

# Modal Regression for Skewed, Truncated, or Contaminated Data with Outliers



Sijia Xiang and Weixin Yao

**Abstract** Built on the ideas of mean and quantile, mean regression and quantile regression are extensively investigated and popularly used to model the relationship between a dependent variable  $Y$  and covariates  $\mathbf{x}$ . However, the research about the regression model built on the mode is rather limited. In this article, we introduce a new regression tool, named modal regression, that aims to find the most probable conditional value (mode) of a dependent variable  $Y$  given covariates  $\mathbf{x}$  rather than the mean that is used by the traditional mean regression. The modal regression can reveal new interesting data structure that is possibly missed by the conditional mean or quantiles. In addition, modal regression is resistant to outliers and heavy-tailed data and can provide shorter prediction intervals when the data are skewed. Furthermore, unlike traditional mean regression, the modal regression can be directly applied to the truncated data. Modal regression could be a potentially very useful regression tool that can complement the traditional mean and quantile regressions.

**Keywords** Modal regression · Mode · Skewed data

## 1 Introduction

When talking about location measurements of a data set or distribution, mean, quantile and mode are most commonly used. They have their own merits and complement each other. Up till now, mean and quantile regressions have been extensively studied and popularly used to model the relationship between a response  $Y$  and covariates  $\mathbf{x}$ . However, there is not much research about the regression built on

---

S. Xiang

School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China  
e-mail: [sjxiang@zufe.edu.cn](mailto:sjxiang@zufe.edu.cn)

W. Yao (✉)

Department of Statistics, University of California at Riverside, Riverside, CA, USA  
e-mail: [weixin.yao@ucr.edu](mailto:weixin.yao@ucr.edu)

the mode (i.e., modal regression). Different from mean/quantile regression, modal regression is another important tool to study the relationship between a response  $Y$  given a set of predictors  $\mathbf{x}$ , which estimates the conditional modes of  $Y$  given  $\mathbf{x}$ . The developed new regression tool complements the mean and quantile regression and is especially useful for skewed and truncated data and has broad applicability throughout science, such as economics, sociology, behavior, medicine, and biology.

Indeed, the skewed data or truncated data can be commonly found in many applications. For example, Cardoso and Portugal (2005) stated that wages, prices, and expenditures are typical examples of skewed data. In sociology, Healy and Moody (2014) showed that “many of the distributions typically studied in sociology are extremely skewed,” for example, church sizes in sociology of religion (Weber 1993), symptoms indices in sociology of mental health (Mirowsky 2013), and so on. Besides, truncated data can be commonly found in many applications such as econometrics (Amemiya 1973; Lewbel & Linton 2002; Park et al. 2008) when dependent variable is an economic index measured within some range. Some examples of truncated data are a sample of Americans whose income is above the poverty line, military height records with a minimum height requirement in many armies, a central bank intervenes to stop an exchange rate falling below or going above certain levels.

We use the following example (Yao & Li 2014) to demonstrate the difference between the modal regression and the mean regression.

*Example 1* Let  $(\mathbf{x}, Y)$  be coming from the model Use the standard equation environment to typeset your equations, e.g.,

$$Y = m(\mathbf{x}) + \sigma(\mathbf{x})\epsilon, \quad (1)$$

where  $\epsilon$  has a density  $q(\cdot)$ , which is a skewed density with mean 0 and mode 1.

1. If  $m(\mathbf{x}) = 0$  and  $\sigma(\mathbf{x}) = \mathbf{x}^\top \alpha$ , then

$$E(Y|\mathbf{x}) = 0 \quad \text{and} \quad \text{Mode}(Y|\mathbf{x}) = \mathbf{x}^\top \alpha.$$

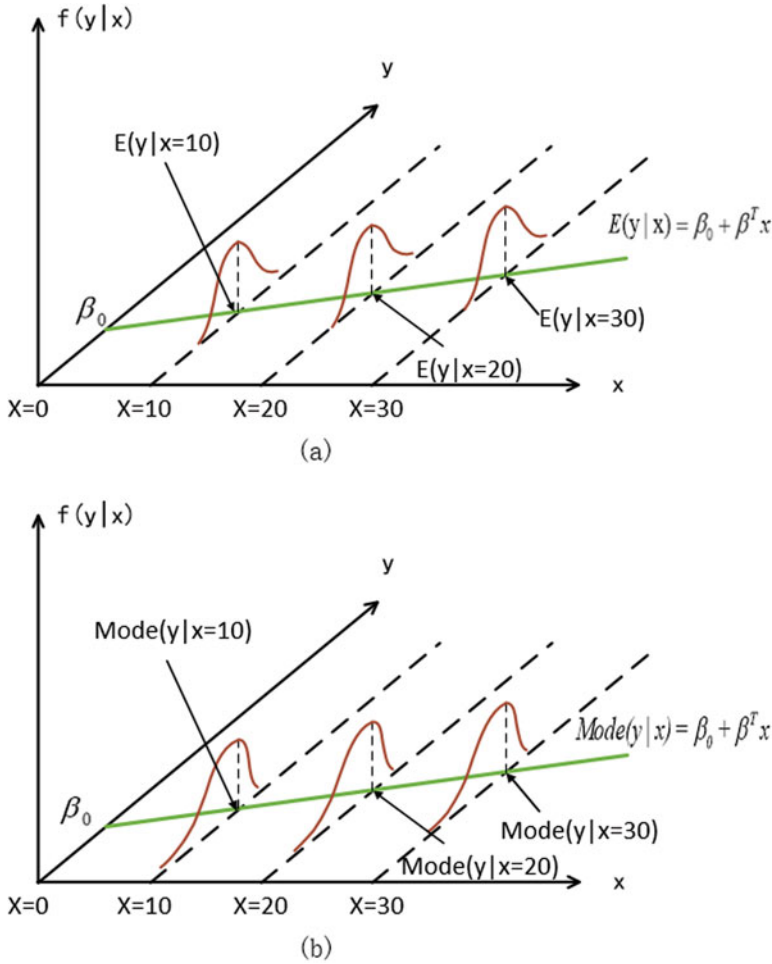
That is to say, the conditional mean does not contain any information of the covariate, while the conditional mode does. As a result, various techniques based on modal regression could reveal more important covariates than conditional mean.

2. If  $\sigma(\mathbf{x}) = \mathbf{x}^\top \alpha - m(\mathbf{x})$  and  $m(x)$  is a nonlinear smooth function, then

$$E(Y|\mathbf{x}) = m(\mathbf{x}) \quad \text{and} \quad \text{Mode}(Y|\mathbf{x}) = \mathbf{x}^\top \alpha.$$

In this case, the conditional mode is linear in  $\mathbf{x}$  while the conditional mean does not. Of course, the opposite situation could also happen.

In Fig. 1, we also use two plots to illustrate the difference between linear mean and linear mode regression.



**Fig. 1** Mean regression vs mode regression

Many authors have made efforts to identify the modes in the one sample problem. See, for example, Parzen (1962); Scott (1992); Friedman and Fisher (1999); Chauduri and Marron (1999); Hall et al. (2004); Ray and Lindsay (2005); Yao and Lindsay (2009); Henderson and Russell (2005); Henderson et al. (2008); Henderson and Parmeter (2015). Modal hunting has received much interest and wide applications in economy and econometrics too. For example, Henderson and Russell (2005) applied a nonparametric production frontier model to show that international polarization (shift from a uni-modal to a bimodal distribution) is brought primarily by technological catch-up. Cardoso and Portugal (2005) studied the impact of union bargaining power and the degrees of employer coordination on the wage distribution in Portugal wage computed by the mode of the contractual wage set by collective

bargaining. Henderson et al. (2008) applied recent advances from statistics literature to test for unconditional multimodality of worldwide distributions of several (unweighted and population-weighted) measures of labor productivity, which is of great interest in economics. They also examined the movements of economies between modal clusters and relationships between certain key development factors and multimodality of the productivity distribution. Einbeck and Tutz (2006) used the value(s) maximizing the conditional kernel density estimate as estimator(s) for the conditional mode(s), and proposed a plug-in estimator using kernel density estimator.

Most of the above modal hunting methods require first nonparametrically estimating the joint density  $f(\mathbf{x}, y)$  and  $f(y | \mathbf{x})$ , and then estimating the mode based on the estimated conditional density  $f(y | \mathbf{x})$ , which is practically challenging when the dimension of  $\mathbf{x}$  is large due to the well-known "curse of dimensionality." Motivated by the result that the conditional mode from the truncated data provides consistent estimates of the conditional mean for the original non-truncated data, Lee (1989) proposed to model  $\text{Mode}(y|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$  and derived the mode regression estimator. The identification of  $\boldsymbol{\beta}$  and strong consistency of its estimator were derived. However, the objective function used by Lee (1989) is based on kernels with bounded support and thus is difficult to implement in practice. This might explain why modal regression has not drawn too much attention in the last century. In addition, the tuning parameter  $h$  used by Lee (1989) is fixed and does not depend on the sample size  $n$ . Therefore, it requires the error to be symmetric to get the consistent modal line. Note that in such cases the modal line is indeed the same as the mean regression line and thus their modal regression estimator is essentially a type of robust regression estimate under the assumption of symmetric error density. This limitation of requiring a symmetric error density also applies to the nonparametric modal regression proposed by Yao et al. (2012).

Kemp and Santos Silva (2012) and Yao and Li (2014) are among the first who proposed consistent linear modal regression estimates without requiring a symmetric error density. They established asymptotic properties of the proposed modal estimates, under very general conditions, allowing a skewed error density and a more general kernel function, by letting the bandwidth  $h$  go to zero. Since the work of Kemp and Santos Silva (2012) and Yao and Li (2014), modal regression has received much attention recently and been widely applied to various problems. Chen et al. (2016) considered a nonparametric modal regression and used it to build confidence sets based on a kernel density estimate of the joint distribution. Zhou and Huang (2016) considered estimating local modes of the food frequency questionnaire (FFQ) intake given one's long-term usual intake using dietary data. Noticing that the neuroimaging features and cognitive assessment are often heavy-tailed and skewed, Wang et al. (2017) argued that a traditional regression approach might fail to capture the relationship, and applied a regularized modal regression to predict for Alzheimer's disease. Yao and Li (2014) also applied the modal linear regression to a forest fire data, and the results showed that the modal regression gave shorter predictive intervals than traditional methodologies. In order to accurately forecast the energy that will be consumed in the evening, so as to optimize the

capacity of storage and consequently to increase the batteries life, Chaouch et al. (2017) applied modal regression to analyze electricity consumption. Kemp et al. (2019) applied both mode- and mean-based autoregressive models to compare the estimates and forecasts of monthly US data on inflation and personal income growth. Please also see Krief (2017); Chen (2018); Li and Huang (2019); Ota et al. (2019); Feng et al. (2020) for some other extensions of the linear modal regression. Ullah et al. (2021) extended the modal regression to the panel data setting.

The rest of the article is organized as follows. In Sect. 2, we formally define the linear modal regression model and discuss its estimator. In Sect. 3, we introduce the nonparametric modal regression. The semiparametric modal regression, which combines the linear modal regression and nonparametric modal regression, is introduced in Sect. 4. A discussion section with some possible future works are presented in Sect. 5.

## 2 Linear Modal Regression

### 2.1 Introduction of Linear Modal Regression

Suppose  $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$  is a random sample, where  $\mathbf{x}_i$  is a  $p$ -dimensional column vector, and  $f(y|\mathbf{x})$  is the conditional density function of  $Y$  given  $\mathbf{x}$ . In conventional regression models, the mean of  $f(y|\mathbf{x})$  is used to investigate the relationship between  $Y$  and  $\mathbf{x}$ . However, when the conditional density of  $Y$  given  $\mathbf{x}$  is skewed, truncated, or contaminated data with outliers, the conditional mean may not provide a good representation of the  $\mathbf{x}$ - $Y$  relationship. In this scenario, it is well-known that the mode provides a more meaningful location estimator than the mean. Therefore, the modal regression model is more preferable in this scenario.

The traditional modal estimation is to first estimate the joint density  $f(\mathbf{x}, y)$  based on kernel density estimation and then derive the conditional density  $f(y|\mathbf{x})$  and its conditional mode. Such method works reasonably well when the dimension of  $\mathbf{x}$  is low, however, it is practically infeasible when the dimension of  $\mathbf{x}$  is large, due to the ‘‘curse of dimensionality’’.

Borrowing the idea from linear mean/quantile regression, Kemp and Santos Silva (2012) and Yao and Li (2014) proposed linear modal regression (LMR), which assumes that the mode of  $f(y|\mathbf{x})$  is a linear function of  $\mathbf{x}$ . Suppose that the mode of  $f(y|\mathbf{x})$  is unique, and denote it by

$$\text{Mode}(Y|\mathbf{x}) = \arg \max_y f(y|\mathbf{x}),$$

then, the LMR assumes that  $\text{Mode}(Y|\mathbf{x})$  is a linear function of  $\mathbf{x}$ , that is,

$$\text{Mode}(Y|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}, \quad (2)$$

where the first element of  $\mathbf{x}$  is assumed to be 1 to represent the intercept. Denote the error term as  $\epsilon = y - \mathbf{x}^\top \boldsymbol{\beta}$ , and let  $q(\epsilon|\mathbf{x})$  to be the conditional distribution of  $\epsilon$  given  $\mathbf{x}$ , which is referred to as the error distribution. Note that we allow the error distribution to depend on  $\mathbf{x}$ . Based on the model assumption of (2), the error density  $q(\epsilon|\mathbf{x})$  has the mode at 0.

Unlike one sample mode estimators, the proposed linear modal regression (Yao and Li 2014) puts some model assumptions on  $\text{Mode}(Y|\mathbf{x})$  to transform the original multivariate problems to a much simpler one-dimensional problem and thereby avoid directly estimating the conditional density  $f(y|\mathbf{x})$ . Note that if the error distribution  $q(\epsilon|\mathbf{x})$  is symmetric, then  $\boldsymbol{\beta}$  in (2) is nothing but the regression coefficient in traditional linear regression model. However, if  $q(\epsilon|\mathbf{x})$  is skewed or heavy-tailed, then, (2) will be quite different from the conventional mean regression model.

Next we explain how we can use a kernel based objective function to estimate the modal regression parameter  $\boldsymbol{\beta}$  in (2) consistently. Note that if  $\boldsymbol{\beta} = \beta_0$  is a scalar, then  $\beta_0$  is the mode of  $f(y)$ , i.e., 0 is the mode of  $f(y - \beta_0)$ . Therefore,  $\beta_0$  can be estimated by the maximizer of

$$Q_h(\beta_0) = \frac{1}{n} \sum_{i=1}^n \phi_h(y_i - \beta_0), \tag{3}$$

which is a kernel density estimate of  $f(y)$ , where  $\phi_h(\cdot) = h^{-1}\phi(\cdot/h)$  with  $\phi(\cdot)$  being a kernel density function symmetric about 0 and  $h$  being a tuning parameter. Such a modal estimator has been proposed by Parzen (1962). It has been proved that as  $n \rightarrow \infty$  and  $h \rightarrow 0$ , the mode of kernel density function will converge to the mode of the distribution of  $Y$ .

If  $\boldsymbol{\beta}$  does include predictors like in the model (2), by extending the objective function (3), we can then estimate  $\boldsymbol{\beta}$  by maximizing

$$Q_h(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \phi_h(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}), \tag{4}$$

which can be also considered as the kernel density estimate of the residual  $\epsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$  at 0. Then, maximizing (4) with respect to  $\boldsymbol{\beta}$  yields  $\mathbf{x}^\top \hat{\boldsymbol{\beta}}$  so that the kernel density function of  $\epsilon_i$  at 0 is maximized. It has been proved by Yao and Li (2014) that as  $h \rightarrow 0$  as  $n \rightarrow \infty$ , the maximizer of (4), named the linear modal regression estimator (LMRE), is a consistent estimate of  $\boldsymbol{\beta}$  in (2) for very general error density without requiring symmetry.

Note that if  $\phi_h(t) = (2h)^{-1}I(|t| \leq h)$ , a uniform kernel, then maximizing (4) is equivalent to maximizing

$$\frac{1}{n} \sum_{i=1}^n I(|y_i - \mathbf{x}_i^\top \boldsymbol{\beta}| \leq h) = \frac{1}{n} \sum_{i=1}^n I(\mathbf{x}_i^\top \boldsymbol{\beta} - h \leq y_i \leq \mathbf{x}_i^\top \boldsymbol{\beta} + h).$$

Therefore, the LMR tries to find the linear regression  $\mathbf{x}^\top \hat{\boldsymbol{\beta}}$  such that the band  $\mathbf{x}^\top \hat{\boldsymbol{\beta}} \pm h$  contains the largest proportion/number of response  $y_i$ . Therefore, modal regression provides more meaningful point predictions, i.e., larger coverage probability of prediction intervals with a fixed small window around the estimate, and shorter predication intervals than the mean and quantile regression for a fixed confidence limit.

### 2.2 Asymptotic Properties

In this section, the consistency, convergence rate and asymptotic distribution of the LMR estimator (Kemp & Santos Silva 2012; Yao & Li 2014) are discussed.

**Theorem 1** *As  $h \rightarrow 0$  and  $nh^5 \rightarrow \infty$ , and under the regularity conditions (A1)–(A3) given in the Appendix, there exists a consistent maximizer of (4) such that*

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

and the asymptotic distribution of the estimator is

$$\sqrt{nh^3} \left[ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \frac{h^2}{2} J^{-1} K \{1 + o_p(1)\} \right] \xrightarrow{D} N\{0, v_2 J^{-1} L J^{-1}\},$$

where  $\boldsymbol{\beta}_0$  denotes the true coefficient of (4),  $v_2 = \int t^2 \phi^2(t) dt$  with  $\phi(\cdot)$  being the standard normal density and  $q(\cdot)$  is the density of the error term.

$$J = E\{q''(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\}, \quad K = E\{q'''(0|\mathbf{x}_i)\mathbf{x}_i\}, \quad L = E\{q(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\}. \tag{5}$$

Readers are referred to Yao and Li (2014) for the proofs. One striking but reasonable finding is that the convergence rate of modal regression estimator is slower than the root- $n$  convergence rate of traditional mean/median regression estimators. That is the cost we need to pay in order to estimate the conditional mode (Parzen 1962). Note that for the distribution of  $Y$  (without conditioning on  $\mathbf{x}$ ), Parzen (1962) and Eddy (1980) have proven similar asymptotic results for kernel estimators of the mode. Therefore, the results of Parzen (1962) and Eddy (1980) can be considered as special cases of the above theorem with no predictor.

Based on the asymptotic bias and asymptotic variance of  $\hat{\boldsymbol{\beta}}$ , a theoretical optimal bandwidth  $h$  for estimating  $\boldsymbol{\beta}$  is to minimize the asymptotic weighted mean squared errors

$$E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top W(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} \approx (4)^{-1} K^\top J^{-1} W J^{-1} K h^4 + (nh^3)^{-1} v_2 \text{tr}(J^{-1} L J^{-1} W),$$

where  $tr(\cdot)$  denotes the trace and  $W$  is a diagonal matrix, whose elements reflect the importance of the accuracy in estimating different coefficients. As a result, an asymptotic optimal bandwidth  $h$  can be calculated as

$$\hat{h}_{opt} = \left[ \frac{3v_2 \text{tr}(J^{-1}LJ^{-1}W)}{K^\top J^{-1}WJ^{-1}K} \right]^{1/7} n^{-1/7},$$

where  $J$ ,  $K$ , and  $L$  are listed in (5).

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

$$\hat{h}_{opt} = \left[ \frac{3v_2(p+1)}{K^\top L^{-1}K} \right]^{1/7} n^{-1/7}.$$

We can then use a plug-in method (Yao and Li 2014) to choose the bandwidth based on the above results.

Another computationally extensive way to choose the bandwidth is to use a cross validation criterion proposed by Zhou and Huang (2019) for modal regression. In addition, instead of just estimating the conditional mode for a chosen value of  $h$ , Kemp and Santos Silva (2012) proposed estimating the parameters of interest for a wide range of  $h$ , and obtain a more detailed picture of how the parameter estimators perform. The authors further argued that since the inference will not be based on a single value of  $h$ , the choice of the limits of  $h$  is not as critical as the choice of an optimal value of  $h$ .

### 2.3 Estimation Algorithm

Since there is no closed-form solution to maximize (4), a modal expectation-maximization (MEM) algorithm (Yao 2013) is extended to find the maximizer, which consists of an E-step and an M-step. Note that the choice of the kernel function is not crucial, and Yao and Li (2014) used the standard Gaussian kernel to simplify the computation in the M-step of a modal EM (MEM) algorithm.

**Algorithm 2.1** For  $t = 0, 1, \dots$ , at the  $(t + 1)$ -th iteration,

*E-step* For  $i = 1, \dots, n$ , calculate the weight as

$$p(i|\boldsymbol{\beta}^{(t)}) = \frac{\phi_h(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})}{\sum_{j=1}^n \phi_h(y_j - \mathbf{x}_j^\top \boldsymbol{\beta}^{(t)})} \propto \phi_h(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}).$$

*M-step* Update the estimate  $\boldsymbol{\beta}^{(t+1)}$  as



$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \arg \max_{\boldsymbol{\beta}} \sum_{i=1}^n \{p(i|\boldsymbol{\beta}^{(t)}) \log \phi_h(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\} \\ &= (\mathbf{X}^\top \mathbf{W}_t \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_t \mathbf{y}, \end{aligned} \tag{6}$$

where  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ ,  $\mathbf{W}_t$  is an  $n \times n$  diagonal matrix whose diagonal element is  $p(i|\boldsymbol{\beta}^{(t)})$  and  $\mathbf{y} = (y_1, \dots, y_n)^\top$ .

*Remark 2.1*

1. From the above algorithm, we can see that the major difference between the mean regression estimated by the least squares (LSE) criterion and the modal regression lies in the weight  $p(i|\boldsymbol{\beta}^{(t)})$ . For LSE, each observation has equal weight  $1/n$ , while for modal regression, the weight  $p(i|\boldsymbol{\beta}^{(t)})$  depends on how close  $y_i$  is to the modal regression curve. This weight scheme allows the modal regression to reduce the effect of observations far away from the regression curve, so as to achieve robustness.
2. Note that when a normal kernel is used in (4), the function optimized in the M-step is a weighted sum of log-likelihoods corresponding to weighted least squares estimator in the ordinary linear regression. In this case, we obtain a closed-form expression for the maximizer in (6). If other kernels are used, then some optimization algorithms are needed in the M-step.
3. It should be noted that the converged value of this MEM algorithm depends on the starting value. Therefore, it is prudent that we start from several different starting values and choose the best local optima.

### 2.4 Prediction Intervals Based on Modal Regression

As we explained after the objective function (4), the modal regression could provide more representative point predictions and shorter prediction intervals. In this section, we explain how to construct asymmetric prediction intervals for new observations based on the linear modal regression. The described methods can be also applied to other nonparametric or semiparametric modal regression models introduced in Sects. 3 and 4.

For the simplicity of explanation, we assume that the error distribution of  $\epsilon$  is independent of  $\mathbf{x}$ . Let  $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$  be the residuals of the linear modal regression estimate, where  $\hat{\epsilon}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ , and  $\hat{\epsilon}_{[i]}$  be the  $i$ th smallest value of the residuals. The traditionally used prediction interval with confidence level  $1 - \alpha$  for a new covariate  $\mathbf{x}_{new}$  is  $(\mathbf{x}_{new}^\top \hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[n_1]}, \mathbf{x}_{new}^\top \hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[n_2]})$ , where  $n_1 = \lfloor n\alpha/2 \rfloor$ , and  $n_2 = n - n_1$ . This symmetric method works best if the error distribution is symmetric. Since the linear modal regression focuses on the highest conditional density region and does not assume a symmetric error density, Yao and Li (2014) proposed the following

method for modal regression to use the information of the skewed error density to construct prediction intervals. Suppose  $\hat{q}(\cdot)$  is a kernel density estimate of  $\epsilon$  based on the residuals  $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ . We find the indexes  $k_1 < k_2$  such that  $k_2 - k_1 = \lceil n(1 - \alpha) \rceil$  and  $\hat{q}(\hat{\epsilon}_{[k_1]}) \approx \hat{q}(\hat{\epsilon}_{[k_2]})$ . The proposed prediction interval by Yao and Li (2014) for a new covariate  $\mathbf{x}_{new}$  is then  $(\mathbf{x}_{new}^\top \hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[k_1]}, \mathbf{x}_{new}^\top \hat{\boldsymbol{\beta}} + \hat{\epsilon}_{[k_2]})$ .

To find indexes  $k_1$  and  $k_2$ , we could use the following iterative algorithm.

**Algorithm 2.2** Starting from  $k_1 = \lfloor n\alpha/2 \rfloor$  and  $k_2 = n - n_1$ ,

- Step 1: If  $\hat{q}(\hat{\epsilon}_{[k_1]}) < \hat{q}(\hat{\epsilon}_{[k_2]})$  and  $\hat{q}(\hat{\epsilon}_{[k_1+1]}) < \hat{q}(\hat{\epsilon}_{[k_2+1]})$ ,  $k_1 = k_1 + 1$  and  $k_2 = k_2 + 1$ ; if  $\hat{q}(\hat{\epsilon}_{[k_1]}) > \hat{q}(\hat{\epsilon}_{[k_2]})$  and  $\hat{q}(\hat{\epsilon}_{[k_1-1]}) > \hat{q}(\hat{\epsilon}_{[k_2-1]})$ ,  $k_1 = k_1 - 1$  and  $k_2 = k_2 - 1$ .
- Step 2: Iterate Step 1 until none of above two conditions is satisfied or  $(k_1 - 1)(k_2 - n) = 0$ .

Based on Yao and Li (2014)’s numerical studies, the above proposed prediction intervals have superior performance to existing symmetric prediction intervals when the data is skewed.

### 3 Nonparametric Modal Regression

Similar to the traditional linear regression, linear modal regression requires a strong parametric assumption which might not hold in practice. To relax the parametric assumption, there are also nonparametric modal regression that is built based on kernel density estimation. Readers are referred to Chen (2018) for a detailed review of nonparametric modal regressions. For simplicity of explanation, in this section, the covariate  $X$  is assumed to be univariate with a compactly supported density function. The estimation procedure can be easily extended to multivariate case but practically difficult due to the ‘‘curse of dimensionality.’’

Let  $f(z)$  denote the probability density function (pdf) of a random variable  $Z$  and be twice differentiable. Then, define the global mode and local modes of  $f(z)$ , respectively, as:

$$\text{UniMode}(Z) = \arg \max_z f(z)$$

and

$$\text{MultiMode}(Z) = \{z : f'(z) = 0, f''(z) < 0\}.$$

UniMode( $Z$ ), which focuses on the conditional global mode, is called the uni-modal regression, as studied by Lee (1989); Manski (1991). MultiMode( $Z$ ), on the other hand, studies the conditional local modes, and is sufficiently investigated by Chen et al. (2016).

The uni-modal regression searches for the function

$$m(x) = \text{UniMode}(Y|X = x) = \arg \max_y f(y|x)$$

and multi-modal regression targets at

$$M(x) = \text{MultiMode}(Y|X = x) = \left\{ y : \frac{\partial}{\partial y} f(y|x) = 0, \frac{\partial^2}{\partial y^2} f(y|x) < 0 \right\},$$

where  $f(y|x) = f(x, y)/f(x)$  is the conditional density of  $Y$  given  $X = x$ . Note that, for a given  $x$ , the modes or local modes of  $f(y|x)$  and  $f(x, y)$  are the same. Therefore, the uni-modal and multi-modal regression can be also defined as

$$m(x) = \text{UniMode}(Y|X = x) = \arg \max_y f(x, y), \tag{7}$$

and

$$M(x) = \left\{ y : \frac{\partial}{\partial y} f(x, y) = 0, \frac{\partial^2}{\partial y^2} f(x, y) < 0 \right\}, \tag{8}$$

respectively.

### 3.1 Estimating Uni-Modal Regression

First, we estimate the joint density  $f(x, y)$  by the kernel density estimator (KDE) as

$$\hat{f}_n(x, y) = \frac{1}{nh_1h_2} \sum_{i=1}^n K_1\left(\frac{x_i - x}{h_1}\right) K_2\left(\frac{y_i - y}{h_2}\right), \tag{9}$$

where  $K_1$  and  $K_2$  are kernel densities such as Gaussian functions and  $h_1 > 0$  and  $h_2 > 0$  are tuning parameters. Then, a nonparametric modal regression estimator of  $m(x)$  in (7) is

$$\hat{m}_n(x) = \arg \max_y \hat{f}_n(x, y).$$

If  $K_2$  is assumed to be a spherical kernel such as  $K_2(z) = \frac{1}{2}I(|z| \leq 1)$ , then it has been shown that the maximization operation is equivalent to the minimization operator on a flattened 0 – 1 loss.

Yao and Xiang (2016) proposed a local polynomial modal regression (LPMR) estimation procedure to estimate the nonparametric modal regression, which maximizes

$$\ell(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_i - x_0) \phi_{h_2} \left( y_i - \sum_{j=1}^p \beta_j (x_i - x_0)^j \right) \tag{10}$$

over  $\boldsymbol{\theta} = (\beta_0, \dots, \beta_p)$ . Similar to Yao et al. (2012), the authors used an EM algorithm to maximize (10) since it has a mixture type form. The asymptotic properties were discussed and proved.

Feng et al. (2020) also studied nonparametric modal regression from a statistical learning viewpoint through the classical empirical risk minimization (ERM) scheme and investigated its theoretical properties.

### 3.2 Estimating Multi-Modal Regression

Similar to the estimation of uni-modal regression, Chen et al. (2016) proposed estimating the multi-modal regression by a plug-in estimate from the KDE, as follows:

$$\hat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y} \hat{f}_n(x, y) = 0, \frac{\partial^2}{\partial y^2} \hat{f}_n(x, y) < 0 \right\},$$

where  $\hat{f}_n(x, y)$  is from (9).

By assuming  $K_1$  and  $K_2$  to be Gaussian kernels,  $\hat{M}_n(x)$  can be estimated through a mean-shift algorithm (Chen et al. 2016) which is actually equivalent to the mode hunting EM algorithm (Yao 2013, MEM). The results can be applied to other radially symmetric kernels as well. The partial mean-shift algorithm is summarized in Algorithm 3.1.

**Algorithm 3.1** *Partial mean-shift*

**Input:** Samples  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , bandwidths  $h_1$  and  $h_2$ .

1. Find a starting set  $\mathcal{M} \in \mathbb{R}^2$ , such as  $\mathcal{D}$ .
2. For each  $(x, y) \in \mathcal{M}$ , fix  $x$  and update  $y$  by
 
$$y \leftarrow \frac{\sum_{i=1}^n y_i K(|x-x_i|/h_1) K(|y-y_i|/h_2)}{\sum_{i=1}^n K(|x-x_i|/h_1) K(|y-y_i|/h_2)}$$
 until convergence. Let  $y^\infty$  be the converged value.

**Output:**  $\mathcal{M}^\infty$ , which contains the points  $(x, y^\infty)$ .

Comparing between uni-modal and multi-modal regression, we can see that multi-modal regression is more preferred in situations where there are hidden heterogeneous relations in the data set. In addition, if there are several modes in the original data, since the uni-modal regression can only detect the main component, the prediction regions tend to be wider than that of the multi-modal regression, as shown in Fig. 2. However, it is obvious that the uni-modal regression is easier to interpret, which is quite important in data applications.

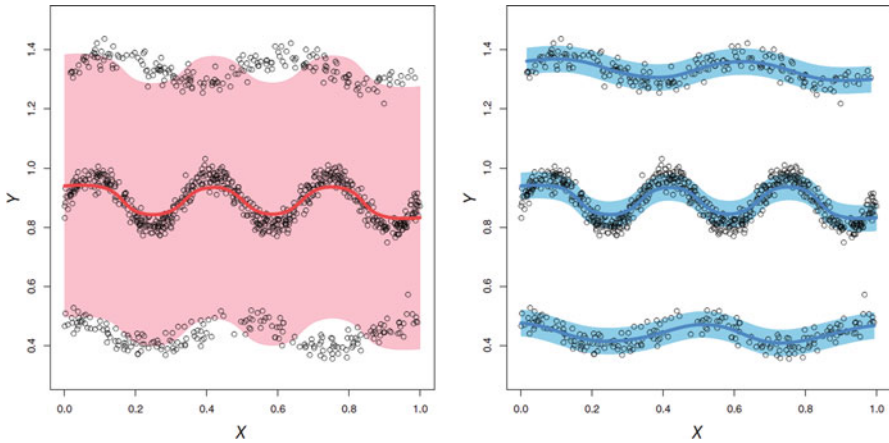


Fig. 2 Uni-modal vs multi-modal regression

### 4 Semiparametric Modal Regression

Many authors have extended the linear modal regression (Kemp & Santos Silva 2012; Yao & Li 2014) to semiparametric models. See, for example, Krief (2017), Ota et al. (2019), and Yao and Xiang (2016). In this section, we explain the idea of semiparametric modal regression using the varying coefficient modal regressions proposed by Yao and Xiang (2016).

To be more specific, given a random sample  $\{(\mathbf{x}_i, u_i, y_i); 1 \leq i \leq n\}$ , where  $y_i$  is the response variable and  $(\mathbf{x}_i, u_i)$  are covariates, Yao and Xiang (2016) proposed a nonparametric varying coefficient modal regression, defined as

$$\text{Mode}(y|\mathbf{x}_i, u_i) = \max_y f(y|\mathbf{x}_i, u_i) = \sum_{j=1}^p g_j(u_i)x_{ij}, \tag{11}$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$  and  $\{g_1(u), \dots, g_p(u)\}^\top$  are unknown smooth functions. If  $g_j(u)$  is constant for all  $j$ , then the above model becomes the linear modal regression (2). In addition, the nonparametric uni-modal regression introduced in Sect. 3 is a special case of (11) when  $p = 1$  and  $x_i = 1$ . Allowing  $g_j(u)$  to depend on some index  $u$ , the varying coefficient modal regression can relax the constant coefficient assumption of the linear modal regression, and also better model how the modal regression coefficients dynamically change over the index  $u$ , which could be a time or location index. Compared to the fully nonparametric modal regression, the above model can easily adopt multivariate covariates by imposing some model assumption on the conditional mode. Therefore, the semiparametric modal regression can combine the benefits of both the parametric modal regression and the nonparametric modal regression.

Yao and Xiang (2016) proposed estimating the varying coefficient modal regression (11) by a local linear approximation of  $g_j(u)$  in a neighborhood of  $u_0$ ,

$$g_j(u) \approx g_j(u_0) + g'_j(u_0)(u - u_0) = b_j + c_j(u - u_0).$$

Let  $\theta = (b_1, \dots, b_p, h_1c_1, \dots, h_1c_p)^\top$ . Then  $\theta$  is found by maximizing

$$\ell(\theta) = \sum_{i=1}^n K_{h_1}(u_i - u_0)\phi_{h_2} \left[ y_i - \sum_{j=1}^p \{b_j + c_j(u_i - u_0)\}x_{ij} \right], \tag{12}$$

over  $\theta$ . Let  $\hat{\theta} = (\hat{b}_1, \dots, \hat{b}_p, h_1\hat{c}_1, \dots, h_1\hat{c}_p)^\top$  be the maximizer of (12). Then  $\hat{\mathbf{g}}(u_0) = (\hat{b}_1, \dots, \hat{b}_p)^\top$  is the estimate of  $\{g_1(u_0), \dots, g_p(u_0)\}^\top$ , and  $\hat{\mathbf{g}}'(u_0) = (\hat{c}_1, \dots, \hat{c}_p)^\top$  is the estimate of  $\{g'_1(u_0), \dots, g'_p(u_0)\}^\top$ .

The algorithm proposed to maximize (12) is summarized as follows.

**Algorithm 4.1** Starting with  $t = 0$ :

E-Step: Update  $\pi(j | \theta^{(t)})$

$$\pi(j | \theta^{(t)}) = \frac{K_{h_1}(u_j - u_0)\phi_{h_2} \left[ y_j - \sum_{l=1}^p \{b_l^{(t)} + c_l^{(t)}(u_j - u_0)\}x_{jl} \right]}{\sum_{i=1}^n K_{h_1}(u_i - u_0)\phi_{h_2} \left[ y_i - \sum_{l=1}^p \{b_l^{(t)} + c_l^{(t)}(u_i - u_0)\}x_{il} \right]},$$

$$j = 1, \dots, n.$$

M-Step: Update  $\theta^{(t+1)}$

$$\theta^{(t+1)} = \arg \max_{\theta} \sum_{j=1}^n \pi(j | \theta^{(t)}) \log \phi_{h_2} \left[ y_j - \sum_{l=1}^p \{b_l^{(t)} + c_l^{(t)}(u_j - u_0)\}x_{jl} \right],$$

which has an explicit solution since  $\phi(\cdot)$  is the Gaussian density.

Denote by  $f_u(u)$  the marginal density of  $u$ ,  $q(\epsilon | \mathbf{x}, u)$  the conditional density of  $\epsilon = y - \sum_{j=1}^p g_j(u)x_j$  given  $\mathbf{x}$  and  $u$ , and  $q^{(v)}(\epsilon | \mathbf{x}, u)$  the  $v$ -th derivative of  $q(\epsilon | \mathbf{x}, u)$ . Let

$$\alpha_j(u) = E\{\mathbf{x}X_j q^{(2)}(0 | \mathbf{x}, u) | u\}, \quad \beta(u) = E\{\mathbf{x}q^{(3)}(0 | \mathbf{x}, u) | u\}$$

$$\Delta(u) = E\{\mathbf{x}\mathbf{x}^\top q^{(2)}(0 | \mathbf{x}, u) | u\}, \quad \tilde{\Delta}(u) = E\{\mathbf{x}\mathbf{x}^\top q(0 | \mathbf{x}, u) | u\}.$$

Yao and Xiang (2016) provided the following asymptotic properties for the proposed varying coefficient modal regression estimator.

**Theorem 2** Under the regularity conditions (A4)|(A6) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1^3h_2^5 \rightarrow \infty$  and  $h_1^2/h_2 \rightarrow 0$  the asymptotic bias of  $\hat{\mathbf{g}}(u_0)$  is given by

$$\text{Bias}\{\hat{\mathbf{g}}(u_0)\} = \frac{1}{2}\Delta^{-1}(u_0)\left\{\mu_2h_1^2\sum_{j=1}^p g_j''(u_0)\alpha_j(u_0) - h_2^2\beta(u_0)\right\}\{1 + o_p(1)\}, \tag{13}$$

and the asymptotic covariance is

$$\text{Cov}\{\hat{\mathbf{g}}(u_0)\} = \frac{\tilde{v}v_0}{nh_1h_2^3f_u(u_0)}\Delta^{-1}(u_0)\tilde{\Delta}(u_0)\Delta^{-1}(u_0)\{1 + o_p(1)\}, \tag{14}$$

where  $\mu_j = \int t^j K(t)dt$ ,  $v_j = \int t^j K^2(t)dt$ , and  $\tilde{v} = \int t^2\phi^2(t)dt$ .

**Theorem 3** Under the same condition as in Theorem 2, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1h_2^5 \rightarrow \infty$  and  $h_1^2/h_2 \rightarrow 0$ , the estimate  $\mathbf{g}(u_0)$  has the following asymptotic distribution

$$[\text{Cov}\{\hat{\mathbf{g}}(u_0)\}]^{-1/2}[\hat{\mathbf{g}}(u_0) - \mathbf{g}_0(u_0) - \text{Bias}\{\hat{\mathbf{g}}(u_0)\}] \xrightarrow{L} N(0, I),$$

where  $\text{Bias}\{\hat{\mathbf{g}}(u_0)\}$  is defined in (13) and  $\text{Cov}\{\hat{\mathbf{g}}(u_0)\}$  is defined in (14).

## 5 Discussion

In this article, we introduced modal regressions, which can be a good complement to mean/quantile regression, and are especially suitable for skewed, truncated, or contaminated data with outliers. Compared to traditional mean regression models, the modal regression models are more robust and have better prediction performance. Simulation studies and real data analysis are done to illustrate the numerical performance of the new methods. Due to the length of the article, the readers are referred to Yao and Li (2014) and Yao and Xiang (2016) for the details.

The development of modal regression is still in its early stage. Parallel to the traditional mean/quantile regression, the modal regression can be extended to a broad variety of parametric, nonparametric, and semiparametric modal regression models. For high dimensional models, it is interesting to investigate how to perform feature screening and variable selection for modal regression. In addition, it also requires more research to extend the modal regression to the longitudinal/panel data (Ullah et al. 2021), time series data, data with measurement errors, and missing data problems.

## Appendix

The conditions used by the theorems are listed below. They are not the weakest possible conditions, but they are imposed to facilitate the proofs.

### Technical Conditions:

- (A1)  $q^{(v)}(t | x)$ ,  $v = 0, 1, 2, 3$  is continuous in a neighborhood of 0, and  $q'(0 | x) = 0$  for any  $x$ .
- (A2)  $n^{-1} \sum_{i=1}^n q''(0 | x_i) x_i x_i^T = J + o_p(1)$ ,  $n^{-1} \sum_{i=1}^n q'''(0 | x_i) x_i = K + o_p(1)$  and  $n^{-1} \sum_{i=1}^n q(0 | x_i) x_i x_i^T = L + o_p(1)$ , where  $J < 0$ , that is,  $-J$  is a positive definite matrix.
- (A3)  $n^{-1} \sum_{i=1}^n \|x_i\|^4 = O_p(1)$ , and  $q'(0 | x) = 0$  any  $x$ .
- (A4)  $g_j(x)$  has continuous  $2^{nd}$  derivative at the point  $x_0$ ,  $j = 1, \dots, p$ .
- (A5)  $q'(0 | \mathbf{x}, u) = 0$ ,  $q''(0 | \mathbf{x}, u) < 0$ ,  $q^{(v)}(t | \mathbf{x}, u)$  is bounded in a neighbor of  $(\mathbf{x}_0, u_0)$  and has continuous first derivative at the point  $(\mathbf{x}_0, u_0)$  as a function of  $(\mathbf{x}, u)$ , for  $v = 0, \dots, 4$ .
- (A6) The  $f_u(u)$  is bounded and has continuous first derivative at the point  $u_0$  and  $f(u_0) > 0$ .

## References

- Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41, 997–1016.
- Cardoso, A. R., & Portugal, P. (2005). Contractual wages and the wage cushion under different bargaining settings. *Journal of Labor Economics*, 23, 875–902.
- Chaouch, P., Laïb, N., & Louani, D. (2017). Rate of uniform consistency for a class of mode regression on functional stationary ergodic data. *Statistical Methods & Applications*, 26(1), 19–47.
- Chauduri, P., & Marron, J. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94, 807–823.
- Chen, Y. (2018). Modal regression using kernel density estimation: a review. *Advanced Review*, 10, 1–14.
- Chen, Y. C., Genovese, C. R., Tibshirani, R. J., & Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, 44, 489–514.
- Eddy, W. P. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics*, 8, 870–882.
- Einbeck, J., & Tutz, G. (2006). Modelling beyond regression functions: an application of multimodal regression to speed-flow data. *Applied Statistics*, 55, 461–475.
- Feng, Y., Fan, J., & Suykens, J. A. (2020). A statistical learning approach to modal regression. *Journal of Machine Learning Research*, 21(2), 1–35.
- 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. *The Annals of Statistics*, 32, 2124–2141.
- Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, 40, 105–128.
- Henderson, D. J., & Parmeter, C. F. (2015). *Applied nonparametric econometrics*. Cambridge University Press.



- Henderson, D. J., & Russell, R. R. (2005). Human capital and convergence: a production frontier approach. *International Economic Review*, *46*, 1167–1205.
- Henderson, D. J., Parmeter, C. F., & Russell, R. R. (2008). Modes, weighted modes, and calibrated modes: evidence of clustering using modality tests. *Journal of Applied Econometrics*, *23*, 607–638.
- Kemp, G. C. R., & Santos Silva, J. M. C. (2012). Regression towards the mode. *Journal of Economics*, *170*, 92–101.
- Kemp, G. C. R., Parente, P., & Santos Silva, J. M. C. (2019). Dynamic vector mode regression. *Journal of Business & Economic Statistics*, *38*, 647–661.
- Krief, J. M. (2017). Semi-linear mode regression. *The Econometrics Journal*, *20*(2), 149–167.
- Lee, M. J. (1989). Mode regression. *Journal of Econometrics*, *42*, 337–349.
- Lewbel, A., & Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica*, *70*, 765–779.
- Li, X., & Huang, X. (2019). Linear mode regression with covariate measurement error. *Canadian Journal of Statistics*, *47*(2), 262–280.
- Manski, C. (1991). Regression. *Journal of Economic Literature*, *29*, 34–50.
- Mirowsky, J. (2013). Analyzing associations between mental health and social circumstances. In *Handbook of the sociology of mental health* (pp. 143–165).
- Ota, H., Kato, K., Hara, S., et al. (2019). Quantile regression approach to conditional mode estimation. *Electronic Journal of Statistics*, *13*(2), 3120–3160.
- Park, B. U., Simar, L., & Zelenyuk, V. (2008). Local likelihood estimation of truncated regression and its partial derivatives: Theory and application. *Journal of Econometrics*, *146*, 185–198.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Journal of American Statistical Association*, *33*, 1065–1076.
- Ray, S., & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *The Annals of Statistics*, 2042–2065.
- Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization*. New York: Wiley.
- Ullah, A., Wang, T., & Yao, W. (2021). Modal regression for fixed effects panel data. *Empirical Economics*, *60*(1), 261–308.
- Wang, X., Chen, H., Shen, D., & Huang, H. (2017). Cognitive impairment prediction in Alzheimer's disease with regularized modal regression. *Advances in Neural Information Processing Systems*, 1447–1457.
- Weber, M. (1993). *The sociology of religion*.
- Yao, W. (2013). A note on EM algorithm for mixture models. *Statistics Probability Letters*, *83*, 519–526.
- Yao, W., & Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, *41*, 656–671.
- Yao, W., & Lindsay, B. G. (2009). Bayesian mixture labelling by highest posterior density. *Journal of American Statistical Association*, *104*, 758–767.
- Yao, W., & Xiang, S. (2016). Nonparametric and varying coefficient modal regression. arXiv:1602.06609.
- Yao, W., Lindsay, B. G., & Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics*, *24*, 647–663.
- Zhou, H., & Huang, X. (2016). Nonparametric modal regression in the presence of measurement error. *Electronic Journal of Statistics*, *10*(2), 3579–3620.
- Zhou, H., & Huang, X. (2019). Bandwidth selection for nonparametric modal regression. *Communications in Statistics-Simulation and Computation*, *48*(4), 968–984.