Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 547874, 13 pages
http://dx.doi.org/10.1155/2013/547874

Research Article

Robust SiZer Approach for Varying Coefficient Models

Hui-Guo Zhang,[1,2] Chang-Lin Mei,[2] and He-Ling Wang[3]

[1] *School of Mathematics and System Science, Xinjiang University, Urumqi 830000, China*
[2] *Department of Statistics, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China*
[3] *School of Applied Mathematics, Xinjiang University of Finance and Economics, Urumqi 830000, China*

Correspondence should be addressed to Chang-Lin Mei; clmei@mail.xjtu.edu.cn

Received 12 January 2013; Accepted 15 April 2013

Academic Editor: Joao B. R. Do Val

Varying coefficient models have widely been applied to many practical fields for exploring dynamic patterns of the regression relationships. In this study, we propose a robust scenario of SiZer (significant zero crossing of derivatives) inference approach based on the local least absolute deviation fitting procedure and the bootstrap confidence interval to uncover the statistically significant features of the coefficient functions in a varying coefficient model under different smoothing scales. The simulation study shows that the proposed SiZer approach is quite robust to outliers and performs well in finding the significant features of the coefficient functions. Furthermore, a real environmental data set is analyzed to demonstrate the application of the proposed approach.

1. Introduction

Varying coefficient models, introduced by Cleveland et al. [1] and Hastie and Tibshirani [2], are a useful extension of the classical linear regression models. By allowing the regression parameters in the linear models to be some nonparametric functions of a covariate, the models can be used to explore dynamic patterns of the regression relationship. Many approaches, including the kernel method [1], splines method [2–4], the local polynomial method [5, 6], and local maximum likelihood estimation method [7, 8], have been proposed to fit the varying coefficient models. Furthermore, Fan and Zhang [9] and Zhang and Peng [10] developed the simultaneous confidence bands of the coefficients in the varying coefficient models and the generalized varying coefficient models, respectively.

Most of the aforementioned methods are based on the mean regression and employ the least-square procedure to estimate the coefficients. As it is well known, the least-square procedure may not be a proper choice in the presence of heavy tailed errors or extreme values in a dataset and the resulting coefficient estimates consequently suffer from a lack of robustness. The effect of outliers may distort the model fitting process and create spurious patterns in the estimates of the coefficients [11]. Therefore, the robust procedures, such as L_1 estimation [12], M-estimation [13, 14], and quantile regression [15–17], have been proposed to handle the effect of outliers in the fits of the varying coefficient models. Although the above studies developed the robust methods to estimate the coefficient functions and established the pointwise asymptotic normality of the estimators, they mainly focused on the construction of the pointwise confidence intervals of the coefficients. This is inadequate for many applications. For instance, in the process of exploring the dynamic features of the regression relationship in the presence of outliers in the varying coefficient models, one often wants to know whether a coefficient function really varies, and if so, how it varies in detail.

The foregoing two questions involve the crucial problem of choosing an optimal bandwidth which may seriously affect the estimates of and the inferences on the regression coefficients. The issue of optimal bandwidth section is far from being solved satisfactorily, although there have been some practical robust criterion of choosing the bandwidth, such as the absolute cross-validation [12], the Schwarz information criterion [16], and the Akaike information criterion [17]. Furthermore, when the coefficient functions in a varying coefficient model possess different degrees of smoothness, the situation becomes more complicated because different bandwidths have to be, respectively, selected for the coefficients

with different degrees of smoothness [5, 6, 8, 9]. It is also worth noting that the selected optimal bandwidth for the estimation may not be proper for hypothesis testing [18, 19].

The SiZer (significant zero crossing of derivatives) method, proposed by Chaudhuri and Marron [20, 21], provides a powerful framework for exploring the features in a family of smooths indexed by different smoothing levels and allows one to handle the bandwidth selection problem in a new way. The SiZer method is based on the idea that the estimated curves under different smoothing levels may provide different information on the variation of the curve and highlights the full family of smooths instead of the "true underlying curve." Due to its flexibility and interpretability, the SiZer method has extensively been studied in its methodology and applications. Hannig and Marron [22] have proposed an improved simultaneous inference version of SiZer. The Bayesian scenario of SiZer has been developed by Erästö and Holmström [23], Godtliebsen and Øigård [24], and Øigård et al. [25]. Ganguli and Wand [26] and Godtliebsen et al. [27, 28] have considered the bivariate smoothing technique in SiZer. SiZer for smoothing splines method has been investigated by Marron and Zhang [29]. The other specific SiZer tools include SiZer for additive model [19], comparison of curves [30], time series analysis [31], SiZer for regression quantiles [32], analysis of random signals [33], and spatially dependent images [34], among others. Recently, Zhang and Mei [35] developed a SiZer approach for the varying coefficient models to explore the statistically significant features of the coefficient functions under different bandwidths. Although the study shows that the SiZer method is efficient at uncovering the genuine features such as monotonicity, peaks, valleys, and even the degree of smoothness of the coefficients, it did not consider the situation where outliers are present in the data. In fact, as pointed out by Hannig and Lee [36], outliers can seriously inflate the estimate of the model variance in the least-squares-based SiZer methods, which accordingly deflate the test statistics. As a result, some important underlying features in the coefficient functions may be missed, which can mislead the varying patterns of the regression relationship.

In this paper, We focus on developing the robust scenario of the SiZer inferential approach to the varying coefficient models in the presence of heavy tailed errors or extreme values in a dataset. The varying coefficient model is of the form

$$Y_i = \sum_{j=1}^{m} \beta_j (U_i) X_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where Y_i and $(U_i, X_{i1}, X_{i2}, \ldots, X_{im})$ are the observations of the response variable and the explanatory variables, respectively and $\beta_j(\cdot)$ $(j = 1, 2, \ldots, m)$ are unknown nonparametric functions and ε_i's are the independent random disturbances with zero median. Note that the model can include a varying intercept by setting $X_{i1} \equiv 1$. The proposed robust SiZer approach simultaneously considers a family of the median estimates of the coefficient functions indexed by a series of bandwidths, and the estimates are obtained by the local least absolute deviation (LAD) procedure [37]. The corresponding SiZer maps summarize a plenty of results of the multiple slope tests based on the bootstrap confidence intervals, which can provide the information on the variation patterns of the coefficients at different scales of smoothing.

The remainder of this paper is organized as follows. Section 2 derives the robust version of the SiZer approach based on the LAD median estimation of the varying coefficient model. Section 3 investigates the performance of the proposed method by simulations. The robust SiZer approach is further used to analyze a real-world dataset in Section 4. Concluding remarks are given in Section 5.

2. Robust SiZer Approach for the Varying Coefficient Model

2.1. LAD Estimates of the Coefficients and Their Derivatives. The local LAD method for nonparametric regression, studied comprehensively by Wang and Scott [37], is much more robust than the local least-squares procedures. Furthermore, the local LAD method can be easily generalized to the case of multiple regression because solving the LAD problem is equivalent to solving a linear programming that is easy to be implemented even for a large-scale linear programming. Tang and Wang [12] have adopted the local LAD method to calibrate the varying coefficient model. Considering that the estimates of the derivatives of the coefficient functions take a key role in the robust SiZer inference for varying coefficient model, we use the local linear LAD fitting procedure to simultaneously obtain the estimates of the coefficients and their derivatives.

Suppose that each coefficient $\beta_j(u)$ in the model in (1) has continuous derivative in the domain of u and let $\{(u_i, x_{i1}, \ldots, x_{im}, y_i)\}_{i=1}^{n}$ be a sample from the model. For each focal point u_0, the coefficient functions $\beta_j(u)$ $(j = 1, \ldots, m)$ can be approximated in the neighborhood of u_0 by

$$\beta_j (u) \approx \beta_j (u_0) + \beta_j' (u_0) (u - u_0), \quad j = 1, \ldots, m, \tag{2}$$

where $\beta_j'(u_0)$ is the derivative of $\beta_j(u)$ at u_0. Given u_0 and a bandwidth h, the local linear LAD estimates of $\beta_j(u_0)$ and $\beta_j'(u_0)$ are, respectively, denoted as $\widehat{\beta}_{jh}(u_0)$ and $\widehat{\beta}'_{jh}(u_0)$ $(j = 1, \ldots, m)$, which are the solution of the following locally weighted LAD problem:

$$\underset{\boldsymbol{\beta}(u_0)}{\text{minimize}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} \left[\beta_j (u_0) + \beta_j' (u_0) (u_i - u_0) \right] x_{ij} \right| \times K_h (u_i - u_0), \tag{3}$$

where $\boldsymbol{\beta}(u_0) = (\beta_1(u_0), \ldots, \beta_m(u_0), \beta_1'(u_0), \ldots, \beta_m'(u_0))^T$, and $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being a kernel function which is taken to be the Gaussian kernel.

Wagner [38] showed that solving an LAD problem is equivalent to solving a linear programming. For the weighted LAD problem in (3), the corresponding linear programming can be formulated as follows.

Let

$$r_{0i} = y_i - \sum_{j=1}^{m} \left[\beta_j(u_0) + \beta_j'(u_0)(u_i - u_0)\right] x_{ij},$$

$$r_{0i}^{+} = r_{0i} I \quad (r_{0i} \geq 0), \qquad r_{0i}^{-} = -r_{0i} I \quad (r_{0i} < 0), \tag{4}$$

$$I(A) = \begin{cases} 1, & \text{if } r_{0i} \in A, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \ldots, n$. We have $r_{0i} = r_{0i}^{+} - r_{0i}^{-}$ and $|r_{0i}| = r_{0i}^{+} + r_{0i}^{-}$ $(i = 1, 2, \ldots, n)$. Then the weighted LAD problem in (3) can equivalently be expressed as the following linear programming:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (r_{0i}^{+} + r_{0i}^{-}) K_h(u_i - u_0) \\ \text{subject to} \quad & \sum_{j=1}^{m} \left[\beta_j(u_0) + \beta_j'(u_0)(u_i - u_0)\right] x_{ij} \\ & + r_{0i}^{+} - r_{0i}^{-} = y_i, \quad i = 1, \ldots, n, \\ & r_{0i}^{+}, r_{0i}^{-} \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{5}$$

The solutions for $\beta_j(u_0)$ and $\beta_j'(u_0)$ $(j = 1, \ldots, m)$ in the linear programming in (5) are the corresponding local linear LAD estimates $\widehat{\beta}_{jh}(u_0)$ and $\widehat{\beta}_{jh}'(u_0)$ $(j = 1, \ldots, m)$, respectively.

The above linear programming can be solved by the so-called simplex method or interior point method (see [39, pages 181–202]). Currently, many computer software packages (e.g., the R software) are available for solving the linear programming problems.

2.2. Robust SiZer Maps for the Coefficient Functions. In order to uncover statistically significant features of a smooth, Chaudhuri and Marron [20] constructed a color map, called the SiZer map, to summarize the inferential results of the confidence intervals of the smooths of the derivatives at each location and each bandwidth. In the SiZer map, the horizontal axis associates with a certain covariable and the vertical axis is the logarithmic value of the bandwidth. When the confidence interval for the smooth of the derivative is completely above (below) zero at the given location and bandwidth, the corresponding pixel in the SiZer map will be colored blue (red) to illustrate that the smooth of the curve has a significantly increasing (decreasing, resp.) trend around the pixel. Furthermore, a purple pixel in the SiZer map means that the corresponding confidence interval for the smooth of the derivative contains zero, which suggests that the smooth of the curve does not significantly vary around that pixel. This graphical device can display the significant features of the smooth of the curve such as monotonicity, peaks, and valleys.

2.2.1. Bootstrap Confidence Intervals for the LAD Derivative Estimates. In order to build the robust SiZer maps for the varying coefficient model, we need to firstly construct the confidence interval for $\widehat{\beta}_{jh}'(u)$ $(j = 1, \ldots, m)$ and the LAD derivative estimates of the coefficient functions in model (1) with the bandwidth h. The asymptotic normality of the LAD estimator derived, for example, in the context of nonparametric regression [37] and in the context of the varying coefficient model [12], makes it possible to construct the asymptotic confidence interval. However, the nominal coverage probability of the confidence interval for LAD estimators may be very different from the true coverage probability in the case of the finite sample [40–42]. Thus, the wild bootstrap techniques have been proposed in the literature to improve the finite-sample performance of the tests in the median regression (e.g., [43, 44]). In this paper, we use a modified version of the residual-based wild bootstrap procedure [44, 45] to construct the confidence interval for the LAD estimator for median regression. Given the i.i.d. sample $\{(u_i, x_{i1}, \ldots, x_{im}, y_i)\}_{i=1}^{n}$, the method can be described in what follows.

Step 1. Define the residuals $\widehat{\varepsilon}_{ih} = y_i - \sum_{j=1}^{m} \widehat{\beta}_{jh}(u_i) x_{ij}$ $(i = 1, \ldots, n)$, where $\widehat{\beta}_{jh}(u_i)$ is the local linear LAD estimator of the jth coefficient $\beta_j(u)$ at u_i with the bandwidth h.

Step 2. For each i, draw the bootstrap residual $\varepsilon_{ih}^{*} = w_i |\widehat{\varepsilon}_{ih}|$, where the weight w_i is independently generated from the Bernoulli distribution with equal probabilities at -1 and 1.

Step 3. Compute the bootstrap sample

$$y_{ih}^{*} = \sum_{j=1}^{m} \widehat{\beta}_{jh}(u_i) x_{ij} + \varepsilon_{ih}^{*} \quad \text{for each } i. \tag{6}$$

Step 4. Refit the model in (1) based on the bootstrap sample $\{(u_i, x_{i1}, \ldots, x_{im}, y_{ih}^{*})\}_{i=1}^{n}$ and let $\widehat{\beta}_{jh}'^{*}(u)$ $(j = 1, \ldots, m)$ denote the bootstrap estimators of the derivatives of the coefficient functions at location u with bandwidth h.

Step 5. Repeat Steps 2–4 for B times at the given location u and bandwidth h and let $T_{jh,\tau}^{*}$ denote the τth sample quantile of $\{(\widehat{\beta}_{jh,1}'^{*}(u), \ldots, \widehat{\beta}_{jh,B}'^{*}(u))$, for all $u\}$. The full bootstrap percentile interval for $\widehat{\beta}_j'(u)$ with bandwidth h and the nominal coverage level $1 - \alpha(h)$ $(0 < \alpha(h) < 1)$ can be constructed as

$$C_{jh}^{\alpha(h)} = \left(T_{jh,(\alpha(h)/2)B}^{*}, T_{jh,(1-(\alpha(h)/2))B}^{*}\right), \tag{7}$$

for $j = 1, \ldots, m$.

2.2.2. Construction of the Robust SiZer Maps of the Coefficient Functions. Suppose that the grid points $\{\widetilde{u}_r;\ r = 1, 2, \ldots, R\}$ are equally spaced in the whole range of the covariate U. We choose a wide range of the bandwidth and denote the set of the bandwidth grids as $H = \{h_l;\ l = 1, 2, \ldots, L,$ with $h_1 < h_2 < \cdots, h_L\}$, where each $\log_{10}(h_l)$ is located at equally spaced L points on the interval $[\log_{10}(h_1), \log_{10}(h_L)]$. Empirically, according to the suggestion by Chaudhuri and Marron [20], h_1 can be taken as $2D_{\min}$ and h_L can be

taken as $1.5D_{\max}$, where $D_{\min}$ and $D_{\max}$ are, respectively, the minimal and maximal distances among all of the distances between each pair of sample points of covariate U. As a result, a SiZer map is covered by $R \times L$ grid pixels denoted by $(\widetilde{u}_r, h_l)$ $(r = 1, \ldots, R,\ l = 1, \ldots, L)$.

In practice, for each of the coefficient functions $\beta_j(u)$ $(j = 1, 2, \ldots, m)$ and the given bandwidth $h_l \in H$, we can perform the test according to the confidence interval in (7) at each pixel $(\widetilde{u}_r, h_l)$ and classify the confidence intervals into three categories: completely above zero, completely below zero, and containing zero. Then the corresponding pixel in the SiZer map is, respectively, colored in blue, red, and purple to illustrate that the smooth of each coefficient has a significantly increasing, significantly decreasing, and no significantly varying trend around each pixel.

3. Simulation Study

In this section, we conduct the simulation experiments to evaluate the performance of the proposed robust SiZer method on discovering the significant features of the coefficient functions in the varying coefficient model under the presence of outliers in the data.

3.1. Design of the Experiments. The experimental data are generated by the following varying coefficient model:

$$y_i = \beta_1(u_i)\, x_{1i} + \beta_2(u_i)\, x_{2i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{8}$$

where the observations u_i $(i = 1, \ldots, n)$ of the covariate U are independently drawn from the uniform distribution on the interval $(0, 1)$, and the observations x_{1i} and x_{2i} $(i = 1, \ldots, n)$ of the explanatory variables are randomly drawn from the uniform distributions on the intervals $(0, 2)$ and $(0, 3)$, respectively. The coefficients $\beta_1(u_i)$ and $\beta_2(u_i)$ are, respectively, chosen from the following three groups of the functions.

Group 1:

$$\beta_1(u) = \frac{1}{2}; \qquad \beta_2(u) = (2u - 1)^3. \tag{9}$$

Group 2:

$$\beta_1(u) = \cos(8\pi u); \qquad \beta_2(u) = 1 - 2(2u - 1)^2. \tag{10}$$

Group 3:

$$\beta_1(u) = \frac{4\cos(4\pi(2u - 1))}{15(2u - 1)^2 + 4}; \tag{11}$$

$$\beta_2(u) = \frac{4}{5}\left[\exp\left(-\frac{9}{25}(10u - 3)^2 + 1\right) + \exp\left(-\frac{9}{25}(10u - 7)^2 + 1\right)\right] - \frac{11}{10}. \tag{12}$$

The coefficients $\beta_1(u)$ and $\beta_2(u)$ in each of the above groups have different degrees of smoothness and they have been used in Zhang and Mei [35] to demonstrate the local least-squares-based SiZer analysis for the varying coefficient model. Figure 1 depicts the true curves of the coefficient functions $\beta_1(u)$ and $\beta_2(u)$ in Groups 1–3.

The random errors ε_{i2} $(i = 1, \ldots, n)$ in the model in (8) are independently drawn from one of the following distributions.

Case 1. The normal distribution $N(0, 0.5^2)$.

Case 2. The Cauchy distribution $C(0, 0.2)$ with the location parameter 0 and the scale parameter 0.2.

Case 3. The contaminated normal distribution $(1 - \delta)N(0, 0.5^2) + \delta N(0, 8^2)$, where the contaminating proportion δ is taken to be 0.3 in the simulations.

In order to give an impression of the signal to noise ratio in the simulation model, Figure 2 depicts the generated data with three error distributions and the coefficients in Group 2. When the model errors are drawn from the distribution $N(0, 0.5^2)$ in Case 1, theoretically, no outliers will be included in the data and this can be seen from Figure 2(a). The Cauchy distribution in Case 2 is of such heavy tails that its mean and variance do not exist. Therefore, the errors from this distribution generally include some extreme values or severe outliers. Indeed, Figure 2(b) displays the simulated data which contains extreme values. Similarly, when the model errors are drawn from the contaminated normal distribution in Case 3, about 30% of them come from $N(0, 8^2)$ and may include some excessively large values. Obviously, the simulated data in Figure 2(c) display a characteristic of fat tail. In summary, when the model errors are drawn from the distribution in Case 2 or 3, the data generated by the model in (8) will include outliers.

For each group of the above coefficient functions, we conduct the simulation with the sample size $n = 1000$. We choose the set of the bandwidth grids as $H = \{h_l;\ l = 1, 2, \ldots, 50\}$, where each $\log_{10}(h_l)$ is located at the equally spaced 50 points on the interval $[\log_{10}(0.01), \log_{10}(1.5)]$. The Gaussian kernel is used in the simulation. For each bandwidth $h_l \in H$, the LAD estimates of the coefficients and their derivatives are obtained by solving the linear programming in (5) at equally spaced 200 points on the range $(0, 1)$ of the covariate U. Given the confidence level $\alpha = 0.05$, the simultaneous bootstrap confidence interval is computed by (7) at each location (u_r, h_l) $(r = 1, 2, \ldots, 200;\ l = 1, 2, \ldots, 50)$ and the consequent SiZer maps are shown in Figures 2, 3, and 5. The white lines in the SiZer maps show the effective window widths, as the distance between the two white lines along the horizontal direction is $2h$.

The simulation is implemented by using the R software (http://cran.r-project.org/) and the codes are available from the authors.

3.2. Simulation Results with Analysis. Figures 3, 4, and 6, respectively, display the robust SiZer maps of the two coefficients $\beta_1(u)$ and $\beta_2(u)$ in Groups 1–3 with the model errors

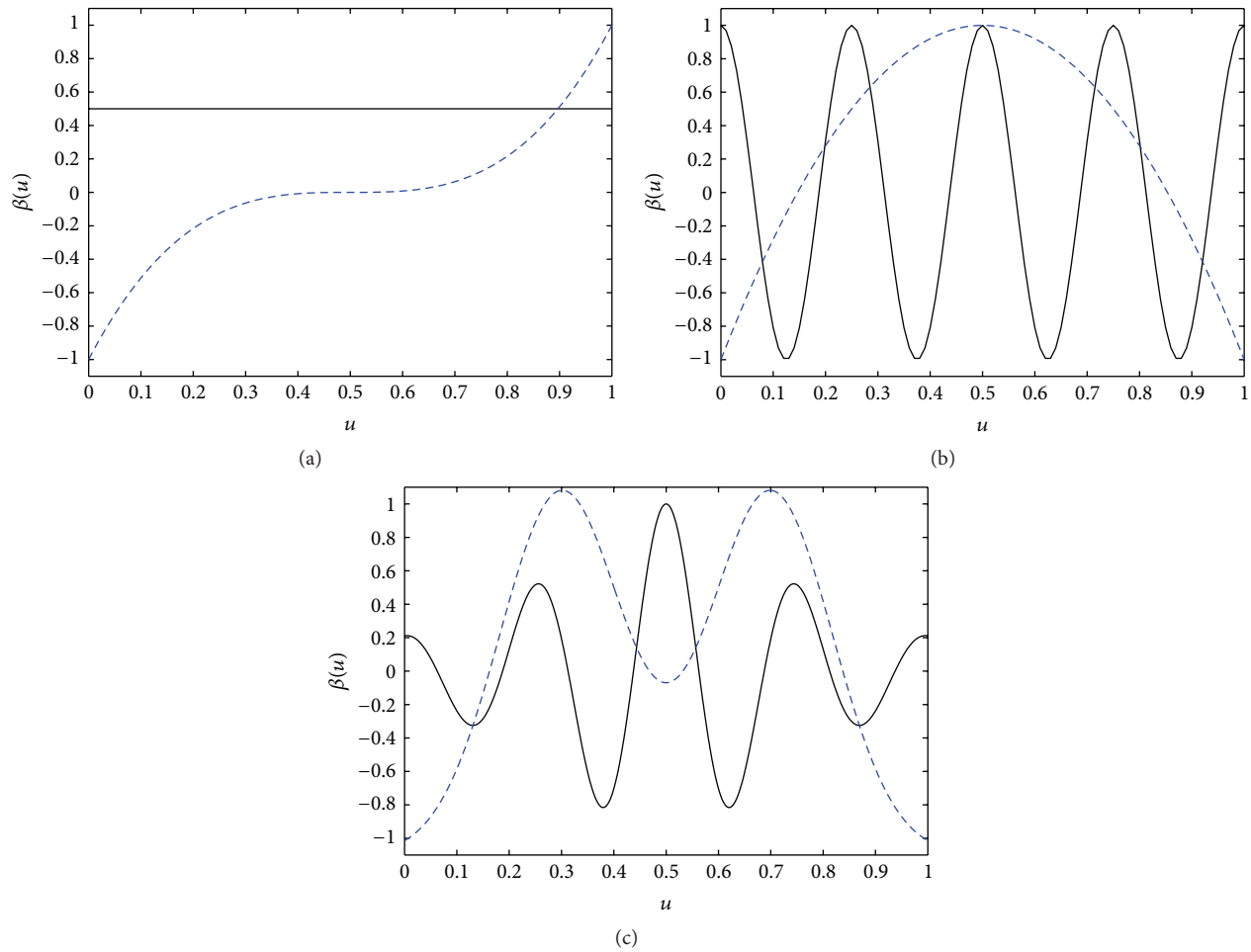

FIGURE 1: The true curves of the coefficients in (a) Group 1; (b) Group 2; and (c) Group 3. The solid curves are for $\beta_1(u)$ and the dashed curves are for $\beta_2(u)$.

respectively drawn from the normal distribution $N(0, 0.5^2)$, the Cauchy distribution $C(0, 0.2)$, and the contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 8^2)$.

We know from Figure 1(a) that the true coefficient $\beta_1(u)$ in Group 1 is constant, and therefore the correct SiZer maps should display no significant features no matter which distribution (the normal distribution, the Cauchy distribution, or the contaminated normal distribution) the model errors are drawn from. It can be seen from the Figures 3(a), 3(c), and 3(e) that the SiZer maps of $\beta_1(u)$ with three types of model errors show only purple color, indicating, as expected, that the coefficient is not varying. Figures 3(b), 3(d) and 3(f) display the SiZer maps of $\beta_2(u)$ with three types of model errors. We can infer from the SiZer maps that the coefficient $\beta_2(u)$ shows a significant increasing trend (blue) across the range of the covariate U for larger values of the bandwidth. Furthermore, it can be observed that, for smaller values of the bandwidth, the SiZer maps firstly show a significant increasing trend (blue) of $\beta_2(u)$ for smaller values of u, then no significant varying pattern (purple) for medium values of u, and a significant increasing trend (blue) for the larger values of u. Clearly, the significant features in each SiZer map correctly reveal the varying behaviors of a cubic function at the different scales.

It is also observed that when the model errors come from the Cauchy distribution and the contaminated normal distribution, the corresponding SiZer maps of both the coefficient functions are almost the same as those with the normally distributed model errors, which indicates that the outliers exert nearly no effects on the local LAD-based SiZer inference for the features of the coefficients.

Figure 4 shows the robust SiZer maps of the coefficients in Group 2 with the model errors drawn from the normal distribution, the Cauchy distribution and the contaminated normal distribution, respectively. We know from Figure 1(b) that $\beta_1(u)$ is a cosine curve and $\beta_2(u)$ is a parabola. The SiZer maps in Figures 4(a), 4(c), and 4(e) display the regular color changes with red-blue or blue-red for a quite wide range of the bandwidths. It is common knowledge that a peak on a smooth curve has the positive derivative (blue) on the left and the negative derivative (red) on the right, while the situation of a valley is reverse. Therefore, the SiZer maps of $\beta_1(u)$ correctly shows all of the peaks and valleys of the cosine curve with all three cases of model error distribution. The SiZer maps

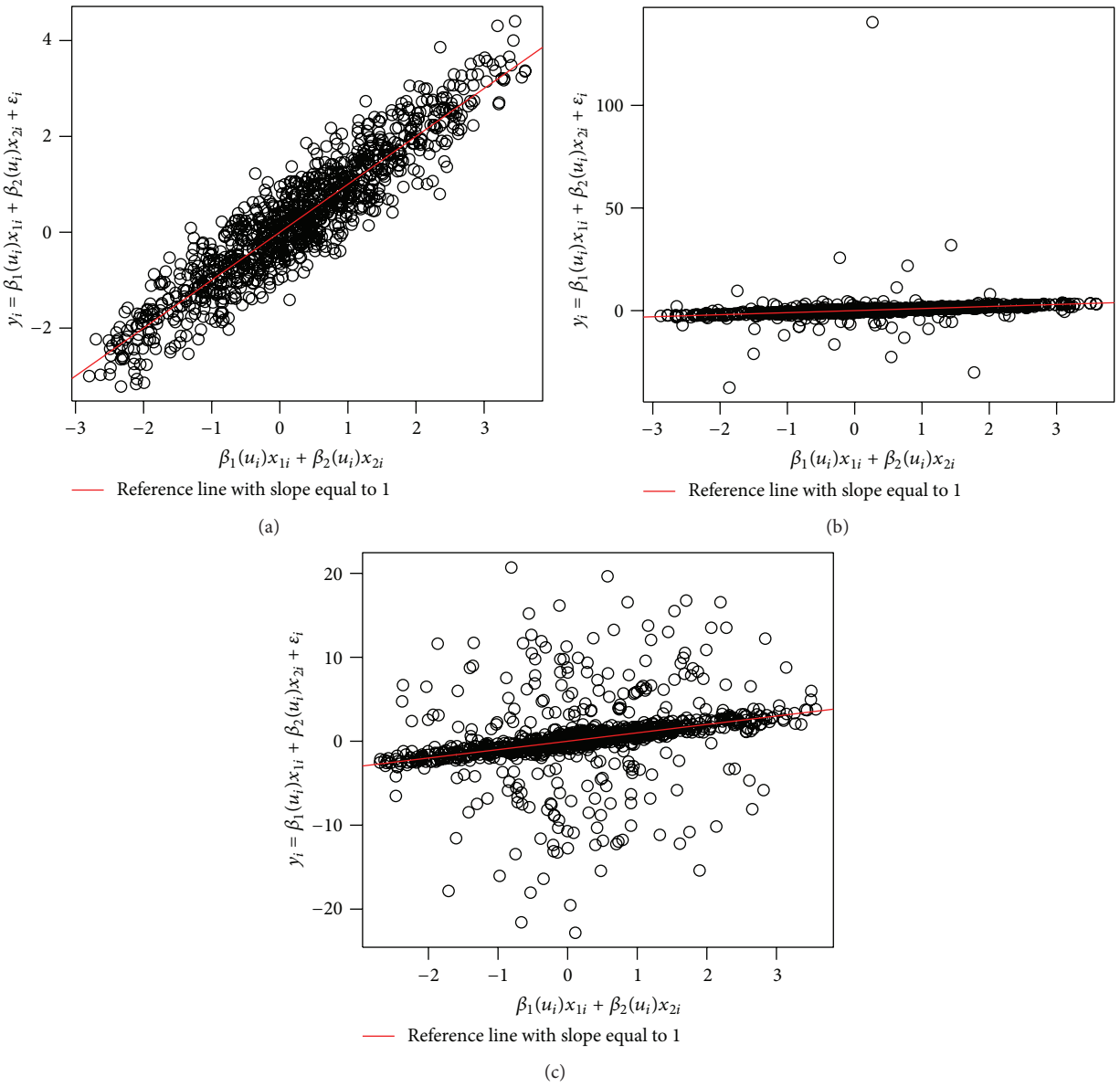

FIGURE 2: The scatter plots of simulated data generated by the varying coefficient model with coefficients in Group 3 and under model error drawn from (a) normal distribution $N(0, 0.5^2)$ in Case 1 (b) Cauchy distribution $C(0, 0.2)$ in Case 2 and (c) contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 16^2)$ in Case 3, respectively.

in Figures 4(b), 4(d), and 4(f) display only one significant peak because of a significant increasing trend (blue) on the left half interval of u and a significant decreasing trend (red) on the right half interval of u. These characteristics shown by the SiZer maps of $\beta_2(u)$ are entirely consistent with the true features of the parabola.

Moreover, it can also be observed that the peaks and valleys in Figure 4(b) are much fewer than those in Figure 4(a), meaning that $\beta_2(u)$ is smoother than $\beta_1(u)$. Therefore, the degrees of smoothness of the coefficients can be clearly visualized by the SiZer maps.

Additionally, in order to make a comparison of the robustness between the local LAD-based SiZer approach and the local least-squares-based SiZer method suggested by Zhang and Mei [35], Figure 5 displays the local least-squares-based SiZer maps and the families of the smooths of the coefficients in Group 2 with the model errors drawn, respectively, from the Cauchy distribution and the contaminated normal distribution. It can be seen that the SiZer maps cannot correctly reflect the features of either the cosine curve or the parabola. The smooths families in Figure 5 show that the estimated coefficient curves are seriously distorted by outliers. The reason is that outliers can inflate the estimate of the model variance and accordingly inflate the lengths of the corresponding confidence intervals. As a result, too many confidence intervals contain zero leading to the SiZer maps missing the important underlying features of coefficient functions.

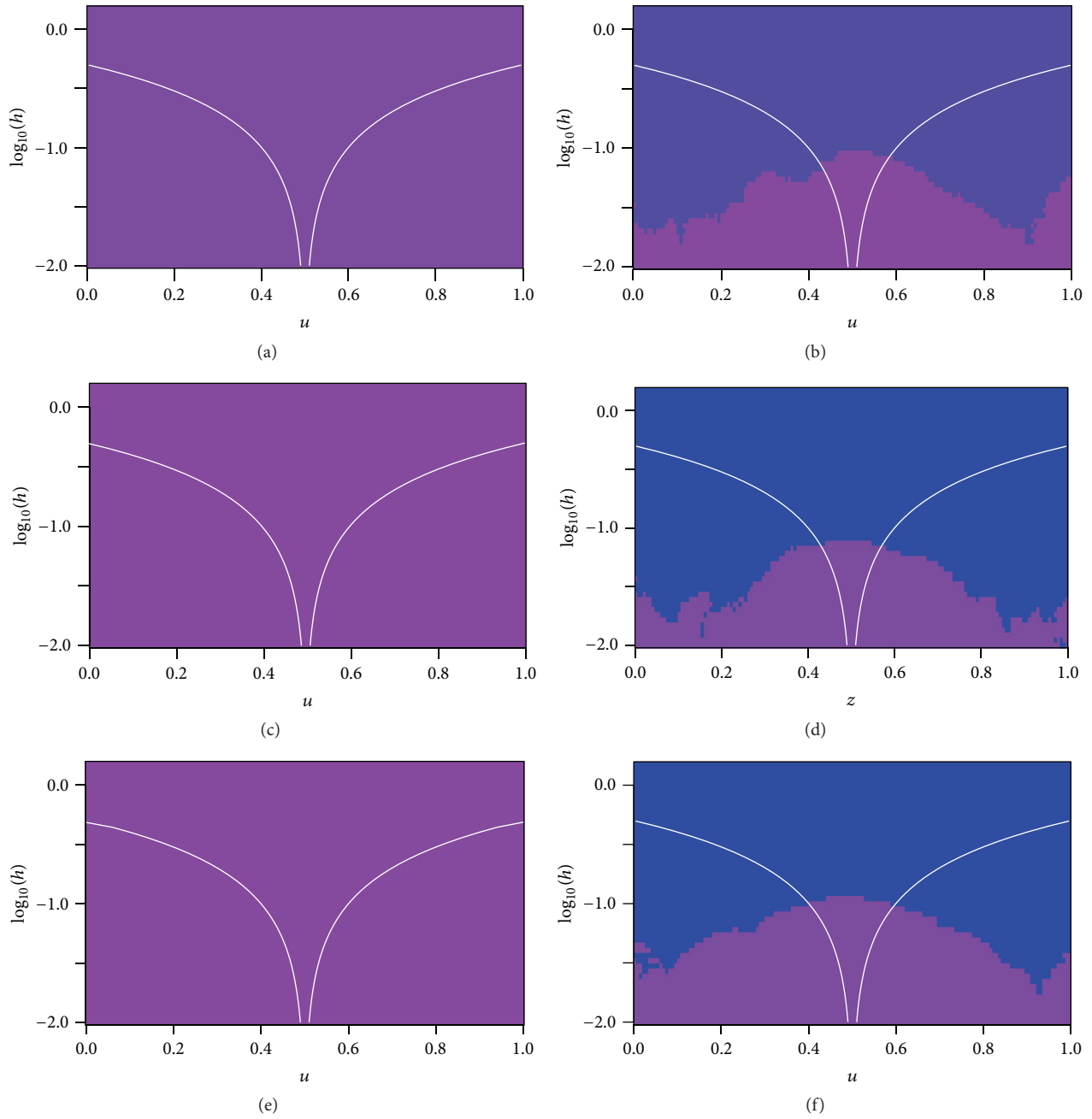

FIGURE 3: The robust SiZer maps of $\beta_1(u)$ (left panel) and $\beta_2(u)$ (right panel) in Group 1 with the model errors drawn from the normal distribution $N(0, 0.5^2)$ ((a) and (b)), the Cauchy distribution $C(0, 0.2)$ ((c) and (d)), and the contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 8^2)$ ((e) and (f)).

Figure 6 depicts the robust SiZer maps for the coefficients in Group 3 with normal, Cauchy, and contaminated normal model errors, respectively. It is known from Figure 1(c) that the two coefficients in Group 3 are respectively an amplitude decay function and a bimodal function. For the amplitude decay function, the robust SiZer maps in Figures 6(a), 6(c), and 6(e) precisely show the decay characteristics of the peaks and valleys from the central to the both sides of the range of u, although the minimal amplitudes are almost missed in Figures 6(c) and 6(e) when outliers appear in the data. In addition, it can be observed that the more permanent significant characteristics along h indicate the larger amplitudes of the peaks and the valleys. This can also be seen from the SiZer maps of the bimodal curve in Figures 6(b), 6(d) and 6(f), in which the significant features on the boundary areas of the maps are more permanent than those in the middle area.

A comparison with the local least-square-based SiZer method in Zhang and Mei [35] is also made in this case to demonstrate the robustness of the proposed SiZer approach.

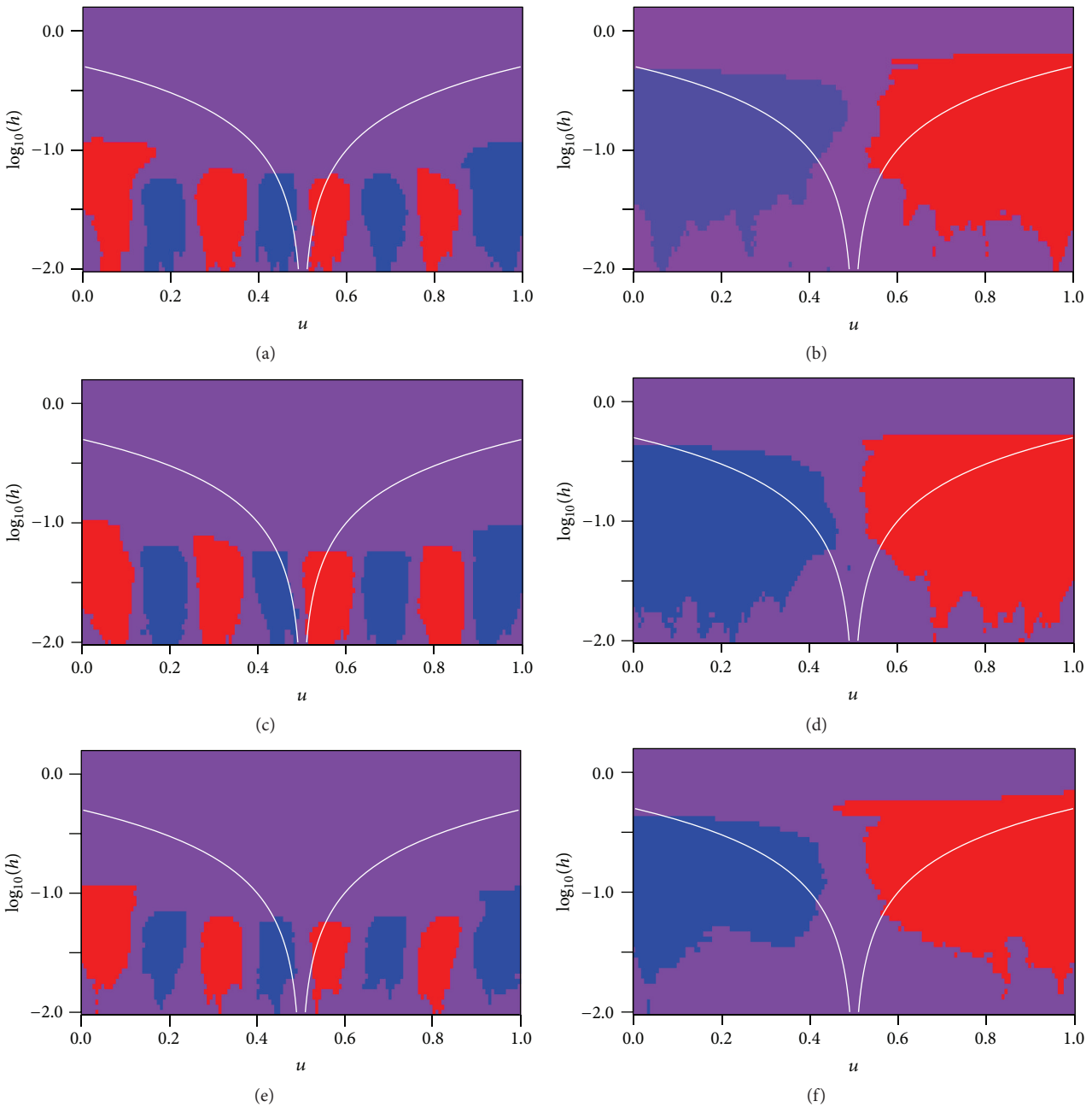

FIGURE 4: The robust SiZer maps of $\beta_1(u)$ (left panel) and $\beta_2(u)$ (right panel) in Group 2 with the model errors drawn from the normal distribution $N(0, 0.5^2)$ ((a) and (b)), the Cauchy distribution $C(0, 0.2)$ ((c) and (d)), and the contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 8^2)$ ((e) and (f)).

Like the situation of the second group coefficient functions, the local least-squares-based SiZer maps of the coefficient functions are seriously distorted by outliers. The corresponding SiZer maps provided in supplementary material available online at http://dx.doi.org/10.1155/2013/547874.

In summary, when the data set does not include outliers, the local LAD-based SiZer approach performs as well as the local least-squares-based SiZer method in revealing the features such as monotonicity, peaks, valleys, and the degrees of smoothness of the coefficients in a varying coefficient model. When the data set does include outliers, the local least-squares-based SiZer method is very sensitive to the outliers and produces distorted SiZer maps. In contrast, the proposed robust scenario of the SiZer approach is very robust to outliers even if severe outliers exist in the data.

4. A Real-Data Example

In this section, we demonstrate the application of the proposed robust SiZer approach via applying the environmental

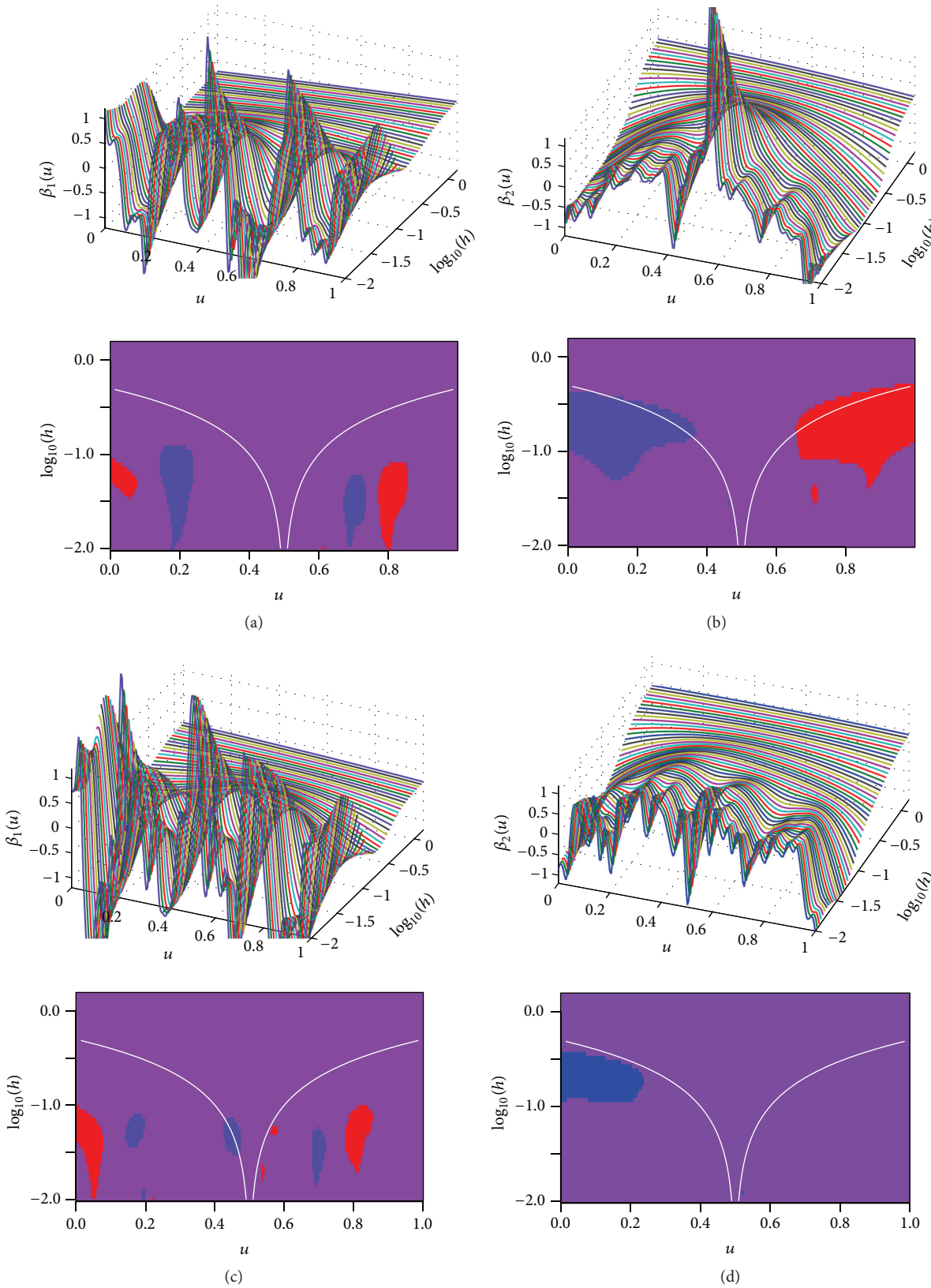

FIGURE 5: The local least-squares-based families of the smooths (top panel) and the SiZer maps (bottom panel) of the two coefficients in Group 2 with the model errors drawn from the Cauchy distribution $C(0, 0.2)$ ((a) and (b)) and the contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 8^2)$ ((c) and (d)).

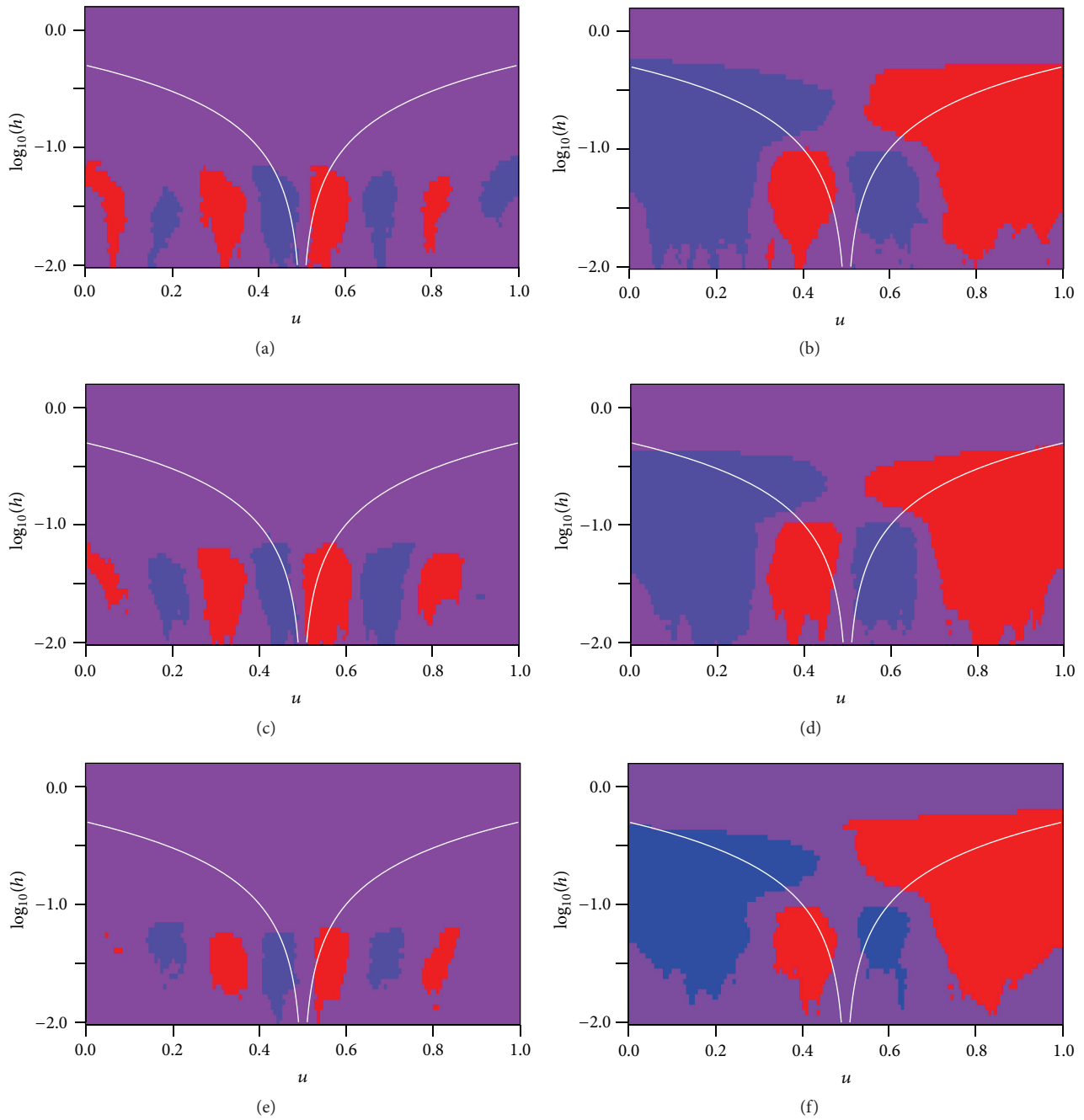

FIGURE 6: The robust SiZer maps of $\beta_1(u)$ (left panel) and $\beta_2(u)$ (right panel) in Group 3 with the model errors drawn from the normal distribution $N(0, 0.5^2)$ ((a) and (b)), the Cauchy distribution $C(0, 0.2)$ ((c) and (d)), and the contaminated normal distribution $0.7N(0, 0.5^2) + 0.3N(0, 8^2)$ ((e) and (f)).

data set that has been analyzed by Fan and Zhang [5, 9], and Cai et al. [7] for different purposes. The data were collected in Hong Kong from January 1 1994 to December 31, 1995 and consists of daily measurements of air pollutants along with the daily number of hospital admissions for circulatory and respiratory problems. The study aims at examining the association between the levels of the pollutants and the number of daily total hospital admissions for circulatory and respiratory problems. Like the model in Fan and Zhang [5, 9], we consider the following varying coefficient model:

$$Y = \beta_1(t) + \beta_2(t) X_2 + \beta_3(t) X_3 + \beta_4(t) X_4 + \varepsilon, \quad (13)$$

where the response variable Y is the number of daily hospital admissions, the explanatory variables X_2, X_3, and X_4 are the levels of sulphur dioxide (SO_2, in $\mu g/m^3$), Nitrogen Dioxide (NO_2, in $\mu g/m^3$), and dust (in $\mu g/m^3$), respectively, and

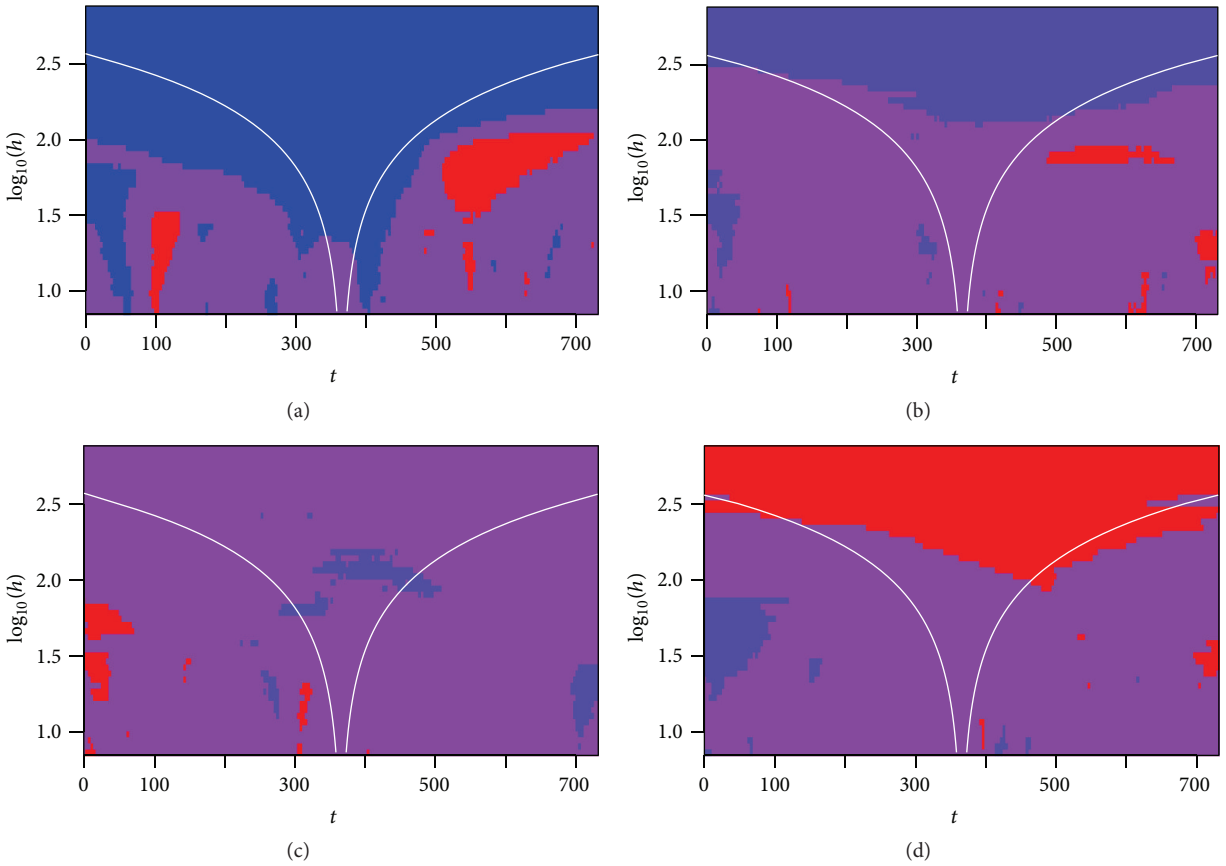

FIGURE 7: The robust SiZer maps of (a) $\beta_1(t)$, (b) $\beta_2(t)$, (c) $\beta_3(t)$, and (d) $\beta_4(t)$ in the model of the environmental data.

the covariate t is the time that the observations were collected. Additionally, each level of the three pollutants is centered by their respective averages.

We use the SiZer maps settings described in Section 2.2.2 and the proposed SiZer approach is performed to explore the significant features of the coefficients in the model in (13). The robust SiZer inference is carried out at the locations (t_r, h_l) for $r = 1, 2, \ldots, 200$ and $l = 1, 2, \ldots, 50$, where t_r are equally spaced 200 points in the range $[0, T]$ of the covariate t with $T = 730$, and $\log_{10}(h_l)$ are equally spaced 50 points on the bandwidth range $[\log_{10}(0.01T), \log_{10}(T)]$. Figure 7 shows the robust SiZer maps of the four coefficients.

The SiZer maps of the intercept $\beta_1(t)$ in Figure 7(a) and the coefficient $\beta_2(t)$ of the variable X_2 (i.e., SO_2) in Figure 7(b) show a significant increasing trend (blue) for larger bandwidths across the time range. However, for smaller bandwidths, the SiZer map of $\beta_1(t)$ displays the color changes with blue-red or red-blue meaning significant peaks and valleys, whereas the SiZer map of $\beta_2(t)$ does not show any permanent characteristics.

The SiZer map in Figure 7(c) displays a few weak and irregular features in the coefficient $\beta_3(t)$ of the variable X_3 (i.e., NO_2) for medium bandwidths, as the SiZer map shows a weak decrease trend (red) on the left but an irregular increase trend (blue) at the center. For the coefficient $\beta_4(t)$ of the variable X_4 (i.e., dust), its SiZer map in Figure 7(d) shows significantly decreasing (red) characteristics across the range of t for larger values of the bandwidth and a significant increase trend (blue) for smaller values of t at medium scales.

The above findings are very similar to those obtained by the local least-squares-based SiZer method in Zhang and Mei [35] except the SiZer map of the coefficient $\beta_3(t)$. For intermediate bandwidths, the local least-squares-based SiZer map of $\beta_3(t)$ displays a strong increase trend on the time range $(250, 450)$, but this significant characteristic does not appear in the robust SiZer map in Figure 7(c), which may indicate that there are outliers in the data. The outliers can be identified, with the help of the scatter plot of standardized data of response variable Y and explanatory variable X_3, which is shown in Figure 8. Figure 8 displays a cluster of outlier on the time range $(350, 450)$. The outliers distort the smoothing curve of $\beta_3(t)$, and the local least-squares-based SiZer map consequently shows the spurious increase trend.

The smoothed curves under particularly different bandwidths will represent different information in the data and the influences of potential outliers will also be reduced. The results indicate that the robust SiZer approach is useful for mining relatively full information in data and drawing more comprehensive conclusions.

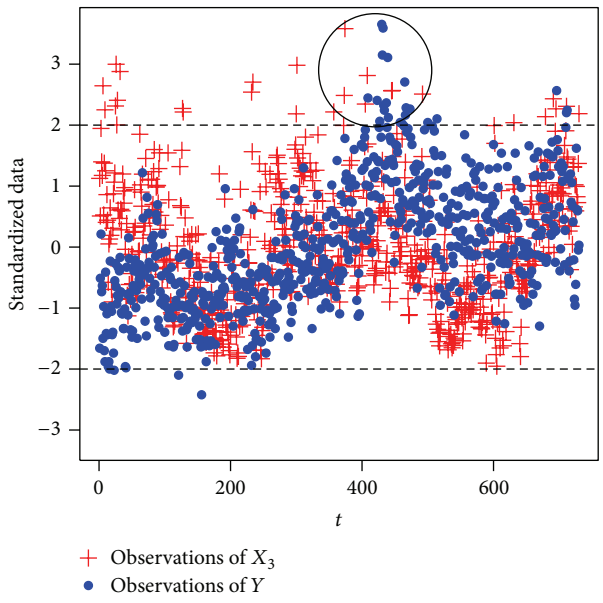

FIGURE 8: The scatter plot of standardize data of response variable Y and explanatory variable X_3 in the environmental data.

5. Concluding Remarks

The varying coefficient model is a useful statistical tool to explore dynamic patterns of a regression relationship, in which the variation features of the regression coefficients are taken as the main evidence to reflect the dynamic relationship between the response and the explanatory variables. However, outliers commonly exist in data and may lead to the distorted estimates of the coefficients and misleading inference on the underlying regression relationship. In this paper, we propose a robust scenario of the SiZer approach based on the local linear LAD procedure and the wild bootstrap confidence interval to uncover the genuine features such as monotonicity, peaks, valleys, and the degree of smoothness of the coefficients under different smoothing scales. The simulation study and the real environmental data analysis demonstrate that the proposed SiZer approach is very robust to outliers and has good performance in uncovering significant features of the coefficient functions in the varying coefficient model.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant no. 41261087), and the work was also supported by the National Natural Science Foundation of China (Grant nos. 11271296 and 10971161). He-Ling Wang's work was supported by the Humanities and Social Science Foundation of the Ministry of Education of China (Grant no. 12XJJC910001).

References

[1] W. S. Cleveland, E. Grosse, and M. J. Shyu, "Local regression models," in *Statistical Models in S. Pacific Grove*, J. M. Chambers and T. Hastie, Eds., pp. 309–376, Wadsworth, Belmont, Calif, USA, 1992.

[2] T. Hastie and R. Tibshirani, "Varying-coefficient models," *Journal of the Royal Statistical Society B*, vol. 55, no. 4, pp. 757–796, 1993.

[3] C.-T. Chiang, J. A. Rice, and C. O. Wu, "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 605–619, 2001.

[4] R. L. Eubank, C. Huang, Y. Muñoz Maldonado, N. Wang, S. Wang, and R. J. Buchanan, "Smoothing spline estimation in varying-coefficient models," *Journal of the Royal Statistical Society B*, vol. 66, no. 3, pp. 653–667, 2004.

[5] J. Fan and W. Zhang, "Statistical estimation in varying coefficient models," *The Annals of Statistics*, vol. 27, no. 5, pp. 1491–1518, 1999.

[6] J. Fan and J.-T. Zhang, "Two-step estimation of functional linear models with applications to longitudinal data," *Journal of the Royal Statistical Society B*, vol. 62, no. 2, pp. 303–322, 2000.

[7] Z. Cai, J. Fan, and R. Li, "Efficient estimation and inferences for varying-coefficient models," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 888–902, 2000.

[8] Z. Cai, "Two-step likelihood estimation procedure for varying-coefficient models," *Journal of Multivariate Analysis*, vol. 82, no. 1, pp. 189–209, 2002.

[9] J. Fan and W. Zhang, "Simultaneous confidence bands and hypothesis testing in varying-coefficient models," *Scandinavian Journal of Statistics*, vol. 27, no. 4, pp. 715–731, 2000.

[10] W. Zhang and H. Peng, "Simultaneous confidence band and hypothesis test in generalised varying-coefficient models," *Journal of Multivariate Analysis*, vol. 101, no. 7, pp. 1656–1680, 2010.

[11] H. G. Zhang and C. L. Mei, "Local least absolute deviation estimation of spatially varying coefficient models: robust geographically weighted regression approaches," *International Journal of Geographical Information Science*, vol. 25, pp. 1467–1489, 2011.

[12] Q. Tang and J. Wang, "L_1-estimation for varying coefficient models," *Statistics*, vol. 39, no. 5, pp. 389–404, 2005.

[13] T. Qingguo and C. Longsheng, "M-estimation and B-spline approximation for varying coefficient models with longitudinal data," *Journal of Nonparametric Statistics*, vol. 20, no. 7, pp. 611–625, 2008.

[14] T. Qingguo and C. Longsheng, "Asymptotic normality of M-estimators for varying coefficient models with longitudinal data," *Communications in Statistics*, vol. 38, no. 8–10, pp. 1422–1440, 2009.

[15] T. Honda, "Quantile regression in varying coefficient models," *Journal of Statistical Planning and Inference*, vol. 121, no. 1, pp. 113–125, 2004.

[16] M.-O. Kim, "Quantile regression with varying coefficients," *The Annals of Statistics*, vol. 35, no. 1, pp. 92–108, 2007.

[17] Z. Cai and X. Xu, "Nonparametric quantile estimations for dynamic smooth coefficient models," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 371–383, 2009.

[18] A. Azzalini and A. W. Bowman, "On the use of nonparametric regression for checking linear relationships," *Journal of the Royal Statistical Society B*, vol. 55, no. 2, pp. 549–557, 1993.

[19] W. González-Manteiga, M. D. Martínez-Miranda, and R. Raya-Miranda, "SiZer map for inference with additive models," *Statistics and Computing*, vol. 18, no. 3, pp. 297–312, 2008.

[20] P. Chaudhuri and J. S. Marron, "SiZer for exploration of structures in curves," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 807–823, 1999.

[21] P. Chaudhuri and J. S. Marron, "Scale space view of curve estimation," *The Annals of Statistics*, vol. 28, no. 2, pp. 408–428, 2000.

[22] J. Hannig and J. S. Marron, "Advanced distribution theory for SiZer," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 484–499, 2006.

[23] P. Erästö and L. Holmström, "Bayesian multiscale smoothing for making inferences about features in scatterplots," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 569–589, 2005.

[24] F. Godtliebsen and T. A. Øigård, "A visual display device for significant features in complicated signals," *Computational Statistics & Data Analysis*, vol. 48, no. 2, pp. 317–343, 2005.

[25] T. A. Øigård, H. Rue, and F. Godtliebsen, "Bayesian multiscale analysis for time series data," *Computational Statistics & Data Analysis*, vol. 51, no. 3, pp. 1719–1730, 2006.

[26] B. Ganguli and M. P. Wand, "Feature significance in geostatistics," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 954–973, 2004.

[27] F. Godtliebsen, J. S. Marron, and P. Chaudhuri, "Significance in scale space for bivariate density estimation," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, pp. 1–21, 2002.

[28] F. Godtliebsen, J. S. Marron, and P. Chaudhuri, "Statistical significance of features in digital images," *Image and Vision Computing*, vol. 22, pp. 1093–1104, 2004.

[29] J. S. Marron and J.-T. Zhang, "SiZer for smoothing splines," *Computational Statistics*, vol. 20, no. 3, pp. 481–502, 2005.

[30] C. Park and K.-H. Kang, "SiZer analysis for the comparison of regression curves," *Computational Statistics & Data Analysis*, vol. 52, no. 8, pp. 3954–3970, 2008.

[31] C. Park, J. Hannig, and K.-H. Kang, "Improved SiZer for time series," *Statistica Sinica*, vol. 19, no. 4, pp. 1511–1530, 2009.

[32] C. Park, T. C. M. Lee, and J. Hannig, "Multiscale exploratory analysis of regression quantiles using quantile SiZer," *Journal of Computational and Graphical Statistics*, vol. 19, no. 3, pp. 497–513, 2010.

[33] L. Holmström, L. Pasanen, R. Furrer, and S. R. Sain, "Scale space multiresolution analysis of random signals," *Computational Statistics & Data Analysis*, vol. 55, no. 10, pp. 2840–2855, 2011.

[34] A. Vaughan, M. Jun, and C. Park, "Statistical inference and visualization in scale-space for spatially dependent images," *Journal of the Korean Statistical Society*, vol. 41, no. 1, pp. 115–135, 2012.

[35] H.-G. Zhang and C.-L. Mei, "SiZer inference for varying coefficient models," *Communications in Statistics*, vol. 41, no. 10, pp. 1944–1959, 2012.

[36] J. Hannig and T. C. M. Lee, "Robust SiZer for exploration of regression structures and outlier detection," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 101–117, 2006.

[37] F. T. Wang and D. W. Scott, "The L_1 method for robust nonparametric regression," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 65–76, 1994.

[38] H. M. Wagner, "Linear programming techniques for regression analysis," *Journal of the American Statistical Association*, vol. 54, pp. 206–212, 1959.

[39] R. Koenker, *Quantile Regression*, Cambridge University Press, Cambridge, UK, 2005.

[40] J. L. Horowitz, "Bootstrap methods for median regression models," *Econometrica*, vol. 66, no. 6, pp. 1327–1351, 1998.

[41] M. Buchinsky, "Estimating the asymptotic covariance matrix for quantile regression models: a Monte Carlo study," *Joumal of Econometrics*, vol. 68, pp. 303–338, 1995.

[42] D. De Angelis, P. Hall, and G. A. Young, "Analytical and bootstrap approximations to estimator distributions in L^1 regression," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1310–1316, 1993.

[43] Y. Sun, "A consistent nonparametric equality test of conditional quantile functions," *Econometric Theory*, vol. 22, no. 4, pp. 614–632, 2006.

[44] X. Feng, X. He, and J. Hu, "Wild bootstrap for quantile regression," *Biometrika*, vol. 98, no. 4, pp. 995–999, 2011.

[45] R. Cao-Abad, "Rate of convergence for the wild bootstrap in nonparametric regression," *The Annals of Statistics*, vol. 19, no. 4, pp. 2226–2231, 1991.