lambda(x_n) + \frac{h_n(x)}{6}\left\{\frac{\widehat{f}_\lambda(x_n) - \widehat{f}_\lambda(x_{n-1})}{s_{n-1}} + s_{n-1}\widehat{\gamma}_{n-1}\right\},$$

$$\widehat{f}'_\lambda(x) = \frac{1}{6}\left\{\frac{\widehat{f}_\lambda(x_n) - \widehat{f}_\lambda(x_{n-1})}{s_{n-1}} + s_{n-1}\widehat{\gamma}_{n-1}\right\}.$$

The variance of $\widehat{f}'_\lambda(x)$ can be calculated easily if the estimator of σ^2, the variance of the error in model (1), is obtained. σ^2 can be estimated by the sum of squared residuals $\sum\left(y_i - \widehat{f}_\lambda(x_i)\right)^2$. If σ^2 is a function of x, $\sigma^2(x)$ can be estimated by $\left(y_i - \widehat{f}_\lambda(x)\right)^2$. The confidence interval of $f_\lambda{}'(x)$ are of the form:

$$\widehat{f}'_\lambda(x) \pm q.\widehat{\mathrm{SD}}\left(\widehat{f}_\lambda{}'(x)\right), \tag{29}$$

where q is based on the nominal level. For details, see Section 3 of Chaudhuri and Marron [25].

SiZerSS can be constructed as SiZerLS. For different values of x, if interval (29) contains zero, pixel (x, λ) is colored purple; if confidence interval is on the right side of zero, blue is used to indicate increasing; otherwise, red is used to imply decreasing. Gray is used to indicate that there is no sufficient data to do reliable inference. The sufficiency can be found in Chaudhuri and Marron [25].

The simulated SiZerLS and SiZerSS maps are displayed in **Figure 2**, where the red and blue regions locate the bumps of regression curve accurately. This simulation illustrates the good behavior of SiZerLS and SiZerSS in exploring features in data.

3.3. Construction of SiZer-RS map for comparing multiple regression curves

The comparison of two or more populations is a common problem and is of great practical interest in statistics. In this subsection, comparison of multiple regression curves in a general regression setting is developed based on regression spline. Suppose we have $n = \sum_{i=1}^{k} n_i$ independent observations from the following k regression models:

$$y_{ij} = f_i(x_{ij}) + \sigma_i(x_{ij})\varepsilon_{ij}, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n_i, \tag{30}$$

where x_{ij} s are covariates, the errors $\varepsilon_{ij} \sim N(0, 1)$ s are independent and identically distributed errors, $f_i(\cdot)$ is the regression function, and $\sigma_i^2(\cdot)$ is the conditional variance function of the ith population. We are concerned about whether the k populations in model (30) are equal; if not, what is the difference? To this end, a multi-scale method, SiZer-RS, based on regression spline is proposed to compare $f_i(\cdot)$ across multiple scales and locations.

As described in Park and Kang [32], the choice of smoothing parameter is also important for comparing regression curves. They developed SiZer for the comparison of regression curves based on local linear smoother. SiZer map for comparing regression curves is a 2D color map, which consists of a large number of pixels. Each pixel is indexed by a scale (smoothing parameter) and a location; the color of a pixel indicates the result for testing the equality of two or more multiple regression curves at the corresponding location and scale. SiZer provides us with more information about the locations of the differences among the regression curves if they do exist. Park et al. [46] developed an ANOVA-type test statistic and conducted it in scale space for testing the equality of more than two regression curves.

The works mentioned above are kernel-based method. Besides it, regression spline is an important smoothing device and is used widely in applications. For a given smoothing parameter m (the number of knots used in regression spline), the p-value for testing the equality of k regression curves at point x is established. Consequently, SiZer-RS is constructed in the same way as SiZerLS for comparing multiple retrogression curves based on higher-order spline interpolation.

For a given smoothing parameter m (the number of knots used in regression spline), the smoothed curve is defined as $f_{i,m}(x) = \mathrm{E}(\widehat{f}_{i,m}(x))$, where $\widehat{f}_{i,m}(x)$ is the regression spline estimator. SiZer-RS for comparing multiple regression curves is based on the testing results for testing null hypothesis:

$$H_{m,x} : f_{1,m}(x) = f_{2,m}(x) = \cdots = f_{k,m}(x), \tag{31}$$

at point x with smoothing parameter m. Without loss of generality, we still suppose that the explanatory variable x takes value from [0, 1]. On the basis of a knot set $\mathbf{T}_m = \{0 = t_1 < t_2, \cdots, < t_m = 1\}$, we have the approximation:

$$f_i(x) \approx \sum_{s=1}^{m+q-1} \beta_{i,s} g_{m,s}(x) \triangleq N_m(x)' \beta_i^m, \tag{32}$$

where $\beta_i^m = \left(\beta_{i,1}, \beta_{i,2}, \cdots, \beta_{i,m+p-1}\right)'$. The estimator of $f_i(x)$ at smoothing level m can be obtained $\widehat{f}_{i,m}(x) = N_m(x)' \widehat{\beta}_i^m$, in which, $N_m(x) = \left\{g_{m,s}(x), s = 1, 2, \cdots, m+q-1\right\}$. If $q = 3$, $N_m(x)'$ is defined below:

$$N_m^l(x) = (t_l - t_{l-4})[t_{l-4}, t_{l-3}, t_{l-2}, t_{l-1}, t_l](t-x)_+^3, \quad l = 2, 3, \cdots, m+3,$$

where $t_l = t_{\min(\max(l,1),m)}$ for $l = -2, -1, \cdots, m+3$:

$$(t-x)_+^3 = \begin{cases} (t-x)^3, t > x \\ 0, \qquad t \le x \end{cases}.$$

For a function $g(\cdot)$, $[t_{l-4}, t_{l-3}, t_{l-2}, t_{l-1}, t_l]g(\cdot)$ denotes the fourth-order divided difference of $g(\cdot)$, that is:

$$\begin{cases} [t_1, t_2]g = g'(t), \quad \text{if} \quad t_1 = t_2 = t \\ [t_1, t_2]g = \dfrac{g(t_2) - g(t_1)}{t_2 - t_1} \text{ otherwise,} \\ [t_1, t_2, \cdots, t_k]g = g^{(k-1)}(t), \textit{if } t_1 = \cdots = t_k \\ [t_1, t_2, \cdots, t_k]g = \dfrac{[t_2, t_3, \cdots, t_k]g - [t_1, t_2, \cdots, t_{k-1}]g}{t_k - t_1}, \text{otherwise.} \end{cases}$$

Then model (31) can be approximately written as the following linear regression model:

$$Y_i = G_i^m \beta_i^m + \Sigma_i E_i, \tag{33}$$

where

$$Y_i = \left(y_{i1}, y_{i2}, \cdots, y_{in_i}\right)', G_i^m = \left(N_m^l(x_i)\right)_{n_i \times (m+2)'} \ \Sigma_i = \text{diag}\{\sigma_i(x_{ij})\}, E_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{in_i}\right)'.$$

At first, we suppose Σ_i is known and then replace it by its available estimator.

From regression model (33), we can get the estimator $\widehat{\beta}_i^m = \left(G_i^{m\prime}\Sigma_i^{-1}G_i^m\right)^{-1} G_i^{m\prime}\Sigma_i^{-1}Y_i$. Let b_i^m denote the expectation of $\widehat{\beta}_i^m$:

$$\mathrm{b}_i^m = \mathrm{E}\left(\widehat{\beta}_i^m\right) = \left(G_i^{m\prime}\Sigma_i^{-1}G_i^m\right)^{-1} G_i^{m\prime}\Sigma_i^{-1}\mathbf{f}_i,$$

where $\mathbf{f}_i = \left(f_i(x_{i1}), \cdots, f_i(x_{in_i})\right)'$. Therefore, the smoothed curve

$$f_{i,m}(x) = \mathrm{E}\left(\widehat{f}_{i,m}(x)\right) = \mathrm{E}\left[N_m(x)'\left(G_i^{m\prime}\Sigma_i^{-1}G_i^m\right)^{-1} G_i^{m\prime}\Sigma_i^{-1}Y_i\right] = N_m(x)'\mathrm{b}_i^m. \tag{34}$$

Denote $\mathbf{b}^m = \left(\mathrm{b}_1^{m\prime}, \mathrm{b}_2^{m\prime}, \cdots, \mathrm{b}_k^{m\prime}\right)'$, and correspondingly, denote its estimator as $\widehat{\beta}^m = \left(\beta_1^{m\prime}, \beta_2^{m\prime}, \cdots \beta_k^{m\prime}\right)'$. Hypothesis $H_{m,x}$ can be presented as

$$H_{m,x}: \quad L_m(x)\mathbf{b}^m = 0_{k-1}, \tag{35}$$

where

$$L_m(x) = \begin{bmatrix} N_m(x) & N_m(x) & N_m(x) & & N_m(x) \\ -N_m(x) & 0 & 0 & \cdots & 0 \\ 0 & -N_m(x) & 0 & & 0 \\ \cdots & \cdots & \cdots & & \cdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & -N_m(x) \end{bmatrix}$$

is a $(k-1) \times k(m+q-1)$ matrix.

http://dx.doi.org/10.5772/intechopen.74858

The p-value for testing hypothesis $H_{m,x}$ in (35) can be defined as

$$p_{m,x}\left(\widehat{\beta}_i^m, \widehat{\Sigma}_m\right) = P\left\{T^m(x)' L_m(x)' \left[L_m(x)\widehat{\Sigma}_m L_m(x)'\right]^{-1} L_m(x) T^m(x) \right.$$
$$\left. \geq \widehat{\beta}_i^{m\prime} L_m(x)' \left[L_m(x)\widehat{\Sigma}_m L_m(x)'\right]^{-1} L_m(x)\widehat{\beta}_i^m \right\}, \tag{36}$$

where $T^m(x) \triangleq \left\{ \left(G_i^{m\prime}\widehat{\Sigma}_{i,m}^{-1} G_i^m\right)^{-1} G_i^{m\prime}\widehat{\Sigma}_{i,m}^{-\frac{1}{2}} E_i \right\}', i = 1, 2, \cdots, k; \widehat{\Sigma}_{i,m} = \mathrm{diag}\{\widehat{\sigma}_i(x_{ij}), j = 1, 2, \cdots, n_i\}$ is an estimator of the variance matrix of the ith regression model and

$$\widehat{\Sigma}_m = \mathrm{diag}\left\{ \left(G_i^{m\prime}\widehat{\Sigma}_{i,m}^{-1} G_i^m\right)^{-1}, i = 1, 2, \cdots, k \right\}$$

is an estimator of the variance matrix of $T^m(x)$ given $\widehat{\beta}_i^m, \widehat{\sigma}_i^{m2}, i = 1, 2, \cdots, k$. The estimator of $\sigma_i(x_{ij})$ can be found in Li and Xu [36], where the smoothing parameter, m_p, can be used as a pilot smoothing parameter, which is different from m used in $\widehat{f}_{i,m}(x)$. SiZer-RS map can be constructed based on different values of m_p, which represents the different trade-offs between the structure of regression curve and errors.

The two SiZer maps given in **Figure 4** are constructed using the data plotted in **Figure 3** to compare three regression curves $f_1(x) = f_x(x) = 0, f_3(x) = 0.5\sin(2\pi x)$. Since the variance of errors is a constant, it can be estimated by the sum of squares of residues. In this case, pilot smoothing parameter is avoided [47, 48]. The two blue regions in **Figure 4** clearly show their difference across interval (0, 1). The gray color indicates that there is no sufficient data that can be used to get credible testing results at x and nearby. The sufficiency is quantized as $ESS(x, m)$ for SiZer-RS, and pixel (x, m) is colored gray if $ESS(x, m) < 5$:

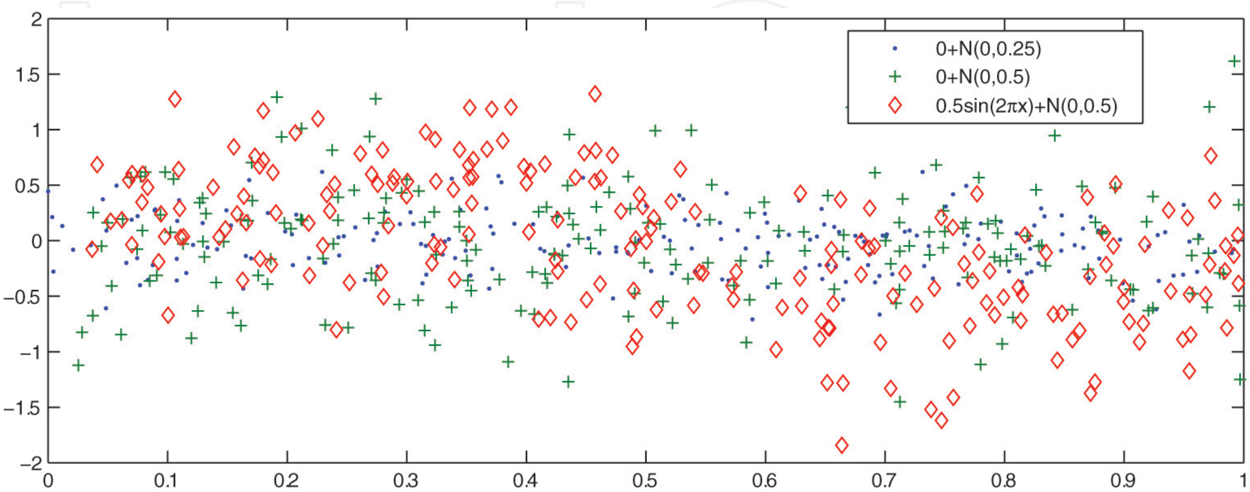

Figure 3. 200 observations.

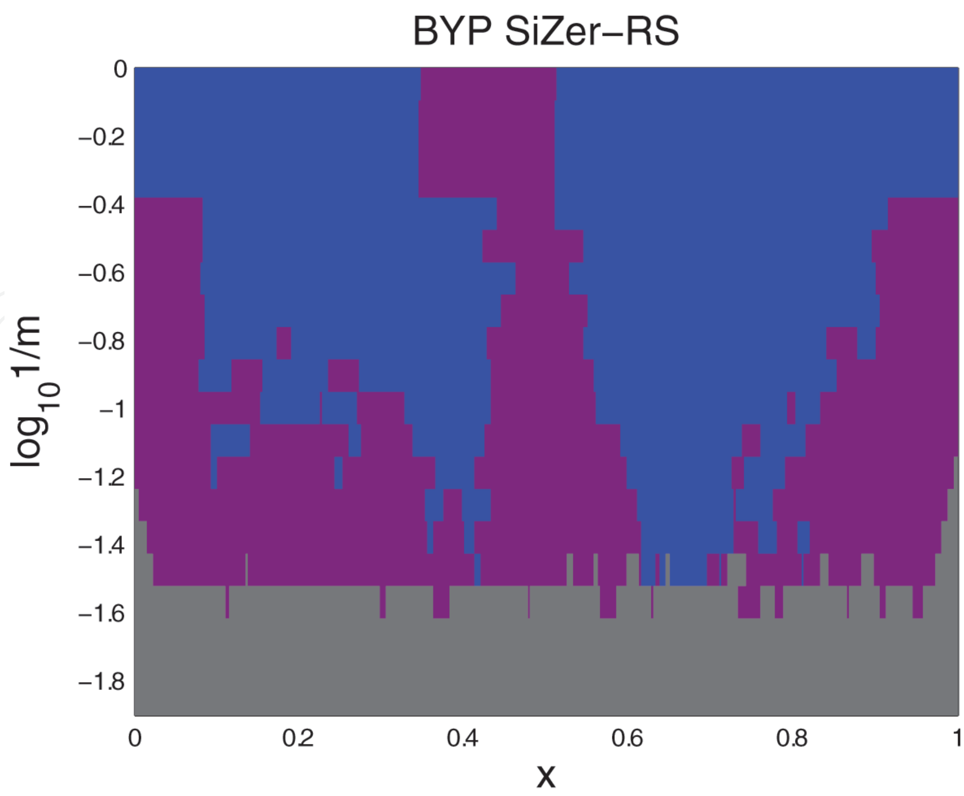

Figure 4. SiZer-RS map.

$$ESS(x, m) \triangleq \min_{i=1,2,\cdots,k} \{N_m(x) G_i^{m\prime} G_i^m (1, 1, \cdots, 1)'\}.$$

Figure 4 shows that SiZer-RS map can explore the differences between regression curves accurately.

It is worth noting that, for SiZer-RS map, the coarsest smoothing level should be $m = q + 1$ to ensure the effectiveness of the qth regression spline and the finest smoothing level is recommend to be the one such that $\mathrm{avg}_{x \in [x_1, x_2, \cdots, x_g]} ESS(x, m) < 5$, where $x_1, x_2, \cdots, x_g$ are points at which hypothesis $H_{m,x}$ is tested and pixels are produced by combing different values of m. In applications, a wide range of values of m_p can be used to generate a family of SiZer-RS maps. Particularly, m_p and m can both be used as smoothing parameters simultaneously to construct a 3D SiZer-RS map [47, 48].

4. Conclusion

This chapter introduces regression spline method for testing the parametric form of nonparametric regression function in nonparametric, partial linear, and varying-coefficient models, respectively. The corresponded p-values are established based on fiducial method and spline interpolation. The test procedures on the basis of the proposed p-value are accurate in some cases and are consistent under some mild conditions, which means that the p-value tends to be zero when null hypothesis is false as sample size and the number of knots used in spline interpolation tend to be infinity. Hence, the proposed test procedures are performed well especially in small sample size case.

The spline-based method frequently used smoothing method, which can be used easily with other statistical methods. When using the spline-based method, the smoothing level is controlled by the number of knots and their positions. In order to sidestep the determination of knots and obtain more reliable results, multi-scale smoothing methods are proposed based on spline regression to infer structures of regression function. The multi-scale method is a visual method to do inference at different locations and smoothing levels. In addition, the smoothing spline version of multi-scale method is also introduced. The proposed multi-scale method can also be used for comparing multiple regression curves. Some real data examples illustrate the practicability of the proposed multi-scale method.

The MATLAB code of SiZerLL and other versions of SiZer based on kernel smoother is available from the homepage of Professor Marron JS; the MATLAB code of SiZerLS can be downloaded from the following website:

http://www.tandfonline.com/doi/suppl/10.1080/10618600.2014.1001069?scroll=top.

Author details

Na Li

Address all correspondence to: nali@amss.ac.cn

School of Econometrics and Management, University of the Chinese Academy of Sciences, Beijing, China

References

[1] Härdle W, Mammen E. Comparing nonparametric regression fits. Annals of Statistics. 1993;**21**:1926-1947

[2] Hart JD. Nonparametric Smoothing and Lack-of-Fit Test. New York: Springer; 1997

[3] Cox D, Koh E, Wahba G, Yandell BS. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. Annals of Statistics. 1988;**16**:113-119

[4] Cox D, Koh EA. Smoothing spline based test of model adequacy in polynomial regression. Annals of the Institute of Statistical Mathematics. 1989;**41**:383-400

[5] Fan J, Zhang C, Zhang J. Generalized likelihood ratio statistics and Wilks phenomenon. Annals of Statistics. 2001;**29**:153-193

[6] Eubank RL, Hart JD. Testing goodness-of-fit in regression via order selection criteria. Annals of Statistics. 1992;**20**:1412-1425

[7] Baraud Y. Model selection for regression on a fixed design. Probability Theory and Related Fields. 2000;**117**:467-493

[8] Bianco A, Boente G. Robust estimators in semi-parametric partly linear regression models. Journal of Statistical Planning and Inference. 2004;**22**:229-252

[9] Liang H, Wang S, Robin JM, Carroll RJ. Estimation in partially linear models with missing covariates. Journal of the American Mathematical Society. 2004;**99**:357-367

[10] Fan J, Huang L. Goodness-of-fit tests for parametric regression models. Journal of the American Mathematical Society. 2001;**96**:640-652

[11] Baraud Y, Huet S, Laurent B. Adaptive tests of linear hypotheses by model selection. Annals of Statistics. 2003;**31**:225-251

[12] Claeskens G. Restricted likelihood ratio lack-of-fit tests using mixed spline models. Journal of the Royal Statistical Society: Series B. 2004;**66**:909-926

[13] Härdel W, Sperlich S, Spokoiny V. Structural tests in additive regression. Journal of the American Mathematical Society. 2001;**96**:1333-1347

[14] Meyer MC. A test for linear versus convex regression function using shape-restricted regression. Biometrika. 2003;**90**:223-232

[15] Stute W. Nonparametric model checks for regression. Annals of Statistics. 1997;**5**:613-641

[16] Liang H. Checking linearity of non-parametric component in partially linear models with an application in systemic inflammatory response syndrome study. Statistical Methods in Medical Research. 2006;**15**:273-284

[17] Fan J, Zhang W. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. Scandinavian Journal of Statistics. 2000;**27**:715-731

[18] Cai Z, Fan J, Li R. Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association. 2000;**95**:888-902

[19] Fan J, Huang T. Profile likelihood inferences on semi-parametric varying-coefficient partially linear models. Bernoulli. 2005;**11**:1031-1057

[20] You J, Zhou Y. Empirical likelihood for semi-parametric varying-coefficient partially linear regression models. Statistics and Probability Letters. 2006;**76**:412-422

[21] Tang Q, Cheng L. M-estimation and B-spline approximation for varying coefficient models with longitudinal data. Journal of Nonparametric Statistics. 2008;**20**:611-625

[22] Hoover DR, Rice JA, Wu C, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998;**85**:809-822

[23] Wu C, Yu K, Chiang CT. A two-step smoothing method for varying coefficient models with repeated measurements. Annals of Statistics. 2000;**52**:519-543

[24] Xu X, Li G. Fiducial inference in the pivotal family of distributions. Science in China, Series A: Mathematics. 2006;**49**:410-432

[25] Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. Journal of the American Statistical Association. 1999;**94**:807-823

[26] Chaudhuri P, Marron JS. Scale space view of curve estimation. Annals of Statistics. 2000;**28**:408-428

[27] Li R, Marron JS. Local likelihood SiZer map. The Indian Journal of Statistics. 2005;**67**:476-498

[28] Hannig J, Lee TCM. Robust SiZer for exploration of regression structures and outlier detection. Journal of Computational and Graphical Statistics. 2006;**15**:101-117

[29] Park C, Lee TCM, Hannig J. Multiscale exploratory analysis of regression quantiles using quantile SiZer. Journal of Computational and Graphical Statistics. 2010;**3**:497-513

[30] Marron JS, deŨnáAlvarez J. SiZer for length biased, censored density and hazard estimation. Journal of Statistical Planning and Inference. 2004;**121**:149-161

[31] Kim CS, Marron JS. SiZer for jump detection. Journal of Nonparametric Statistics. 2006;**18**:13-20

[32] Park C, Kang KH. SiZer analysis for the comparison of regression curves. Computational Statistics and Data Analysis. 2008;**52**:3954-3970

[33] Marron JS, Zhang JT. SiZer for smoothing spline. Computational Statistics. 2005;**20**:481-502

[34] Dawid AP, Stone M. The functional model basis of fiducial inference. Annals of Statistics. 1982;**10**:1054-1067

[35] Hannig J, Iyer H, Patterson P. Fiducial generalized confidence intervals. Journal of the American Mathematical Society. 2006;**101**:254-269

[36] Li N, Xu X, Jin P. Testing the linearity in partially linear models. Journal of Nonparametric Statistics. 2011;**23**:99-114

[37] Li N, Xu X, Liu X. Testing the constancy in varying-coefficient regression models. Metrika. 2011;**74**:409-438

[38] Shi P, Li G. Global convergence rates of B-spline M-estimators in nonparametric regression. Statistica Sinica. 1995;**5**:303-318

[39] Mao W, Zhao LH. Free-knot polynomial splines with confidence intervals. Journal of the Royal Statistical Society: Series B. 2003;**65**:901-909

[40] DiMatteo I, Genovese CR, Kass RE. Bayesian curve-fitting with free-knot splines. Biometrika. 2001;**88**:1055-1071

[41] Sonderegger DL, Hannig J. Fiducial Theory for Free-Knot Splines. Vol. 68. New York: Springer International Publishing; 2014. pp. 155-189

[42] Lindstrom MJ. Penalized estimation of free-knot splines. Journal of Computational and Graphical Statistics. 1999;**8**:333-352

[43] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics. 2001;**29**:1165-1188

[44] Li N, Xu X. Spline multiscale smoothing to control FDR for exploring features of regression curves. Journal of Computational and Graphical Statistics. 2016;**25**:325-343

[45] Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models. London: Chapman and Hall; 1994

[46] Park C, Hannig J, Kang KH. Nonparametric comparison of multiple regression curves in scale-space. Journal of Computational and Graphical Statistics. 2014;**23**:657-677

[47] Hannig J, Marron JS. Advanced distribution theory for SiZer. Journal of the American Statistical Association. 2006;**101**:484-499

[48] Weerahandi S. Generalized confidence intervals. Journal of the American Statistical Association. 1993;**88**:899-905