Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 41: 656–671, 2014
doi: 10.1111/sjos.12054

A New Regression Model: Modal Linear Regression

WEIXIN YAO

Department of Statistics, Kansas State University

LONGHAI LI

Department of Mathematics and Statistics, University of Saskatchewan

ABSTRACT. The mode of a distribution provides an important summary of data and is often estimated on the basis of some non-parametric kernel density estimator. This article develops a new data analysis tool called modal linear regression in order to explore high-dimensional data. Modal linear regression models the conditional mode of a response Y given a set of predictors $\boldsymbol{x}$ as a linear function of $\boldsymbol{x}$. Modal linear regression differs from standard linear regression in that standard linear regression models the conditional mean (as opposed to mode) of Y as a linear function of $\boldsymbol{x}$. We propose an expectation–maximization algorithm in order to estimate the regression coefficients of modal linear regression. We also provide asymptotic properties for the proposed estimator without the symmetric assumption of the error density. Our empirical studies with simulated data and real data demonstrate that the proposed modal regression gives shorter predictive intervals than mean linear regression, median linear regression and MM-estimators.

Key words: forest fire data, linear regression, modal regression, mode

1. Introduction

The mode provides an important summary of data. Many authors have made efforts to identify modes of population distributions for low-dimensional data (see, for example, Muller & Sawitzki (1991), Scott (1992), Friedman & Fisher (1999), Chaudhuri & Marron (1999), Fisher & Marron (2001), Davies & Kovac (2004), Hall *et al.* (2004), Ray & Lindsay (2005), & Yao & Lindsay (2009), as well as documentations of the R package 'np' for non-parametric mode estimation). In high-dimensional data, it is often of interest to impose some model structure assumption on conditional distributions in order to identify associations between a response and a set of predictors. To the best of our knowledge, little research has been carried out to hunt conditional modes in regression problems.

Suppose we have collected a random sample $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is a p-dimensional column vector and y_i is observation of a continuous response variable Y. Conventional regression methods usually model the mean or median of $f(y \mid \boldsymbol{x})$ as a linear function of $\boldsymbol{x}$, where $f(y \mid \boldsymbol{x})$ is the conditional density function of Y given $\boldsymbol{x}$. In this article, we propose a new regression model called *modal linear regression* (MODLR) that assumes that the *mode* of $f(y \mid \boldsymbol{x})$ is a linear function of the predictor $\boldsymbol{x}$. MODLR measures the centre using the 'most likely' conditional values rather than the conditional average. Compared with other regression models, the proposed MODLR has the following features:

(i) MODLR attempts to capture the 'most probable' value—the mode (instead of the mean, median or quantile) of the conditional distribution of Y given $\boldsymbol{x}$. The conditional mode may be a more useful summary than the conditional mean when the conditional distribution of Y given $\boldsymbol{x}$ is asymmetric.

(ii) MODLR may provide shorter prediction intervals than other linear regression approaches for a nominal confidence level, because an interval around a conditional mode can cover more samples than an interval of the same length around a conditional mean.
(iii) MODLR is robust to outliers that do not follow the same relationship exhibited by the majority of a sample and is also robust to heavy-tailed conditional error distributions.
(iv) MODLR is well justified in situations where conditional distributions are highly skewed. Many robust regression methods still target the mean regression function and require symmetries in conditional distributions. Quantile regression is another alternative data analysis tool when the data is skewed. However, quantile regression cannot reveal the modal information and might produce a low density point prediction.

Modal linear regression is potentially a very useful addition to current data analysis tools. However, estimation of modal regression coefficients is not trivial. In this article, we propose an expectation–maximization (EM) algorithm that minimizes a kernel-based objective function for estimating modal regression coefficients. We have studied asymptotic and other theoretical properties of the proposed estimation procedure. We also propose a method for constructing asymmetric prediction intervals that can have better coverage than symmetric prediction intervals when conditional distributions are highly skewed.

The rest of this article is organized as follows. In Section 2, we introduce the kernel-based objective function and the EM algorithm for maximizing it; we also provide the theoretical properties of the estimation procedure. In Section 3, we use simulated datasets to compare the proposed MODLR with least square regression, median regression (MEDREG) and MM-estimators. We also compare these regression methods using forest fire data. Our empirical results show that MODLR provides significantly shorter prediction intervals than other regression methods. The article is concluded in Section 4 with discussions of possible future work. Proofs of the consistency of our estimators and necessary technical conditions are given in the Appendix.

2. Modal linear regression

2.1. Introduction to modal linear regression

Suppose that a response variable Y given a set of predictor $\boldsymbol{x}$ is distributed with a probability density function $f(y \mid \boldsymbol{x})$. Assume that the mode of $f(y \mid \boldsymbol{x})$, denoted by $\text{Mode}(Y \mid \boldsymbol{x}) = \arg\max_y (f(y \mid \boldsymbol{x}))$, is unique. The proposed MODLR method assumes that $\text{Mode}(Y \mid \boldsymbol{x})$ is a linear function of $\boldsymbol{x}$, that is,

$$\text{Mode}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}. \tag{1}$$

In (1), we assume that the first element of $\boldsymbol{x}$ is 1; this represents the intercept term. Let $\epsilon = y - \boldsymbol{x}^T \boldsymbol{\beta}$; we denote the conditional density of ϵ given $\boldsymbol{x}$ by $g(\epsilon \mid \boldsymbol{x})$ and refer to it as the *error distribution*. Note that the estimation method (and its asymptotic justification) that we will propose next allows for the error distribution to depend on $\boldsymbol{x}$.

If $g(\epsilon \mid \boldsymbol{x})$ is symmetric about 0, the $\boldsymbol{\beta}$ in (1) will be the same as the coefficients obtained by conventional mean linear regression; however, if $g(\epsilon \mid \boldsymbol{x})$ is skewed, modal regression coefficients and conventional linear regression coefficients will be different. It is even possible that the mode of Y given $\boldsymbol{x}$ is a linear function of $\boldsymbol{x}$, but the conventional mean is non-linear. The following example illustrates the difference between modal regression function and conventional mean regression function when the error distribution is skewed.

Example 1. Let $(\boldsymbol{x}, Y)$ satisfy the following model assumption

$$Y = m(\boldsymbol{x}) + \sigma(\boldsymbol{x})\varepsilon, \tag{2}$$

where ε has density $h(\cdot)$. Suppose $h(\cdot)$ is a skewed density with mean 0 and mode 1.

(i) If $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ and $\sigma(x) = \boldsymbol{x}^T\boldsymbol{\alpha}$, then

$$\mathrm{E}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Mode}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{\beta} + \boldsymbol{\alpha}).$$

Thus, Y depends on $\boldsymbol{x}$ linearly from the point of view of both mean regression and modal regression even though their regression parameters are different.

(ii) If $m(\boldsymbol{x}) = 0$ and $\sigma(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\alpha}$, then

$$\mathrm{E}(Y \mid \boldsymbol{x}) = 0 \quad \text{and} \quad \mathrm{Mode}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\alpha}.$$

Therefore, in terms of conditional mean, Y does not depend on $\boldsymbol{x}$; however, in terms of conditional mode, Y does depend linearly on $\boldsymbol{x}$. From this example, we see that variable selection techniques based on modal regression might reveal some useful predictors that mean regression cannot.

To estimate the modal regression parameter $\boldsymbol{\beta}$ in (1), we propose *maximizing* the kernel-based objective function

$$Q_h(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n} \phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right), \tag{3}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$ and $\phi(t)$ is a kernel density function symmetric about 0. For the remainder of the paper, we will assume that ϕ is the standard normal density (for the simplicity of computation). On the basis of this choice of kernel, the M-step of the modal EM (MEM) algorithm presented next has the closed-form solution shown in (6). It should be noted that all the asymptotic results presented in this article still hold if other kernels are used. We will denote the maximizer of (3) by $\hat{\boldsymbol{\beta}}$ and call it the MODLR estimator, shortened by MODLRE.

We now explain why (3) can be used to estimate the modal regression coefficients. We first look at the simplest case in which there is no predictor, that is, $\boldsymbol{\beta} = \beta_0$. For such cases, (3) is simplified to

$$Q_h(\beta_0) \equiv \frac{1}{n}\sum_{i=1}^{n} \phi_h(y_i - \beta_0). \tag{4}$$

Note that $Q_h(.)$ is the kernel estimate of the density function of Y. Therefore, the maximizer of (4) is the mode of the kernel density function based on $y_1, \ldots, y_n$. When $n \to \infty$ and $h \to 0$, the mode of this kernel density function will converge to the mode of the distribution of Y. Such a modal estimator has been proposed by Parzen (1962). When there are predictors present, for any fixed $\boldsymbol{\beta}$, $Q_h(\boldsymbol{\beta})$ in (3) is the value of the kernel density function based on the residuals $\epsilon_i = y_i - x_i\boldsymbol{\beta}$ at $\epsilon = 0$. Maximizing (3) with respect to $\boldsymbol{\beta}$ yields the line $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ such that the kernel density function of residuals ϵ_i has the highest value at 0. In the special case that $\phi_h(t) = (2h)^{-1}I(|t| \le h)$, a uniform kernel, maximizing (3) yields the line $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ such that the band $\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \pm h$ covers the largest number of response values y_i.

Lee (1989) used a uniform kernel to estimate modal regression coefficients. In his theoretical investigation, h is fixed and does not depend on the sample size n. In order to get consistency results for the estimator, Lee assumed the error distribution to be symmetric. Note that in such cases, the modal line is the same as the traditional mean regression line. Thus, Lee's theoretical results did not justify applications of MODLR for situations with skewed error distributions (where MODLR is more useful than other regression methods). In this article, we prove (see Appendix for details) that if we let $h \to 0$ when $n \to \infty$, the $\hat{\boldsymbol{\beta}}$ found by maximizing $Q_h(\boldsymbol{\beta})$ in (3) is a consistent estimate of the modal regression parameter in (1) for very general error density functions without symmetry assumptions.

2.2. *Modal expectation–maximization algorithm*

There is no closed-form expression of the maximizer of (3); therefore, we propose to extend the MEM algorithm (Li *et al.*, 2007; Yao, 2013) in order to maximize (3).

Similar to an EM algorithm, the MEM algorithm consists of an E-step and an M-step. Starting with $\boldsymbol{\beta}^{(0)}$, repeat the following two steps until it converges:

E-Step: In this step, we calculate weights $\pi\left(j \mid \boldsymbol{\beta}^{(k)}\right), j = 1, \ldots, n$ as

$$\pi\left(j \mid \boldsymbol{\beta}^{(k)}\right) = \frac{\phi_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}^{(k)}\right)}{\sum_{i=1}^n \phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)} \propto \phi_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}^{(k)}\right). \quad (5)$$

M-Step: In this step, we update $\boldsymbol{\beta}^{(k+1)}$

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} \left\{\pi\left(j \mid \boldsymbol{\beta}^{(k)}\right) \log \phi_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right)\right\} \\ &= \left(\boldsymbol{X}^T\boldsymbol{W}_k\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}_k\boldsymbol{y},\end{aligned} \quad (6)$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{W}_k$ is an $n \times n$ diagonal matrix with diagonal elements $\pi(j \mid \boldsymbol{\beta}^{(k)})$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$.

Some remarks on the proposed MEM algorithm:

(i) the major difference between the least squares estimate (LSE) and the MODLRE lies in the E-step. For the LSE, each observation has equal weights, whereas for MODLRE, the weights depend on how close y_i is to the modal regression line. This weighting scheme allows MODLRE to reduce the effect of observations far away from the modal regression line and thus robustness is achieved.

(ii) when the normal kernel is used for ϕ in (3), the function optimized in the M-step is a weighted sum of log likelihoods corresponding to ordinary linear regression. In this case, we obtain a closed-form expression for the maximizer in (6). If other kernels are used, some optimization algorithms are needed in the M-step.

(iii) the converged value obtained by the MEM algorithm depends on the starting point chosen, and there is no guarantee that the algorithm will converge to the global optimal solution of (3). Therefore, it is prudent to run the algorithm multiple times using several different starting points and choose the best local optima found.

We have proven (see Appendix) the ascending property of the proposed MEM for any choice of kernel for ϕ in (3):

Theorem 2.1. *Each iteration of (5) and (6) will monotonically non-decrease the objective function (3), that is,* $Q_h\left(\boldsymbol{\beta}^{(k+1)}\right) \geq Q_h\left(\boldsymbol{\beta}^{(k)}\right)$, *for all k.*

The iteratively reweighted least squares (IRWLS) algorithm has been commonly used for general M-estimators. Because the maximizer of (3) can be considered as a special case of M-estimators, the IRWLS algorithm can be applied to find $\hat{\boldsymbol{\beta}}$. When the normal kernel $\phi(\cdot)$ is used, the IRWLS algorithm is indeed equivalent to the proposed MEM algorithm, but when other kernels are used, the two algorithms are different. IRWLS has been proven to be ascending (i.e. monotonically non-decreases the objective function) if $-\phi(x)/x$ is non-increasing (Huber, 1981). However, when $\phi(x)$ is a normal density function, $-\phi(x)/x$ is not non-increasing. Therefore, the existing theories of IRWLS cannot justify theorem 2.1 if the normal kernel $\phi(\cdot)$ is used. Because the proof of theorem 2.1 is for any kernel density $\phi(\cdot)$, including the normal kernel, theorem 2.1 provides an extension to existing IRWLS theories.

2.3. Asymptotic properties of $\hat{\boldsymbol{\beta}}$

The asymptotic properties established for traditional M-estimators are based on assumptions that the error density is symmetric and the objective function is fixed. In addition, the target of traditional M-estimators is the conditional mean. For our proposed modal regression, we will allow that the tuning parameter h in the objective function goes to zero and the error density can be skewed. Therefore, the theoretical results on the traditional M-estimators cannot be directly applied to the proposed MODLRE. In this section, we will give the results about the consistency of the proposed modal regression estimator $\hat{\boldsymbol{\beta}}$ for model (1), its convergence rate and its asymptotic distribution. Their proofs are given in the Appendix.

Theorem 2.2. *When $h \to 0$ and $nh^5 \to \infty$, under the regularity conditions (A1)–(A3) in the Appendix, there exists a consistent maximizer of (3) such that*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p\left\{h^2 + \left(nh^3\right)^{-1/2}\right\},$$

where $\boldsymbol{\beta}_0$ is the true coefficient of the modal regression function defined in (1).

Theorem 2.3. *Under the same assumptions as theorem 2.2, the $\hat{\boldsymbol{\beta}}$ that satisfies the consistency result given in theorem 2.2 has the following asymptotic normality result*

$$\sqrt{nh^3}\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \frac{h^2}{2}J^{-1}K\{1 + o_p(1)\}\right] \xrightarrow{D} N\left\{0, \nu_2 J^{-1}LJ^{-1}\right\}, \tag{7}$$

where $\nu_2 = \int t^2\phi^2(t)dt$ and

$$J = \mathrm{E}\left\{g''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right\}; K = \mathrm{E}\left\{g'''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\right\}; L = \mathrm{E}\left\{g(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right\}.$$

Parzen (1962) and Eddy (1980) have proven similar asymptotic results for kernel estimators of the mode of the distribution of Y without conditioning on $\boldsymbol{x}$. Therefore, the results of Parzen (1962) and Eddy (1980) can be considered as special cases of theorem 2.3 when there is no predictor involved, that is, $\boldsymbol{x} = 1$.

By theorem 2.3, the asymptotic bias of $\hat{\boldsymbol{\beta}}$ is $h^2J^{-1}K/2$, and the asymptotic variance is $\nu_2 J^{-1}LJ^{-1}/(nh^3)$. A theoretic optimal bandwidth h for estimating $\boldsymbol{\beta}$ can be obtained by minimizing the asymptotic weighted mean squared errors

$$\mathrm{E}\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T W(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\} \approx K^T J^{-1}WJ^{-1}Kh^4/4 + (nh^3)^{-1}\nu_2 \mathrm{tr}\left(J^{-1}LJ^{-1}W\right),$$

where $\mathrm{tr}(A)$ is the trace of A and W is a diagonal matrix, whose diagonal elements reflect the importance of the accuracy in estimating different coefficients. Therefore, the asymptotic optimal bandwidth h is

$$\hat{h}_{opt} = \left[\frac{3\nu_2 \mathrm{tr}\left(J^{-1}LJ^{-1}W\right)}{K^T J^{-1} W J^{-1} K} \right]^{1/7} n^{-1/7}. \tag{8}$$

If $W = (J^{-1}LJ^{-1})^{-1} = JL^{-1}J$, which is proportional to the inverse of the asymptotic variance of $\hat{\boldsymbol{\beta}}$, then

$$\hat{h}_{opt} = \left[\frac{3\nu_2(p+1)}{K^T L^{-1} K} \right]^{1/7} n^{-1/7}. \tag{9}$$

Let $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_s)^T$, where β_0 is a scalar intercept parameter and $\boldsymbol{\beta}_s$ is the slope parameter. If ϵ is independent of $\boldsymbol{x}$, then

$$J^{-1}K = (1, 0, \ldots, 0)^T g'''(0) / \left\{2g''(0)\right\},$$

and thus, the asymptotic bias of the slope parameter $\boldsymbol{\beta}_s$ is 0. Therefore, the optimal bandwidth h for estimating $\boldsymbol{\beta}_s$ should go to infinity, which implies that the resulting estimate $\hat{\boldsymbol{\beta}}_s$ is an LSE with root n consistency. This is expected because when ϵ is independent of $\boldsymbol{x}$, the slope parameter $\boldsymbol{\beta}_s$ of modal regression line is the same as the slope parameter of conventional mean regression line and thus can be estimated at root n convergence rate.

Given the root n consistent estimate $\hat{\boldsymbol{\beta}}_s$ (using LSE, for example), we propose to further estimate β_0 by

$$\hat{\beta}_0 = \arg\max_{\beta_0} \frac{1}{n} \sum_{i=1}^{n} \phi_h \left(y_i - x_i^T \hat{\boldsymbol{\beta}}_s - \beta_0 \right). \tag{10}$$

The aforementioned maximization can be carried out similarly using the MEM algorithm proposed in Section 2.2 We have the following result for $\hat{\beta}_0$. Its proof is given in the Appendix.

Theorem 2.4. *Under the same assumption as theorem 2.2, if ϵ is independent of $\boldsymbol{x}$ and $g''(0) \neq 0$, the $\hat{\beta}_0$ defined in (10) has the following asymptotic distribution:*

$$\sqrt{nh^3} \left\{ \hat{\beta}_0 - \beta_0 - \frac{g'''(0)h^2}{2g''(0)} + o_P(h^2) \right\} \xrightarrow{D} N \left\{ 0, \frac{g(0)\nu_2}{[g''(0)]^2} \right\}. \tag{11}$$

Note that when ϵ is independent of $\boldsymbol{x}$, $J^{-1}LJ^{-1} = g''(0)^{-2} g(0) \mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)^{-1}$. Let

$$A = \mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right) = \begin{pmatrix} 1 & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}$$

and a^{11} be the (1,1) element of A^{-1}. Noting that $a^{11} = \left(1 - A_{12}A_{22}^{-1}A_{12}^T\right)^{-1}$ and A_{22} is positive definite, we have $a^{11} \geq 1$. Therefore, on the basis of theorems 2.3 and 2.4, we can see that using the root n consistent estimate $\hat{\boldsymbol{\beta}}_s$ as initial, we can get more efficient estimate of the intercept β_0 than the one found by maximizing (3) directly. This is reasonable because the estimate $\hat{\beta}_0$ in (10) need not account for the uncertainty of $\hat{\boldsymbol{\beta}}_s$ because of its root n consistency, and thus, $\hat{\beta}_0$ is asymptotically as efficient as if $\boldsymbol{\beta}_s$ were known.

From theorem 2.4, we can see that the asymptotic bias of $\hat{\beta}_0$ is $\{2g''(0)\}^{-1} g'''(0)h^2$ and its asymptotic variance is $[\{g''(0)\}^2 nh^3]^{-1} g(0)\nu_2$. By minimizing the asymptotic MSE, we can get the asymptotic optimal bandwidth h for estimating β_0:

$$\hat{h}_{opt} = \left[\frac{3g(0)\nu_2}{\{g'''(0)\}^2} \right]^{1/7} n^{-1/7}. \tag{12}$$

2.4. Finite sample breakdown point

To investigate robustness of the MODLRE, we also calculate its finite sample breakdown point. A breakdown point is used to quantify the proportion of bad data in a sample that an estimator can tolerate before returning arbitrary values. Because usually the breakdown point is most useful in a small sample set-up (Donoho, , 1982; Donoho & Huber, , 1983), we will mainly focus on the finite sample breakdown point. A number of definitions for the finite sample breakdown point have been proposed (see, for example, Hampel, , 1971, 1974; Donoho, , 1982; Donoho & Huber, , 1983). In this paper, we shall work with the finite sample contamination breakdown point. Let $z_i = (\boldsymbol{x}_i, y_i)$. Given the sample $\boldsymbol{Z} = (z_1, \ldots, z_n)$, denote $T(\boldsymbol{Z})$ the MODLRE $\hat{\boldsymbol{\beta}}$, as defined as the maximizer of (3). We can corrupt the original sample $\boldsymbol{Z}$ by adding m arbitrary points $\boldsymbol{Z}' = (z_{n+1}, \ldots, z_{n+m})$. The corrupted sample $\boldsymbol{Z} \cup \boldsymbol{Z}'$ then has sample size $n+m$ and contains a fraction $\delta = m/(m+n)$ of bad values. The finite sample contamination breakdown point δ^* is defined as

$$\delta^*(\boldsymbol{Z}, T) = \min_{1 \le m \le n} \left\{ \frac{m}{n+m} : \sup_{\boldsymbol{Z}'} ||T(\boldsymbol{Z} \cup \boldsymbol{Z}')|| = \infty \right\}, \tag{13}$$

where $||\cdot||$ is Euclidean norm.

Theorem 2.5. *Given observations $\boldsymbol{Z} = (z_1, \ldots, z_n)$, suppose $T(\boldsymbol{Z}) = \hat{\boldsymbol{\beta}}$, the MODLRE defined as the maximizer of (3). Let*

$$M = \sqrt{2\pi} h \sum_{i=1}^{n} \phi_h \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \right). \tag{14}$$

Then the finite sample contamination breakdown point of MODLRE is

$$\delta^*(\boldsymbol{Z}, T) = \frac{m^*}{n+m^*}, \tag{15}$$

where m^* is an integer satisfying $\lceil M \rceil \le m^* \le \lfloor M \rfloor + 1$, $\lfloor a \rfloor$ is the largest integer not greater than a and $\lceil a \rceil$ is the smallest integer not less than a.

The proof of theorem 2.5 is given in the Appendix. From the aforementioned theorem, we can see that the breakdown point depends not only on $\phi(\cdot)$, and the tuning parameter h, but also on the sample configuration. (However, Huber (1984) pointed out if the scale (contained in the bandwidth h of the MODLRE) is determined from the sample itself, empirically, the breakdown point is quite high.)

3. Simulation study and application

In this section, we conduct a Monte Carlo simulation study in order to assess the performance of our proposed MODLR under a finite sample size scenario. We will compare MODLR with some other regression methods. A real data application is also provided.

3.1. Bandwidth selection

The modal regression estimator requires a selection of the bandwidth. The asymptotically optimal bandwidth formula (9) contains the unknown quantities $g^{(v)}(0 \mid \boldsymbol{x})$, $v = 0, 2, 3$, that is, the vth derivative of the conditional density of ϵ given $\boldsymbol{x}$.

Hence, they are not ready to use. A commonly used method is to replace these unknown quantities with estimates. Given the initial residual $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the traditional

LSE (or a robust estimate if there are some outliers) of $\boldsymbol{\beta}$, we can estimate their mode, denoted by $\hat{m}$, by maximizing the kernel density estimator (Parzen, 1962). Under the assumption of independence of ϵ and $\boldsymbol{x}$, $\hat{\epsilon}_i - \hat{m}$ approximately has density $g(\cdot)$, and thus, $g^{(v)}(0 \mid \boldsymbol{x})$ can be estimated by (see, for example, Silverman, , 1986 and Scott, , 1992)

$$\hat{g}^{(v)}(0 \mid \boldsymbol{x}) = \frac{1}{nh^{v+1}} \sum_{i=1}^{n} K^{(v)} \left\{ \frac{\hat{\epsilon}_i - \hat{m}}{h} \right\}, \; v = 0, 2, 3,$$

where h is chosen using the method reported by Botev *et al.* (2010) and $K^{(v)}(\cdot)$ is the vth derivative of kernel density function $K(\cdot)$. Then, we can estimate J, K and L by

$$\hat{J} = n^{-1} \sum_{i=1}^{n} \hat{g}''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T, \; \hat{K} = n^{-1} \sum_{i=1}^{n} \hat{g}'''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i \text{ and}$$
$$\hat{L} = n^{-1} \sum_{i=1}^{n} \hat{g}(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

and apply (9) to estimate $\hat{h}_{opt}$. To refine the bandwidth selection, one might further iteratively update a chosen bandwidth by recalculating the residual $\hat{\epsilon}_i$ given by the MODLR estimate.

3.2. A Monte Carlo simulation study

We generated independently and identically distributed sample $\{(x_i, y_i), i = 1, \ldots, n\}$ from the following model:

$$Y = 1 + 3X + \sigma(X)\varepsilon,$$

where $X \sim U(0, 1)$, $\varepsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ and $\sigma(X) = 1 + 2X$. Note that $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Mode}(\varepsilon) = 1$ and $\mathrm{Median}(\varepsilon) = 0.67$ (the last two quantities are approximate). For this model, the conditional mean regression function is $\mathrm{E}(Y \mid X) = 1 + 3X$, and the conditional modal regression function is $\mathrm{Mode}(Y \mid X) = 2 + 5X$. The modal regression residual is $Y - \mathrm{Mode}(Y|X) = (1+2X)(\varepsilon - 1)$, whose distribution has mode at 0 but is negatively skewed. The conditional median function is $\mathrm{Median}(Y|X) = 1.67 + 4.34X$. We consider and compare the regression parameter estimates by the following four methods: (i) traditional mean regression based on the LSE; (ii) MEDREG; (iii) MM-estimate (M-estimate with a initial robust M-estimate of scale, Yohai, , 1987) based on Tukey's biweight ψ-function; and (iv) the proposed MODLR.

Figure 1 shows the scatter plot of a generated sample with $n = 200$, as well as regression lines corresponding to the four regression methods. From the plot, we can see that the modal regression line goes through the area containing the most number of points. A small prediction band around this line is expected to contain the most number of future points. In contrast, the mean regression line based on LSE is *skewed* to a flatter line and lies in a much less dense area for capturing the conditional mean. The regression lines based on the MEDREG and the MM-estimate lie in higher density areas than the regression line based on LSE.

Table 1 reports the average and standard error (Std) of the parameter estimates for each method on the basis of 1000 replicates. From this table, we see that LSE, MEDREG and MODLR estimate their target parameters well. However, the MM-estimate does not estimate the conditional mean function well; this is because the assumption of symmetric error density is violated. Surprisingly, the MODLR has smaller Std than the other methods in this example (when the error is skewed), especially when $n = 200$ or $n = 400$. Therefore, for finite samples,

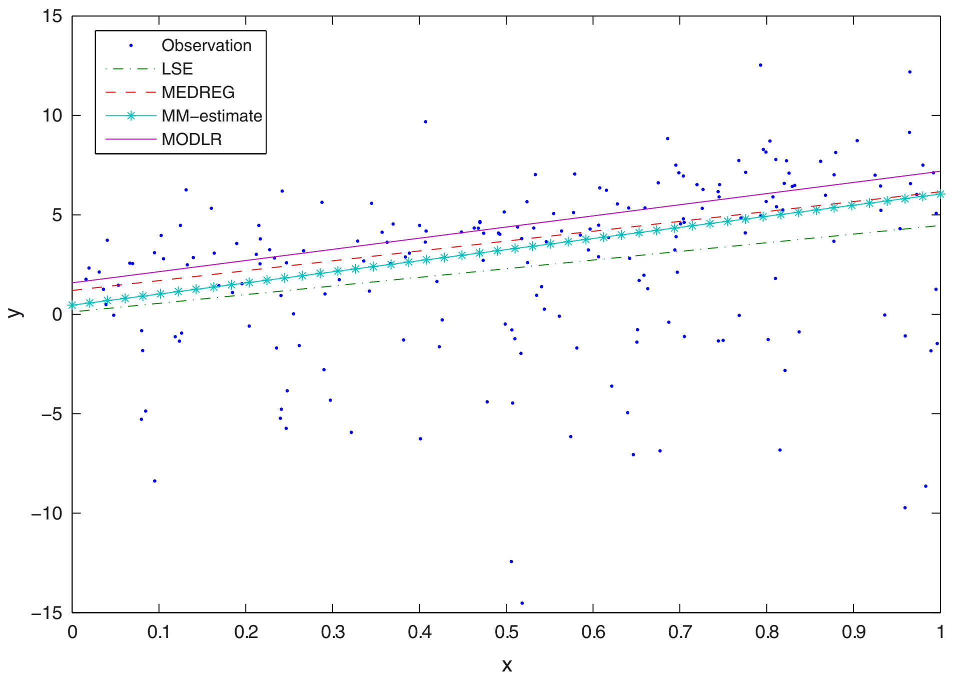

Fig. 1. Scatter plot of a typical sample with $n = 200$ for Example 1 with different estimated regression lines: '—.' denotes the mean regression line based on LSE; '——' denotes the median regression line; '—∗' denotes the regression line based on MM-estimate; and '—' denotes the modal regression line.

Table 1. *Average (standard error) of parameter estimates over 1000 repetitions*

Method	Parameter	$n = 50$	$n = 100$	$n = 200$	$n = 400$
LSE	$\beta_0 = 1$	1.022(0.964)	0.989(0.659)	1.007(0.490)	1.009(0.322)
	$\beta_1 = 3$	2.890(2.260)	3.063(1.500)	2.977(1.160)	2.976(0.733)
MEDREG	$\beta_0 = 1.67$	1.587(0.707)	1.613(0.422)	1.636(0.301)	1.667(0.188)
	$\beta_1 = 4.34$	4.226(1.670)	4.372(0.981)	4.339(0.705)	4.312(0.457)
MM-estimate	$\beta_0 = 1$	1.051(0.782)	1.040(0.530)	1.022(0.376)	1.035(0.265)
	$\beta_1 = 3$	5.123(1.640)	5.234(1.060)	5.271(0.744)	5.271(0.512)
MODLR	$\beta_0 = 2$	1.789(0.670)	1.841(0.372)	1.875(0.229)	1.912(0.140)
	$\beta_1 = 5$	4.829(1.750)	5.024(0.948)	5.044(0.574)	5.020(0.387)

LSE, least squares estimate; MEDREG, median regression; MODLR, modal linear regression.

the MODLR not only has a good modal explanation but also might have better estimation accuracy than other methods when the error is skewed.

Table 2 reports the average (and Std) of the coverage probabilities of prediction intervals of similar lengths centred around each estimated regression line in 1000 replicates. We consider three different lengths of intervals: $0.1\sigma, 0.2\sigma$ and 0.5σ, where $\sigma = 2$ is the approximate Std of ε. For each of the 1000 replications, the coverage probability is estimated from 1000 new cases where the predictor $\boldsymbol{x}$ is equally spaced between 0.1 and 0.9. From Table 2, we see that MODLR provides higher coverage probabilities than the other three methods. In addition, MEDREG provides larger coverage probabilities than the MM-estimate and LSE, whereas the MM-estimate provides larger coverage probabilities than LSE. Note that when the lengths of these intervals are large enough, the different methods will provide similar coverage probabilities.

Table 2. *Average (standard error) of coverage probabilities over 1000 repetitions with $\sigma = 2$*

Width	Method	$n = 50$	$n = 100$	$n = 200$	n=400
0.1σ	LSE	0.034(0.015)	0.032(0.011)	0.030(0.009)	0.029(0.007)
	MEDREG	0.073(0.018)	0.077(0.014)	0.078(0.012)	0.080(0.010)
	MM-estimate	0.065(0.023)	0.067(0.019)	0.066(0.015)	0.067(0.012)
	MODLR	0.087(0.016)	0.092(0.012)	0.095(0.010)	0.095(0.009)
0.2σ	LSE	0.069(0.028)	0.065(0.022)	0.061(0.015)	0.059(0.013)
	MEDREG	0.144(0.033)	0.153(0.024)	0.155(0.019)	0.158(0.015)
	MM-estimate	0.129(0.042)	0.133(0.034)	0.132(0.027)	0.134(0.021)
	MODLR	0.170(0.027)	0.179(0.018)	0.184(0.013)	0.186(0.012)
0.5σ	LSE	0.186(0.062)	0.181(0.047)	0.174(0.035)	0.171(0.028)
	MEDREG	0.338(0.061)	0.355(0.040)	0.360(0.029)	0.365(0.022)
	MM-estimate	0.313(0.080)	0.322(0.062)	0.325(0.046)	0.330(0.036)
	MODLR	0.378(0.049)	0.395(0.029)	0.404(0.018)	0.407(0.015)

LSE, least squares estimate; MEDREG, median regression; MODLR, modal linear regression.

3.3. Application to forest fire data

Forest fires, also called wildfires, cause great ecological and economical damage. Fast detection of a forest fire is vital for successful fire fighting, but traditional human or automatic surveillance (such as by satellites, infrared or smoke scanners) is expensive. Recently the use of low-cost meteorological data (such as temperature, wind and precipitation data) to warn the public of a potential wildfire has received a lot of attention. This inexpensive form of information can also be used to get a quick estimate of postfire damage.

In this section, we compare the proposed MODLR and other regression techniques with a forest fire dataset (Cortez & Morais, 2007). The data were downloaded from http://www.dsi.uminho.pt/~pcortez/forestfires. These forest fire data contain 517 observations and were collected between January 2000 and December 2003 from the Montesinho Natural Park of the Trás-os-Montes northeast region of Portugal. On a daily basis, every time a forest fire occurred, many features were recorded, such as the time, date, spatial location and weather conditions. Following Cortez & Morais (2007), we use four meteorological variables: outside temperature (temp), outside relative humidity (RH), outside wind speed (wind) and outside rain (rain), as predictors for the total burned area (area). We fit the data by LSE, MEDREG, MM-estimate and MODLR. One important feature of this dataset is that it contains outliers and a positively-skewed response variable (area); therefore, it is expected that the proposed MODLR will compare favourably with the mean regression.

To compare the four regression methods, we look at the widths of each prediction interval (with the same confidence level). For constructing confidence intervals, we assume that the error distribution of ϵ is independent of $\boldsymbol{x}$. Suppose we have obtained the parameter estimate $\hat{\boldsymbol{\beta}}$ and the corresponding error (residual) $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$, we will use $\hat{\epsilon}_{[i]}$ to denote the ith smallest value of the residuals. The traditional prediction interval with confidence level α for the new predictor $\boldsymbol{x}_{new}$ is symmetric about the point prediction of y_{new}: $(\boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}} - \hat{\epsilon}_{[n_1]}, \boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[n_2]})$, where $n_1 = \lfloor n\alpha/2 \rfloor$ and $n_2 = n - n_1$. This symmetric method will be ideal if the regression error distribution is symmetric. To consider and make use of the skewness of the error distribution, we propose to construct *asymmetric prediction intervals* as follows. Suppose $\hat{g}(\cdot)$ is the kernel density estimate of ϵ based on the residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ that are estimated by MODLR. We propose to find the indexes $k_1 < k_2$ such that $k_2 - k_1 = n_2 - n_1 = \lceil n(1-\alpha) \rceil$ and $\hat{g}(\hat{\epsilon}_{[k_1]}) \approx \hat{g}(\hat{\epsilon}_{[k_2]})$. The proposed prediction interval for the new predictor $\boldsymbol{x}_{new}$ is $(\boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}} - \hat{\epsilon}_{[k_1]}, \boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[k_2]})$. We propose the following iterative algorithm to find indexes k_1 and k_2: Let $k_1 = n_1$ and $k_2 = n_2$ be the initial values for k_1 and k_2.

Table 3. *Average widths (coverage rates) of the prediction intervals*

Methods	Nominal confidence levels			
	10%	30%	50%	90%
LSE	2.166(0.101)	6.687(0.294)	12.70(0.493)	53.03(0.896)
MEDREG	0.975(0.091)	2.638(0.292)	6.506(0.491)	48.52(0.894)
MM-estimate	1.144(0.099)	2.910(0.294)	6.499(0.497)	48.49(0.906)
MODLR	0.012(0.112)	0.035(0.311)	0.571(0.499)	26.44(0.899)

LSE, least squares estimate; MEDREG, median regression; MODLR, modal linear regression.

Step 1. If $\hat{g}(\hat{\epsilon}_{[k_1]}) < \hat{g}(\hat{\epsilon}_{[k_2]})$ and $\hat{g}(\hat{\epsilon}_{[k_1+1]}) < \hat{g}(\hat{\epsilon}_{[k_2+1]})$, $k_1 \leftarrow k_1 + 1$ and $k_2 \leftarrow k_2 + 1$; if $\hat{g}(\hat{\epsilon}_{[k_1]}) > \hat{g}(\hat{\epsilon}_{[k_2]})$ and $\hat{g}(\hat{\epsilon}_{[k_1-1]}) > \hat{g}(\hat{\epsilon}_{[k_2-1]})$, $k_1 \leftarrow k_1 - 1$ and $k_2 \leftarrow k_2 - 1$.

Step 2. Iterate the aforementioned procedure until none of the aforementioned two conditions is satisfied or $(k_1 - 1)(k_2 - n) = 0$.

We use this method to construct prediction intervals for MODLR.

In Table 3, we report the average widths and the actual coverage rates of the prediction intervals for 10%, 30%, 50% and 90% confidence levels.

The actual coverage rates are estimated on the basis of leave-one-out cross validation. From Table 3, we have the following findings:

(i) all the prediction intervals are well-calibrated—the actual coverage rates are very close to the nominal confidence levels.
(ii) the average widths of prediction intervals constructed around the point prediction defined by MODLR are significantly shorter than the prediction intervals constructed around the other three estimates.
(iii) both MEDREG and MM-estimate have shorter prediction intervals than LSE.

4. Summary and discussions

In this article, we proposed a new data analysis tool called MODLR in order to explore the relationship between a response variable and a set of predictors. MODLR investigates this relationship using the conditional mode instead of the conditional mean or other summaries used by traditional regression techniques. When the error distribution is skewed, MODLR provides a more meaningful prediction than LSE. Our empirical results show that the MODLR provides significantly shorter prediction intervals than other regression methods.

In the application to the forest fire dataset, we provided one possible way to construct asymmetric prediction intervals for MODLR. On the basis of cross-validation results, the proposed skewed prediction intervals for MODLR were much shorter than the prediction intervals constructed by some of the other commonly used regression methods for forest fire data. Further research can be conducted to find out how to construct the shortest (skewed) prediction interval for a given confidence level using the information of skewed error density. One related work is by Kim & Lindsay (2011), who proposed to use confidence distribution sampling to visualize confidence sets.

Modal linear regression assumes that the mode of the conditional density of Y given $\boldsymbol{x}$ is a linear function of $\boldsymbol{x}$. The idea of MODLR can be easily generalized to other models such as non-linear regression, non-parametric regression and varying coefficient partial linear regression. In addition, it would also be interesting to see how to select the most informative variables on the basis of this modal regression idea. This will comprise our future research work.

Appendix

The following technical conditions are imposed in this section.

(A1) $g^{(v)}(t \mid \boldsymbol{x}), v = 0, 1, 2, 3$ is continuous in a neighbourhood of 0, and $g'(0 \mid \boldsymbol{x}) = 0$ for any $\boldsymbol{x}$.
(A2) $n^{-1}\sum_{i=1}^{n} g''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T = J + o_p(1), n^{-1}\sum_{i=1}^{n} g'''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i = K + o_p(1)$ and $n^{-1}\sum_{i=1}^{n} g(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T = L + o_p(1)$, where $J < 0$, that is, $-J$ is a positive definite matrix.
(A3) $n^{-1}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^4 = O_p(1)$.
0, and $g'(0 \mid \boldsymbol{x}) = 0$ any $\boldsymbol{x}$.

The aforementioned conditions are mild and are fulfilled in many applications. Note that the J, K and L are defined in theorem 2.3. All the results proved in this section also hold if general kernels are used for ϕ in (3) under some mild conditions adopted for traditional kernel density estimator (for example, ϕ is symmetric about 0 and has bounded continuous third derivative. In addition, ϕ has finite second moment with $\int t^2\phi^2(t)dt < \infty$).

Proof of theorem 2.1: Note that

$$\begin{aligned}
\log\left\{Q_h\left(\boldsymbol{\beta}^{(k+1)}\right)\right\} - \log\left\{Q_h\left(\boldsymbol{\beta}^{(k)}\right)\right\} &= \log\left\{\sum_{i=1}^{n}\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+1)}\right)\right\} - \log\left\{\sum_{i=1}^{n}\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)\right\} \\
&= \log\left[\sum_{i=1}^{n}\frac{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+1)}\right)}{\sum_{i=1}^{n}\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}\right] \\
&= \log\left[\sum_{i=1}^{n}\frac{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}{\sum_{i=1}^{n}\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}\frac{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+1)}\right)}{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}\right] \\
&= \log\left[\sum_{i=1}^{n}\pi\left(i \mid \boldsymbol{\beta}^{(k)}\right)\frac{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+1)}\right)}{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}\right]
\end{aligned}$$

On the basis of the Jensen's inequality, we have

$$\log\left\{Q_h\left(\boldsymbol{\beta}^{(k+1)}\right)\right\} - \log\left\{Q_h\left(\boldsymbol{\beta}^{(k)}\right)\right\} \geq \sum_{i=1}^{n}\pi\left(i \mid \boldsymbol{\beta}^{(k)}\right)\log\left\{\frac{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+1)}\right)}{\phi_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)}\right\}.$$

On the basis of the property of M-step in (6), we have

$$\log\left\{Q_h\left(\boldsymbol{\beta}^{(k+1)}\right)\right\} - \log\left\{Q_h\left(\boldsymbol{\beta}^{(k)}\right)\right\} \geq 0,$$

and thus, $Q_h\left(\boldsymbol{\beta}^{(k+1)}\right) \geq Q_h\left(\boldsymbol{\beta}^{(k)}\right)$. □

Proof of theorem 2.2: Note that

$$\phi_h''(t) = h^{-3}\left(\frac{t^2}{h^2} - 1\right)\phi(t/h) \text{ and } \phi_h'(t) = -\frac{t}{h^3}\phi(t/h).$$

Let $a_n = (nh^3)^{-1/2} + h^2$. It is sufficient to show that for any given $\eta > 0$, there exists a large constant c such that

$$P\left\{\sup_{\|\mu\|=c} Q_h\left(\boldsymbol{\beta}_0 + a_n\mu\right) < Q_h\left(\boldsymbol{\beta}_0\right)\right\} \geq 1 - \eta. \tag{16}$$

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Denote

$$K_n \equiv \frac{\partial Q_h(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = -\frac{1}{n}\sum_{i=1}^{n} \phi_h'\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_0\right)\boldsymbol{x}_i \tag{17}$$

$$J_n \equiv \frac{\partial^2 Q_h(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} = \frac{1}{n}\sum_{i=1}^{n} \phi_h''\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_0\right)\boldsymbol{x}_i\boldsymbol{x}_i^T, \tag{18}$$

where $Q_h(\boldsymbol{\beta})$ is defined in (3) and $\boldsymbol{\beta}_0$ is the true parameter value.

On the basis of Taylor expansion and symmetric property of $\phi(t)$, we can get the mean and variance of J_n and K_n:

$$\begin{aligned}
\mathrm{E}(J_n \mid \boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} g''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\{1 + o_p(1)\} = J\{1 + o_p(1)\},\\
\mathrm{Var}(J_n \mid \boldsymbol{x}) &= O_p\left\{\left(nh^5\right)^{-1}\right\},\\
\mathrm{E}(K_n \mid \boldsymbol{x}) &= \frac{h^2}{2n}\sum_{i=1}^{n} g'''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i(1 + o_p(1)) = \frac{h^2}{2}K\{1 + o_p(1)\},\\
\mathrm{Cov}(K_n \mid \boldsymbol{x}) &= \frac{1}{n^2h^3}\nu_2\sum_{i=1}^{n} g(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\{1 + o(1)\} = \frac{1}{nh^3}\nu_2 L\{1 + o_p(1)\},
\end{aligned} \tag{19}$$

where $J = \lim n^{-1}\sum_{i=1}^{n} g''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$, $K = \lim n^{-1}\sum_{i=1}^{n} g'''(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i$ and $L = \lim n^{-1}\sum_{i=1}^{n} g(0 \mid \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$. By default, when calculating the variance of a matrix, we find the variance of each element of the matrix. Using the result $X = \mathrm{E}(X) + O_p(\{\mathrm{Var}(X)\}^{1/2})$, because $nh^5 \to \infty$, $J_n = J + o_p(1)$, notice that

$$\begin{aligned}
Q_h(\boldsymbol{\beta}_0 + a_n\mu) - Q_h(\boldsymbol{\beta}_0) &= a_n K_n^T\mu + \frac{a_n^2}{2}\mu^T J_n\mu\\
&\quad - \frac{a_n^3}{6nh^4}\sum_{i=1}^{n}\phi'''\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*}{h}\right)\left(\boldsymbol{x}_i^T\mu\right)^3\\
&= M_1 + M_2 + M_3,
\end{aligned} \tag{20}$$

where $||u|| = c$ and $||\boldsymbol{\beta}^* - \boldsymbol{\beta}_0|| \leq ca_n$. From (19), we get $K_n = O_p(a_n)$ and hence $M_1 = O_p(a_n^2)$. Note that $M_2 = 0.5a_n^2\mu^T J\mu\{1 + o_p(1)\}$. On the basis of the boundness of $\phi^{(4)}(t)$ and $||\boldsymbol{\beta}^* - \boldsymbol{\beta}_0|| \leq ca_n$, we have

$$\phi'''\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*}{h}\right) = \phi'''\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h}\right)(1 + o_p(1)).$$

Noting that $\phi'''(t) = (3t - t^3)\phi(t)$, on the basis of the Taylor expansion and the symmetric property of $\phi(t)$, we have that

$$\mathrm{E}\left\{\phi'''\left(\frac{Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h}\right)\middle|\,\boldsymbol{x}\right\} = O_p(h^4),\ \mathrm{Var}\left\{\phi'''\left(\frac{Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h}\right)\middle|\,\boldsymbol{x}\right\} = O_p(h). \tag{21}$$

Because $nh^5 \to \infty$, we can prove that $M_3 = o_p(a_n^2)$.

For any $\eta > 0$, we can choose c big enough, such that the second term M_2 dominates the other two terms in (20) with probability $1 - \eta$. Because $J < 0$, $Q_h(\boldsymbol{\beta}_0 + a_n\mu) - Q_h(\boldsymbol{\beta}_0) < 0$ with probability $1 - \eta$. The result of theorem 2.2 follows. □

Proof of theorem 2.3: Suppose $\hat{\boldsymbol{\beta}}$ is the consistent solution to $\partial Q_h(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ found in theorem 2.2. On the basis of the Taylor expansion, we have

$$0 = \frac{\partial Q_h(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}} = K_n + (J_n + L_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{22}$$

where

$$L_n = -\frac{1}{2nh^4}\sum_{i=1}^{n}\left[\phi'''\left(\frac{Y_i - \boldsymbol{\beta}^{*T}\boldsymbol{x}_i}{h}\right)\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\boldsymbol{x}_i\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T\right],$$

where $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| \le \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$.

On the basis of the result of (22), we have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (J_n + L_n)^{-1}K_n$. Because $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(a_n)$, where $a_n = (nh^3)^{-1/2} + h^2$, similar to the proof of M_3 in (20), we have $L_n = o_p(1)$. Hence, on the basis of (19), we have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = J^{-1}K_n(1 + o_p(1))$. Next, we prove the asymptotic normality for $K_n^* = \sqrt{nh^3}K_n$.

For any unit vector $d \in \mathbb{R}^{p+1}$, we prove

$$\left\{d^T\mathrm{Cov}(K_n^*)d\right\}^{-\frac{1}{2}}\left\{d^TK_n^* - d^T\mathrm{E}(K_n^*)\right\} \xrightarrow{D} N(0,1)$$

Let

$$\xi_i = -\frac{1}{\sqrt{nh}}\phi'\left(\frac{Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0^T}{h}\right)d^T\boldsymbol{x}_i.$$

Then $d^TK_n^* = \sum_{i=1}^n \xi_i$. We check the Lyapunov's condition. On the basis of the results (19), we know

$$\mathrm{Cov}(K_n) = \frac{L}{nh^3}\nu_2\{1 + o(1)\}. \tag{23}$$

Hence, $\mathrm{Var}(d^TK_n^*) = nh^3d^T\mathrm{Cov}(K_n)d = g(0)\nu_2d^TLd + o(1)$. So we only need to prove $n\mathrm{E}|\xi_1|^3 \to 0$. Noticing that $(d^T\boldsymbol{x}_i)^2 \le \|\boldsymbol{x}_i\|^2\|d\|^2 = \|\boldsymbol{x}_i\|^2$ and $\phi'(\cdot)$ is bounded, we have $n\mathrm{E}|\xi_1|^3 \le O\{(nh^3)^{-1/2}\} \to 0$. So, the asymptotic normality for K_n^* holds, that is, we have

$$\sqrt{nh^3}\left\{K_n - h^2K/2(1 + o_p(h^2))\right\} \xrightarrow{D} N(0, \nu_2L).$$

On the basis of the Slutsky's theorem, we have

$$\sqrt{nh^3}\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \frac{h^2}{2}J^{-1}K\{1 + o_p(h^2)\}\right] \xrightarrow{D} N\left\{0, \nu_2J^{-1}LJ^{-1}\right\}.$$

□

Proof of theorem 2.4: Because $\hat{\boldsymbol{\beta}}_s$ has root n consistency, the asymptotic result of $\hat{\beta}_0$ is the same as if $\hat{\boldsymbol{\beta}}_s$ were known, and its asymptotic distribution can be derived from theorem 2.3 by assuming $\boldsymbol{x} = 1$ and the independence of ϵ and $\boldsymbol{x}$, under which we have $J^{-1}K = g'''(0)/(2g''(0))$ and $J^{-1}LJ^{-1} = g''(0)^{-2}g(0)$. Then the result follows. □

Proof of theorem 2.5: Let $\phi^*(t) = \sqrt{2\pi}h\phi_h(t)$. Then $M = \sum_{i=1}^n \phi^*(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}} = T(\boldsymbol{Z})$. Notice that $\phi^*(\cdot)$ has a maximum at 0 with $\phi^*(0) = 1$ and $\phi^*(\cdot)$ decreases monotonely toward both sides and that $\lim\phi^*(t) = 0$ for $|t| \to \infty$.

We first prove that $T(\boldsymbol{Z} \cup \boldsymbol{Z}')$ stays bounded if $m < M$. Let $\xi > 0$ be such that $m + n\xi < M$, and let C be such that $\phi^*(t) \le \xi$ for $|t| \ge C$. Let $\boldsymbol{\beta}$ be any real vector such that $|y - \boldsymbol{x}^T\boldsymbol{\beta}| \ge C$ for all $z = (\boldsymbol{x}, y)$ in $\boldsymbol{Z}$. Then

$$\sum_{i=1}^{m+n} \phi^* \left(y_i - \boldsymbol{x}_i^T T(\boldsymbol{Z}) \right) \ge M \tag{24}$$

and

$$\sum_{i=1}^{m+n} \phi^* \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right) \le n\xi + m. \tag{25}$$

From (24) and (25), one knows that $T(\boldsymbol{Z} \cup \boldsymbol{Z}')$ must satisfy $|y - \boldsymbol{x}^T T(\boldsymbol{Z} \cup \boldsymbol{Z}')| < C$ for a point in $\boldsymbol{Z}$, and thus, $T(\boldsymbol{Z} \cup \boldsymbol{Z}')$ is bounded.

On the other hand, if $m > M$, let $\xi > 0$ such that $m - m\xi > M$, and let C be such that $\phi^*(t) \le \xi$ for $|t| \ge C$. Assume that all points $\{(\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_{n+m}, y_{n+m})\}$ in $\boldsymbol{Z}'$ are the same and satisfy a linear relationship $y = \boldsymbol{x}^T\beta^*$. Let $\boldsymbol{\beta}$ be any vector such that $|y_{n+1} - \boldsymbol{x}_{n+1}^T\beta| < C$. Then

$$\sum_{i=1}^{m+n} \phi^* \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right) \le M + m\xi, \tag{26}$$

and

$$\sum_{i=1}^{m+n} \phi^* \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^* \right) \ge m. \tag{27}$$

From (26) and (27), one knows that $T(\boldsymbol{Z} \cup \boldsymbol{Z}')$ must satisfy $|y_{n+1} - \boldsymbol{x}_{n+1}^T T(Z \cup Z')| \le C$. If we let $y_{n+1} \to \infty$ with $\boldsymbol{x}_{n+1}$ fixed, $||T(\boldsymbol{Z} \cup \boldsymbol{Z}')||$ must go off to infinity, and we have breakdown. □

Acknowledgements

The authors would like to thank the editor, an associate editor and the anonymous referees for their valuable comments and suggestions, which led to substantial improvements in the manuscript.

The research of Longhai Li is supported by fundings from the Natural Sciences and Engineering Research Council of Canada and Canadian Foundation of Innovations.

References

Botev, Z. I., Grotowski, J. F. & Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Ann. Statist.* **38**, 2916–2957.

Chaudhuri, P. & Marron, J. S. (1999). Sizer for exploration of structures in curves. *J. Amer. Statist. Assoc.* **94**, 807–823.

Cortez, P. & Morais, A. (2007). A data mining approach to predict forest fires using meterological data. In *Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, 512–523, December, Guimaraes, Portugal.

Davies, P. L. & Kovac, A. (2004). Densities, spectral densities and modality. *Ann. Statist.* **32**, 1093–1136.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*, Ph.D. Qualifying paper, Department of Statistics, Harvard University.

Donoho, D. L. & Huber, P. J. (1983). The notion of breakdown point. In *Festchrift in Honor of Erich Lehman* (eds Doksum, K. & Hodges, J. L.), Wadsworth, Belmont, CA; 157–184.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* **8**, 870–882.

Fisher, N. I. & Marron, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika* **88**, 499–517.

Friedman, J. H. & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* **9**, 123–143.

Hall, P., Minnotte, M. C. & Zhang, C. (2004). Bump hunting with non-Gaussian kernels. *Ann. Statist.* **32**, 2124–2141.

Hampel, F. R. (1971). A general quantitative definition of robustness. *Ann. Math. Statist.* **42**, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 909–927.

Huber, P. J. (1981). *Robust statistics*, Wiley, New York.

Huber, P. J. (1984). Finite sample breakdown points of M- and P-estimators. *Ann. Statist.* **12**, 119–126.

Kim, D. & Lindsay, B. G. (2011). Using confidence distribution sampling to visualize confidence sets. *Statist. Sinica* **21**, (2), 923–948.

Lee, M. J. (1989). Mode regression. *J. Econometrics* **42**, 337ndash 349.

Li, J., Ray, S. & Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8**, (8), 1687–1723.

Muller, D. W. & Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Ann. Math. Statist.* **33**, 1065–1076.

Ray, S. & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Ann. Statist.* **33**, 2042–2065.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice and visualization*, Wiley, New York.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Yao, W. (2013). A note on EM algorithm for mixture models. *Stat. Probabil. Lett.* **83**, 519–526.

Yao, W. & Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.* **104**, 758–767.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Ann. Statist.* **15**, 642–656.

Received 11 September 2011, in final form 15 September 2013

Weixin Yao, Department of Statistics, Kansas State University, Manhattan, Kansas 66506, USA.
E-mail: wxyao@ksu.edu