IEICE TRANS. FUNDAMENTALS, VOL.E106–A, NO.2 FEBRUARY 2023

PAPER

Modal Interval Regression Based on Spline Quantile Regression

Sai YAO[†a)], ***Nonmember***, **Daichi KITAHARA**[†*b)], **Hiroki KURODA**[†**c)], ***and*** **Akira HIRABAYASHI**[†], ***Members***

SUMMARY The *mean*, *median*, and *mode* are usually calculated from univariate observations as the most basic representative values of a random variable. To measure the spread of the distribution, the *standard deviation*, *interquartile range*, and *modal interval* are also calculated. When we analyze continuous relations between a pair of random variables from bivariate observations, *regression analysis* is often used. By minimizing appropriate costs evaluating regression errors, we estimate the conditional mean, median, and mode. The conditional standard deviation can be estimated if the bivariate observations are obtained from a Gaussian process. Moreover, the conditional interquartile range can be calculated for various distributions by the *quantile regression* that estimates any conditional quantile (percentile). Meanwhile, the study of the modal interval regression is relatively new, and *spline regression models*, known as flexible models having the optimality on the smoothness for bivariate data, are not yet used. In this paper, we propose a modal interval regression method based on spline quantile regression. The proposed method consists of two steps. In the first step, we divide the bivariate observations into bins for one random variable, then detect the modal interval for the other random variable as the lower and upper quantiles in each bin. In the second step, we estimate the conditional modal interval by constructing both lower and upper quantile curves as spline functions. By using the spline quantile regression, the proposed method is widely applicable to various distributions and formulated as a *convex optimization problem* on the coefficient vectors of the lower and upper spline functions. Extensive experiments, including settings of the bin width, the smoothing parameter and weights in the cost function, show the effectiveness of the proposed modal interval regression in terms of accuracy and visual shape for synthetic data generated from various distributions. Experiments for real-world meteorological data also demonstrate a good performance of the proposed method.

key words: *regression analysis, modal interval, spline function, quantile regression, convex optimization*

1. Introduction

In the big data era, data analysis [1] is becoming more and more important. When we obtain a large number of univariate observations, we compute the *mean*, *median*, and *mode* as the most basic representative values of the observations. We also compute the *standard deviation*, *interquartile range*, and *modal interval* to analyze the spread of the observations. For non-Gaussian asymmetric data, the mean with the standard deviation is not appropriate as a representative value, and the median with the interquartile range is used instead. The mode and the modal interval are used to directly represent the region where observations appear most frequently.

When we obtain multivariate observations, it is important to analyze the relation between multiple components in addition to the univariate analysis for each component. This paper focuses on the simplest case where we analyze the relation between a pair of random variables from bivariate observations. In this case, regression analysis [1]–[7] is widely used because it can give a continuous relation by constructing a univariate continuous function that maps one random variable to the other one. This continuous function is often constructed as a low-order univariate polynomial having the *least square errors*, which represents the *conditional mean*.

In robust statistics [8]–[12], we construct a polynomial having the *least absolute errors*, which represents the *conditional median* and suppresses the influence of outliers. The quantile regression [13]–[16] is regarded as a generalization of the median regression and constructs a polynomial minimizing certain *asymmetric absolute errors*, which represents the *conditional quantile (percentile)*. The modal regression [17]–[22] can estimate the *conditional mode*, i.e., the curve where bivariate observations appear most frequently. However, the modal regression is relatively difficult to use since it has to minimize certain *kernel-based nonconvex errors*.

Though polynomial regression models are easy to handle, such simple models cannot flexibly express the true conditional mean, median, quantile, and mode of various shapes. To make the most of each regression analysis, spline regression models are used [7], [9], [15], [22]. Spline functions are smooth piecewise polynomials and widely used to estimate smooth functions because of their flexibility and optimality for bivariate data [23]–[28]. The spline functions are locally low-order polynomials, and hence they are fairly tractable.

Compared with regression of a certain single curve, regression of a certain interval is more informative since it can also express the change of the spread of observations. If bivariate observations are given from a Gaussian process, the conditional mean and the *conditional standard deviation* can be estimated by the Gaussian process regression [29]–[32]. For non-Gaussian asymmetric observations, the *conditional interquartile range* is more important, and it can be estimated by the quantile regression for the first and third quartiles. The *conditional modal interval* (see Definition 2) is very useful to visualize the trend of bivariate observations since it shows the

Manuscript received March 15, 2022.
Manuscript revised June 22, 2022.
Manuscript publicized July 26, 2022.

[†]The authors are with the Dept. of Information Science and Engineering, Ritsumeikan University, Kusatsu-shi, 525-8577 Japan.
[*]Presently, with the Division of Electrical, Electronic and Infocommunications Engineering, Osaka University, Suita-shi, 565-0871 Japan.
[**]Presently, with the Dept. of Information and Management Systems Engineering, Nagaoka University of Technology, Nagaoka-shi, 940-2188 Japan.
a) E-mail: yaosai96@outlook.com
b) E-mail: d-kita@media.ritsumei.ac.jp (Corresponding author)
c) E-mail: kuroda@vos.nagaokaut.ac.jp
DOI: 10.1587/transfun.2022EAP1031

region where the observations appear most frequently for a user-specified coverage probability. However, the number of existing works on the modal interval regression is relatively small [33]–[41] compared to the other regressions [1]–[22], [29]–[32], and spline regression models are not yet used.

To flexibly express the smooth modal intervals of various shapes, this paper proposes a modal interval regression method using spline regression models similar to [42]. The proposed modal interval regression consists of two steps. In the first step, we divide bivariate observations into some bins for one random variable, then we detect the *local empirical modal interval* for the other random variable as the lower and upper quantiles in each bin. In the second step, we estimate the lower and upper curves of the conditional modal interval by the simultaneous spline quantile regression that guarantees the *non-crossing property* for the two curves. Thanks to the simultaneous quantile regression, we can obtain more reliable results compared to a straightforward least squares approach using the local empirical modal intervals. In addition, the second step is formulated as a *convex optimization problem* on the coefficient vectors of the lower and upper spline functions and can be solved by several algorithms [43]–[49].

Numerical experiments, including settings of the bins, the smoothing parameter and weights in the cost function, show the effectiveness of the proposed method, compared to the straightforward least squares method and kernel density estimation methods, in terms of accuracy and visual shape for various synthetic data. Experiments for real-world meteorological data also show that the proposed method well visualizes the region where observations appear most frequently.

2. Basics of Regression Analysis

Let $\mathbb{R}$ and $\mathbb{N}$ be the sets of all real numbers and nonnegative integers, respectively. For any nonnegative integer or infinity $\rho \in \mathbb{N} \cup \{\infty\}$ and any open interval[†] $(a, b) \subseteq \mathbb{R}$ (s.t. $a < b$), $C^\rho(a, b)$ denotes the set of all ρ-times continuously differentiable functions on (a, b), and $\mathrm{cl}(a, b)$ denotes the closure of (a, b). For any $d \in \mathbb{N}$, $\mathbb{P}_d$ ($\subsetneq C^\infty(-\infty, \infty)$) stands for the set of all univariate real polynomials of degree d at most, i.e., $\mathbb{P}_d := \{u : \mathbb{R} \ni x \mapsto \sum_{k=0}^{d} c_k x^k \in \mathbb{R} \mid c_k \in \mathbb{R}\}$. We write vectors and matrices by boldface small and capital letters, respectively, and random variables by normal capital letters.

2.1 Least Squares: Mean Regression

Suppose that we have finite observations $(x_n, y_n) \in \mathbb{R}^2$ ($n = 1, 2, \ldots, N$) of a pair of real-valued random variables (X, Y) whose joint probability density function $f_{X,Y}(x, y)$ is continuous and satisfies $\iint_\Omega f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y = 1$ and $f_{X,Y}(x, y) > 0$ for all (x, y) in a simply connected domain $\Omega \subseteq \mathbb{R}^2$. We use a rectangular domain $\Omega := (x_{\inf}, x_{\sup}) \times (y_{\inf}, y_{\sup})$ in this paper for simplicity. The *conditional probability density function of Y given X* is defined by $f_{Y|X}(y|x) := f_{X,Y}(x, y)/f_X(x) := f_{X,Y}(x, y)/\int_{y_{\inf}}^{y_{\sup}} f_{X,Y}(x, y)\,\mathrm{d}y > 0$ for all $(x, y) \in \Omega$.

[†]Note that values of a and b can be $-\infty$ and ∞, respectively.

When we analyze continuous relations between the two random variables X and Y from the bivariate observations $\{(x_n, y_n)\}_{n=1}^N$, regression analysis [1]–[22] has been widely used. In particular, the *least squares regression*

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{n=1}^{N} |y_n - r_{\boldsymbol{\theta}}(x_n)|^2 \tag{1}$$

is often used due to its low computational cost [1]–[7], where the function $r_{\boldsymbol{\theta}}(x)$ is some regression model such as a low-order polynomial and the vector $\boldsymbol{\theta}$ denotes adjustable parameters to be optimized. For example, $\boldsymbol{\theta}$ is the coefficient vector if $r_{\boldsymbol{\theta}}(x)$ is a polynomial. We find the minimizer $\boldsymbol{\theta}^*$ of (1), and the function $r_{\boldsymbol{\theta}^*}(x)$ expresses one average relation between X and Y. This is because, when N approaches infinity, $r_{\boldsymbol{\theta}^*}(x)$ converges[††] to the *conditional mean $\mu_Y(x)$ of Y given $X = x$*:

$$r_{\boldsymbol{\theta}^*}(x) \xrightarrow{N\to\infty} \mu_Y(x) := E[Y|X = x] = \int_{y_{\inf}}^{y_{\sup}} y f_{Y|X}(y|x)\,\mathrm{d}y$$

under the assumption that $r_{\boldsymbol{\theta}}(x)$ can exactly express $\mu_Y(x)$ if we choose the appropriate vector $\boldsymbol{\theta}$, i.e., $r_{\boldsymbol{\theta}}(x) = \mu_Y(x)$ holds for some $\boldsymbol{\theta}$ (see, e.g., [1] for proof). Hence, the least squares regression can be regarded as the *mean regression*.

2.2 Least Absolute Deviations: Median Regression

It is known that the square errors used in (1) are sensitive to outliers, and hence the reliability of the least squares regression greatly decreases for observations from an asymmetric long-tailed distribution. For data of such a non-Gaussian distribution, the *least absolute deviations regression* [8]–[12]

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{n=1}^{N} |y_n - r_{\boldsymbol{\theta}}(x_n)| \tag{2}$$

is used instead. The absolute errors do not over-evaluate outliers differently from the square errors, and the minimizer $\boldsymbol{\theta}^*$ of (2) is robust for long-tailed data even if N is small. Moreover, when N approaches infinity, the solution $r_{\boldsymbol{\theta}^*}(x)$ to (2) converges to the *conditional median $\varrho_Y(x)$* of Y given $X = x$:

$$r_{\boldsymbol{\theta}^*}(x) \xrightarrow{N\to\infty} \varrho_Y(x) \text{ satisfying } \int_{y_{\inf}}^{\varrho_Y(x)} f_{Y|X}(y|x)\,\mathrm{d}y = 0.5$$

under the assumption that $r_{\boldsymbol{\theta}}(x)$ can exactly express $\varrho_Y(x)$ if we choose the appropriate $\boldsymbol{\theta}$ (see [8]). The least absolute deviations regression can be regarded as the *median regression*.

2.3 Quantile Regression

By generalizing the fact that the least absolute deviations regression is the same as the median regression, we can estimate any *quantile (percentile)* as follows. Define the *conditional cumulative distribution function of Y given $X = x$* by

[††]For readability, we have simply stated that $r_{\boldsymbol{\theta}^*}(x)$ converges to $\mu_Y(x)$. In fact, the minimizer $\boldsymbol{\theta}^*$ is also a random variable, and convergence of a random variable sequence requires careful discussion.

$$F_{Y|x}(y) := \int_{y_{\inf}}^{y} f_{Y|X}(t|x)\,\mathrm{d}t \quad \text{for } y \in (y_{\inf}, y_{\sup}).$$

Since $F_{Y|x}(y)$ is a strictly increasing function because of the positivity of $f_{Y|X}(y|x)$, its inverse function $F^{-1}_{Y|x}(p)$ is well-defined for $p \in (0,1)$. This inverse function $F^{-1}_{Y|x}(p)$ equals the *conditional quantile function of Y given $X = x$* [3], [50]:

$$Q_{Y|x}(p) := F^{-1}_{Y|x}(p) \quad \text{for } p \in (0,1).$$

The value of $q_{p,Y}(x) := Q_{Y|x}(p)$ for $x \in (x_{\inf}, x_{\sup})$ is called the *pth conditional quantile of Y given $X = x$*. Note that the pth quantile $q_{p,Y}(x)$ is also called the *$100p$th percentile (or centile)*, e.g., in [51]–[53]. If $p = 0.5$, the conditional quantile $q_{p,Y}(x)$ is equivalent to the conditional median $\varrho_Y(x)$.

To generalize the median regression to the *quantile regression*, we extend the (symmetric) absolute value function to an *asymmetric absolute value function* defined by

$$\mathcal{J}_p(t) := \begin{cases} pt & \text{if } t \geq 0, \\ -(1-p)t & \text{if } t < 0, \end{cases}$$

with $p \in (0,1)$. The function $\mathcal{J}_p(t)$ is symmetric only when $p = 0.5$, and $\mathcal{J}_{0.5}(t) = \frac{1}{2}|t|$ holds. The quantile regression is formulated as the following optimization problem [13]–[16]

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{n=1}^{N} \mathcal{J}_p(y_n - r_{\boldsymbol{\theta}}(x_n)). \tag{3}$$

By finding the minimizer $\boldsymbol{\theta}^*$ of (3) for some fixed $p \in (0,1)$, we obtain an estimate of $q_{p,Y}(x)$ as $r_{\boldsymbol{\theta}^*}(x)$. Indeed, when N approaches infinity, the solution $r_{\boldsymbol{\theta}^*}(x)$ to (3) converges to the pth conditional quantile $q_{p,Y}(x)$ of Y given $X = x$:

$$r_{\boldsymbol{\theta}^*}(x) \xrightarrow{N\to\infty} q_{p,Y}(x) \text{ satisfying } F_{Y|x}\big(q_{p,Y}(x)\big) = p$$

under the assumption that $r_{\boldsymbol{\theta}}(x)$ can exactly express $q_{p,Y}(x)$ if we choose the appropriate $\boldsymbol{\theta}$ (see [15]). This is because the optimal solution $r_{\boldsymbol{\theta}^*}(x)$ to (3) yields lower-side errors $\epsilon_n^- = y_n - r_{\boldsymbol{\theta}}(x_n) < 0$ and upper-side errors $\epsilon_n^+ = y_n - r_{\boldsymbol{\theta}}(x_n) > 0$ at a ratio of $p : 1-p$. Therefore, we can easily estimate any quantile curve $q_{p,Y}(x)$ by just changing the value of p in (3).

2.4 Modal Regression

Both mean regression in Sect. 2.1 and median regression in Sect. 2.2 implicitly suppose that the bivariate observations $\{(x_n, y_n)\}_{n=1}^N$ frequently appear around the conditional mean and median, respectively. If this implicit assumption holds, the solution $r_{\boldsymbol{\theta}^*}(x)$ to (1) or (2) can adequately represent the trend of the bivariate observations. However, otherwise, both regression results are not so appropriate to describe continuous relations between a pair of random variables (X, Y).

To directly estimate the curve where the bivariate observations appear most frequently, the *modal regression* [17]–[22] is used. The goal of the modal regression is to estimate

$$m_Y(x) := \underset{y \in \mathrm{cl}(y_{\inf}, y_{\sup})}{\operatorname{argmax}} f_{Y|X}(y|x), \tag{4}$$

and $m_Y(x)$ is called the *conditional mode of Y given $X = x$*. Some papers [54]–[57] define the mode by a local maximum, and thus a function is called *unimodal* if it has only one local maximum and *multimodal* if it has multiple local maxima. Meanwhile, as in [17]–[22], this paper defines the mode by the global maximum of $f_{Y|X}(y|x)$, while supposing that the maximizer exists uniquely in (4) for any $x \in (x_{\inf}, x_{\sup})$.

The modal regression can be formulated as

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \frac{1}{\sigma} \sum_{n=1}^{N} \left(\mathcal{K}(0) - \mathcal{K}\left(\frac{y_n - r_{\boldsymbol{\theta}}(x_n)}{\sigma} \right) \right), \tag{5}$$

where $\mathcal{K} : \mathbb{R} \to [0,1]$ is a *nonnegative symmetric unimodal kernel* s.t. $\int_{-\infty}^{\infty} \mathcal{K}(t)\,\mathrm{d}t = 1$ and $\mathcal{K}(t) = \mathcal{K}(-t) \leq \mathcal{K}(0)$ for all $t \in \mathbb{R}$, and $\sigma > 0$ is the *bandwidth parameter*. As such a kernel function, we often use the Gaussian kernel

$$\mathcal{K}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \tag{6}$$

Since the cost function in (5) evaluates regression errors for outliers as small values, the minimizer $\boldsymbol{\theta}^*$ of (5) is more robust for long-tailed data than that of (3). When N approaches infinity and $\sigma > 0$ approaches zero, the solution $r_{\boldsymbol{\theta}^*}(x)$ to (5) converges to the conditional mode $m_Y(x)$ of Y given $X = x$:

$$r_{\boldsymbol{\theta}^*}(x) \xrightarrow{N\to\infty,\, \sigma\to+0} m_Y(x) = \underset{y \in \mathrm{cl}(y_{\inf}, y_{\sup})}{\operatorname{argmax}} f_{Y|X}(y|x)$$

under the assumption that $r_{\boldsymbol{\theta}}(x)$ can exactly express $m_Y(x)$ if we choose the appropriate $\boldsymbol{\theta}$ (see [19]). This is because the function $\frac{1}{\sigma N} \sum_{n=1}^{N} \mathcal{K}(\frac{y_n - r_{\boldsymbol{\theta}}(x_n)}{\sigma})$ converges to the *modal regression risk* $\int_{x_{\inf}}^{x_{\sup}} f_{Y|X}(r_{\boldsymbol{\theta}}(x)|x) f_X(x)\,\mathrm{d}x$ whose maximizer is also the conditional mode $m_Y(x)$ of Y given $X = x$.

Although the minimizer $\boldsymbol{\theta}^*$ itself is robust for outliers, actually it is difficult to find the minimizer $\boldsymbol{\theta}^*$ due to the *nonconvexity* of the cost function in (5). In particular, when σ is close to zero, the gradient vanishes for most of observations, and hence we tend to get stuck in a poor local minimum [58]. As a result, since careful setting of σ and an appropriate initial value of $\boldsymbol{\theta}$ are needed, the modal regression is more difficult to use than the mean, median, and quantile regressions.

2.5 Spline Regression Model

In regression analysis, the most commonly used regression model is a polynomial $r_{\boldsymbol{\theta}}(x) = \sum_{k=0}^{d} c_k x^k$, and its degree is usually set to $d = 1$ or $d = 2$ [1], [8], [13], [17], where each coefficient $c_k \in \mathbb{R}$ of the polynomial is an adjustable parameter, i.e., $\boldsymbol{\theta} = (c_0, c_1, \ldots, c_d)^{\mathrm{T}} \in \mathbb{R}^{d+1}$. However, such a simple regression model often cannot express the true functions $\mu_Y(x)$, $\varrho_Y(x)$, $q_{p,Y}(x)$, and $m_Y(x)$ with sufficient accuracy.

To express curves of various shapes flexibly, *spline functions* are widely exploited as generalizations of polynomials [7], [9], [15], [22]. The spline function is a piecewise polynomial and possesses certain-times continuous differentiability, including at locations called *knots* where the polynomial pieces connect. In this paper, the knots are denoted by ξ_j s.t.

$-\infty =: \xi_{-1} < \xi_0 < \xi_1 < \cdots < \xi_b < \xi_{b+1} := \infty$, and subintervals divided by the knots are defined as $I_0 := (-\infty, \xi_0]$, $I_j := [\xi_{j-1}, \xi_j]$ $(j = 1, 2, \ldots, b)$, and $I_{b+1} := [\xi_b, \infty)$. For a partition $\sqcup := \{I_j\}_{j=0}^{b+1}$ and $\rho, d \in \mathbb{N}$ s.t. $0 \le \rho \le d$, define

$$\mathcal{S}_d^\rho(\sqcup) := \{s \in C^\rho(-\infty, \infty) \mid s = u_j \in \mathbb{P}_d \text{ on } I_j \in \sqcup\}$$

as the set of all univariate spline functions of degree d at most and smoothness ρ on $\sqcup$. The optimality of the spline function for estimation of a smooth curve is shown below [23]–[28].

Fact 1 (Spline smoothing): Suppose that bivariate observations $\{(x_n, y_n)\}_{n=1}^N$ satisfy $x_1 < x_2 < \cdots < x_N$. There is the optimal solution $g^* \in C^2(-\infty, \infty)$ to the following problem

$$\underset{g \in C^2(-\infty,\infty)}{\text{minimize}} \sum_{n=1}^N w_n \mathcal{D}(y_n - g(x_n)) + \lambda \int_{\mathbb{R}} |g''(x)|^2 \,\mathrm{d}x, \quad (7)$$

and it is a so-called *natural cubic spline function* $g^* \in \mathcal{S}_3^2(\sqcup)$ s.t. $\xi_j := x_{j+1}$ $(j = 0, 1, \ldots, N-1 =: b)$ [23], where $g''(x)$ is the second derivative of $g(x)$, $\mathcal{D} : \mathbb{R} \to \mathbb{R}$ is a data fidelity term such as the square error with a wight $w_n > 0$, and the smoothing parameter $\lambda > 0$ controls the trade-off between the data fidelity and the smoothness of the solution. Note that this solution is well-known when $\mathcal{D}$ is the square error, but for any $\mathcal{D}$ including nonconvex and asymmetric errors, the optimal solution is always a natural cubic spline function as long as the roughness penalty of the second term is added.

In general, the extrapolation results on I_0 and I_{b+1} are seldom used. In what follows, for $\sqcup_b := \{I_j\}_{j=1}^b$, we define

$$\mathcal{S}_d^\rho(\sqcup_b) := \{s \in C^\rho(\xi_0, \xi_b) \mid s = u_j \in \mathbb{P}_d \text{ on } I_j \in \sqcup_b\}$$

as the set of all univariate spline functions of degree d at most and smoothness ρ on $\sqcup_b$, where the domain of interest is restricted from $\mathbb{R} = (-\infty, \infty)$ to $I := (\xi_0, \xi_b)$ $(\supseteq (x_{\min}, x_{\max}) := (\min\{x_n\}, \max\{x_n\}))$. Although a spline regression model $r_{\boldsymbol{\theta}} = s \in \mathcal{S}_d^\rho(\sqcup_b)$ is really flexible by setting $h_j := \xi_j - \xi_{j-1}$ $(j = 1, 2, \ldots, b)$ to small values, overfitting will arise when the number N of observations is not enough. Therefore, by assuming that the energy of the curvature of the target $\mu_Y(x)$, $\varrho_Y(x)$, $q_{p,Y}(x)$, or $m_Y(x)$ is small and adding the roughness penalty in (7) as regularization for $r_{\boldsymbol{\theta}}(x) = s(x)$, we solve

$$\underset{s \in \mathcal{S}_d^\rho(\sqcup_b)}{\text{minimize}} \sum_{n=1}^N w_n \mathcal{D}(y_n - s(x_n)) + \lambda \int_I |s''(x)|^2 \,\mathrm{d}x, \quad (8)$$

where $2 \le \rho \le d$,[†] $w_n > 0$, and $\mathcal{D} : \mathbb{R} \to [0, \infty)$ is the data fidelity term in (1), (2), (3), or (5). In [15], each weight w_n at x_n is set to a value similar to $f_X(x_n)$. Note that when $\rho = d$, w_n is a constant for all n, and $\lambda > 0$ approaches zero in (8), the optimal solution $s^*(x)$ to (8) becomes the optimal polynomial of degree d in (1), (2), (3), or (5), depending on $\mathcal{D}$. In addition, for two sets of spline functions of different degrees and the same ρ on $\sqcup_b$, $d_1 < d_2 \Rightarrow \mathcal{S}_{d_1}^\rho(\sqcup_b) \subsetneq \mathcal{S}_{d_2}^\rho(\sqcup_b)$ holds and the optimal solution to (8) does not get worse when the degree d increases. Actually, if each h_j is sufficiently small, $d = \rho + 1$ is enough in many cases and setting $d \ge \rho + 2$ only gives almost the same result with a longer computation time.

[†]In most papers, the degree of a spline function is set to $d \le 5$.

3. Spline Modal Interval Regression

3.1 Modal Interval

In addition to the mean, median, and modal regressions, it is important to analyze the spread of a distribution. If bivariate observations $\{(x_n, y_n)\}_{n=1}^N$ are given from a certain Gaussian process,[††] the *conditional variance of Y given $X = x$*

$$v_Y(x) := V[Y|X = x] = \int_{y_{\inf}}^{y_{\sup}} |y - \mu_Y(x)|^2 f_{Y|X}(y|x) \,\mathrm{d}y$$

can be estimated [29]–[32]. For non-Gaussian distributions, especially asymmetric $f_{Y|X}(y|x)$, information of the conditional variance or the conditional standard deviation $\sqrt{v_Y(x)}$ is not so beneficial. Instead of the conditional standard deviation, the *conditional interquartile range of Y given $X = x$*

$$\mathrm{IQR}_Y(x) := \left[q_{0.25,Y}(x),\ q_{0.75,Y}(x)\right]$$

is often used to describe the spread of the distribution. For various distributions, $\mathrm{IQR}_Y(x)$ can be estimated by performing the quantile regression in (3) for $p = 0.25$ and $p = 0.75$.

As with the discussion about the modal regression in Sect. 2.4, the *conditional modal interval of Y given $X = x$* is more informative than $\mathrm{IQR}_Y(x)$ since it directly shows the region where bivariate observations appear most frequently. The modal interval is mainly defined in two ways as follows.

Definition 1 (Width-based modal interval): For every $\tau \in (0, \frac{y_{\sup} - y_{\inf}}{2})$, the width-based modal interval is defined by

$$\begin{aligned}\mathrm{MI}_Y^{[\tau]}(x) &:= \left[m_Y^{[\tau]}(x) - \tau,\ m_Y^{[\tau]}(x) + \tau\right] \\ &:= \underset{[y_{\mathrm{mid}}-\tau, y_{\mathrm{mid}}+\tau] \subseteq \mathrm{cl}(y_{\inf}, y_{\sup})}{\operatorname{argmax}} \int_{y_{\mathrm{mid}}-\tau}^{y_{\mathrm{mid}}+\tau} f_{Y|X}(y|x) \,\mathrm{d}y. \quad (9)\end{aligned}$$

Definition 2 (Probability-based modal interval): For every $\alpha \in (0, 1)$, the probability-based modal interval is defined by

$$\begin{aligned}&\mathrm{MI}_{\alpha,Y}(x) := \left[m_{\alpha,Y}^{\mathrm{low}}(x),\ m_{\alpha,Y}^{\mathrm{up}}(x)\right] \\ &:= \underset{[y_{\mathrm{low}}, y_{\mathrm{up}}] \subseteq \mathrm{cl}(y_{\inf}, y_{\sup})}{\operatorname{argmin}} y_{\mathrm{up}} - y_{\mathrm{low}} \text{ s.t. } \int_{y_{\mathrm{low}}}^{y_{\mathrm{up}}} f_{Y|X}(y|x) \,\mathrm{d}y = \alpha. \quad (10)\end{aligned}$$

We suppose that the maximizer in (9) and the minimizer in (10) exist uniquely for any $x \in (x_{\inf}, x_{\sup})$. Note that there are several names for the modal interval in Definition 2 such as the *shorth*, the *shortest prediction interval*, the *minimum volume region*, and the *highest density interval* [33]–[41]. When τ approaches zero in Definition 1, the middle point $m_Y^{[\tau]}(x)$ of $\mathrm{MI}_Y^{[\tau]}(x)$ converges to the conditional mode $m_Y(x)$ of Y

[††]In this situation, $x_{\inf} = y_{\inf} = -\infty$ and $x_{\sup} = y_{\sup} = \infty$ hold.

given $X = x$ in (4). Similarly when α approaches zero in Definition 2, both lower point $m^{\mathrm{low}}_{\alpha,Y}(x)$ and upper point $m^{\mathrm{up}}_{\alpha,Y}(x)$ of $\mathrm{MI}_{\alpha,Y}(x)$ converge to the conditional mode $m_Y(x)$.

A regression problem of the modal interval $\mathrm{MI}^{[\tau]}_Y(x)$ in Definition 1 is reduced to that of its middle point $m^{[\tau]}_Y(x)$. Moreover, if we set the bandwidth parameter σ according to the given interval width 2τ, $m^{[\tau]}_Y(x)$ can be estimated by solving the same optimization problem as in (5). As a result, this regression problem is usually discussed in a similar framework to the modal regression in Sect. 2.4 [59]–[61]. However, the modal interval $\mathrm{MI}^{[\tau]}_Y(x)$ in Definition 1 cannot express the change of the spread of the conditional probability density function $f_{Y|X}(y|x)$ as the value of x changes, since the interval width in (9) is fixed to 2τ for all $x \in (x_{\inf}, x_{\sup})$.

The modal interval $\mathrm{MI}_{\alpha,Y}(x)$ in Definition 2 is more informative than $\mathrm{MI}^{[\tau]}_Y(x)$ in Definition 1 since the change of the spread of $f_{Y|X}(y|x)$ is expressed as the change of the interval width $m^{\mathrm{up}}_{\alpha,Y}(x) - m^{\mathrm{low}}_{\alpha,Y}(x)$. Meanwhile, we have to estimate both lower point $m^{\mathrm{low}}_{\alpha,Y}(x)$ and upper point $m^{\mathrm{up}}_{\alpha,Y}(x)$, and thus a regression problem of the modal interval $\mathrm{MI}_{\alpha,Y}(x)$ in Definition 2 is more difficult than that of $\mathrm{MI}^{[\tau]}_Y(x)$ in Definition 1, i.e., the optimization problem with a fixed σ in (5) cannot be directly applied to the regression of $\mathrm{MI}_{\alpha,Y}(x)$.

To the best of the authors' knowledge, the number of existing papers on regression of $\mathrm{MI}_{\alpha,Y}(x)$ is relatively small, and the existing papers are mainly either *simple line charts* [33]–[35], *kernel density estimation methods* [36], [37], *local linear regression methods* [38], [39], or *random forest methods* [40], [41]. These existing methods are still insufficient to flexibly express the true smooth curves $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ especially when N is not so large. In the next section, we propose a regression method of $\mathrm{MI}_{\alpha,Y}(x)$, where we reconstruct both $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ as smooth spline functions.

3.2 Proposed Modal Interval Regression Method

To express the $100\alpha\%$ conditional modal intervals $\mathrm{MI}_{\alpha,Y}(x)$ of various shapes flexibly, we propose to estimate both lower curve $m^{\mathrm{low}}_{\alpha,Y}(x)$ and upper curve $m^{\mathrm{up}}_{\alpha,Y}(x)$ as smooth spline functions. Differently from estimation of $m^{[\tau]}_Y(x)$, it is very difficult to directly estimate $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ by extending the nonconvex optimization problem in (5). Instead, by expressing the modal interval locally with quantiles, we propose a two-step regression method based on spline quantile regression that avoids complicated nonconvex optimization problems as in (5). The proposed modal interval regression consists of the following *local empirical modal interval detection step* and *simultaneous spline quantile regression step*.

1: Divide bivariate observations $\{(x_n, y_n)\}_{n=1}^N$ into b bins according to the values of x_n. For $j = 1, 2, \ldots, b$, detect the local $100\alpha\%$ empirical modal interval $[\hat{y}^{\langle j\rangle}_{\mathrm{low}}, \hat{y}^{\langle j\rangle}_{\mathrm{up}}]$ of Y in the jth bin. Then, express the lower point $\hat{y}^{\langle j\rangle}_{\mathrm{low}}$ as the local $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$th empirical quantile of Y and the upper point $\hat{y}^{\langle j\rangle}_{\mathrm{up}}$ as the local $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$th empirical quantile of Y.

2: Let $\underline{s}(x)$ and $\bar{s}(x)$ be spline regression models for the lower curve $m^{\mathrm{low}}_{\alpha,Y}(x)$ and the upper curve $m^{\mathrm{up}}_{\alpha,Y}(x)$, respectively. Find $\underline{s}^*(x)$ and $\bar{s}^*(x)$ simultaneously, which minimize a weighted sum of the asymmetric absolute errors with $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$ and the roughness penalty terms under the *non-crossing constraint* s.t. $\forall x\ \underline{s}(x) \le \bar{s}(x)$.

For simplicity, in what follows, the bins in the first step construct a partition $\sqcup_b := \{I_j\}_{j=1}^b$ used in the definition of spline functions $\underline{s}(x)$ and $\bar{s}(x)$, i.e., each bin corresponds to each subinterval $I_j := [\xi_{j-1}, \xi_j]$. Moreover, we suppose that $\xi_0 \le x_{\min} < x_{\max} \le \xi_b$ and $\forall n\ x_n \ne \xi_j$ $(j = 1, 2, \ldots, b-1)$ so that the division of the observations $\{(x_n, y_n)\}_{n=1}^N$ into the bins can be uniquely determined. Define $\mathcal{B}_j := \{n \,|\, x_n \in I_j\}$ $(j = 1, 2, \ldots, b)$ as the set of the indices for the observations in the jth bin. The number of elements of $\mathcal{B}_j$ is denoted by N_j s.t. $\sum_{j=1}^b N_j = N$, where each N_j should not be too small.

In the first step, after the division of $\{(x_n, y_n)\}_{n=1}^N$ into b bins, we detect the local $100\alpha\%$ empirical modal interval in each bin as follows. We suppose that the values of y_n are different for all $n \in \mathcal{B}_j$ for simplicity, and sort $\{y_n\}_{n\in\mathcal{B}_j}$ as $y^{\langle j\rangle}_1 < y^{\langle j\rangle}_2 < \cdots < y^{\langle j\rangle}_{N_j}$. The *empirical cumulative distribution function of* Y in the jth bin is defined by

$$\hat{F}_{Y|\xi_{j-1}\le x\le\xi_j}(y) := \begin{cases} 0 & \text{if } y_{\inf} < y < y^{\langle j\rangle}_1, \\ \dfrac{n}{N_j} & \text{if } y^{\langle j\rangle}_n \le y < y^{\langle j\rangle}_{n+1} \\ & (n = 1, 2, \ldots, N_j - 1), \\ 1 & \text{if } y^{\langle j\rangle}_{N_j} \le y < y_{\sup}. \end{cases} \tag{11}$$

From (11), we see that as the index n of $y^{\langle j\rangle}_n$ increases by 1, the value of $\hat{F}_{Y|\xi_{j-1}\le x\le\xi_j}(y)$ increases by $\frac{1}{N_j}$. Hence, to increase the the value of $\hat{F}_{Y|\xi_{j-1}\le x\le\xi_j}(y)$ by α, we only have to increase the index n of $y^{\langle j\rangle}_n$ by $\lceil \alpha N_j \rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. By finding the following optimal index

$$n^*_j := \operatorname*{argmin}_n\ y^{\langle j\rangle}_{n+\lceil \alpha N_j\rceil} - y^{\langle j\rangle}_n, \tag{12}$$

the $100\alpha\%$ empirical modal interval of Y in the jth bin can be defined as

$$\left[\hat{y}^{\langle j\rangle}_{\mathrm{low}},\ \hat{y}^{\langle j\rangle}_{\mathrm{up}}\right] := \left[y^{\langle j\rangle}_{n^*_j},\ y^{\langle j\rangle}_{n^*_j+\lceil \alpha N_j\rceil}\right]. \tag{13}$$

As a modal interval regression method with $[\hat{y}^{\langle j\rangle}_{\mathrm{low}}, \hat{y}^{\langle j\rangle}_{\mathrm{up}}]$ in (13), we can consider the following spline smoothing

$$\underset{\underline{s}\in C^2(\xi_0,\xi_b)}{\text{minimize}} \sum_{j=1}^b \left|\hat{y}^{\langle j\rangle}_{\mathrm{low}} - \underline{s}\left(\tfrac{\xi_{j-1}+\xi_j}{2}\right)\right|^2 + \lambda \int_I |\underline{s}''(x)|^2\,\mathrm{d}x \tag{14}$$

for estimation of $m^{\mathrm{low}}_{\alpha,Y}(x)$, where the solution $\underline{s}^*(x)$ is a natural cubic spline function. A similar problem whose solution $\bar{s}^*(x)$ is a natural cubic spline function is considered for estimation of $m^{\mathrm{up}}_{\alpha,Y}(x)$. However, the reliability of this method is not so high when the number N of observations is not large since the values of $\hat{y}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{y}^{\langle j\rangle}_{\mathrm{up}}$ may greatly change depending on the bin width h_j. In the worst case, the two regression

results $\underline{s}^*(x)$ and $\bar{s}^*(x)$ cross each other, even though the true curves $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ satisfy $\forall x\, m^{\mathrm{low}}_{\alpha,Y}(x) < m^{\mathrm{up}}_{\alpha,Y}(x)$.

To realize more reliable regression, we propose to represent $\hat{y}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{y}^{\langle j\rangle}_{\mathrm{up}}$ in (13) as the $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$th and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$th empirical quantiles of Y and use $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$ in the second step, where

$$\hat{p}^{\langle j\rangle}_{\mathrm{low}} := \frac{n^*_j}{N_j} \quad \text{and} \quad \hat{p}^{\langle j\rangle}_{\mathrm{up}} := \frac{n^*_j + \lceil \alpha N_j \rceil}{N_j}. \tag{15}$$

To summarize, the first step of the proposed method ends by computing (12) and (15) after sorting $\{y_n\}_{n\in\mathcal{B}_j}$ for each j.

In the second step, we simultaneously estimate $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ with spline quantile regression similar to [42]. Specifically, we solve the following optimization problem

$$\begin{aligned}\underset{\underline{s},\bar{s}\in\mathcal{S}^{\rho}_d(\sqcup_b)}{\text{minimize}} & \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n \mathcal{J}_{\hat{p}^{\langle j\rangle}_{\mathrm{low}}}(y_n - \underline{s}(x_n)) + \lambda\int_I |\underline{s}''(x)|^2\,\mathrm{d}x \\ & + \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n \mathcal{J}_{\hat{p}^{\langle j\rangle}_{\mathrm{up}}}(y_n - \bar{s}(x_n)) + \lambda\int_I |\bar{s}''(x)|^2\,\mathrm{d}x \\ \text{subject to } & \forall x\ \underline{s}(x) + \delta \le \bar{s}(x),\ (y_{\inf} \le \underline{s}(x) \text{ and } \bar{s}(x) \le y_{\sup}),\end{aligned} \tag{16}$$

where $\delta \ge 0$ is the lower bound of $\bar{s}(x) - \underline{s}(x)$, which can be set to any nonnegative value, and the optional constraints $\forall x\ y_{\inf} \le \underline{s}(x)$ and/or $\forall x\ \bar{s}(x) \le y_{\sup}$ can be enforced if $y_{\inf}$ and/or $y_{\sup}$ are† known finite values. There is a possibility that the non-crossing property $\forall x\ \underline{s}^*(x) + \delta \le \bar{s}^*(x)$ will be automatically satisfied even if we find the optimal solutions $\underline{s}^*(x)$ and $\bar{s}^*(x)$ separately without the constraint, but it cannot be confirmed before finding both solutions. Meanwhile, the optimization problem in (16) always guarantees the non-crossing property by finding $\underline{s}^*(x)$ and $\bar{s}^*(x)$ simultaneously.

The optimization problem in (16) can be understood as follows. At each $x \in (x_{\inf}, x_{\sup})$, the lower and upper points $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ in (10) can be locally expressed as the $p_{\mathrm{low}}(x)$th and $p_{\mathrm{up}}(x)$th quantiles of Y, respectively, where

$$p_{\mathrm{low}}(x) := F_{Y|x}\big(m^{\mathrm{low}}_{\alpha,Y}(x)\big) \quad \text{and} \quad p_{\mathrm{up}}(x) := F_{Y|x}\big(m^{\mathrm{up}}_{\alpha,Y}(x)\big). \tag{17}$$

If $p_{\mathrm{low}}(x)$ in (17) is given, for estimation of the lower curve $m^{\mathrm{low}}_{\alpha,Y}(x)$, we can consider the following problem based on (3)

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{n=1}^{N} \mathcal{J}_{p_{\mathrm{low}}(x_n)}(y_n - r_{\boldsymbol{\theta}}(x_n)) \tag{18}$$

with a regression model $r_{\boldsymbol{\theta}}(x)$. Indeed, when N approaches infinity, the solution $r_{\boldsymbol{\theta}^*}(x)$ to (18) converges to $m^{\mathrm{low}}_{\alpha,Y}(x)$:

$$r_{\boldsymbol{\theta}^*}(x) \xrightarrow{N\to\infty} m^{\mathrm{low}}_{\alpha,Y}(x) \text{ satisfying } F_{Y|x}\big(m^{\mathrm{low}}_{\alpha,Y}(x)\big) = p_{\mathrm{low}}(x)$$

under the assumption that $r_{\boldsymbol{\theta}}(x)$ can exactly express $m^{\mathrm{low}}_{\alpha,Y}(x)$ if we choose the appropriate $\boldsymbol{\theta}$. A similar optimization problem can be considered for estimation of $m^{\mathrm{up}}_{\alpha,Y}(x)$ with $p_{\mathrm{up}}(x)$. However, the ideal problem in (18) cannot be used because $p_{\mathrm{low}}(x)$ is usually unknown as well as $F_{Y|x}(y)$ and $m^{\mathrm{low}}_{\alpha,Y}(x)$. Therefore, in the first step of the proposed method, we estimate $p_{\mathrm{low}}(x)$ and $p_{\mathrm{up}}(x)$ in (17) as $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$ in (15) for each bin. Then, in the second step, we estimate $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ simultaneously in (16) as flexible spline regression models $\underline{s}(x)$ and $\bar{s}(x)$ satisfying the non-crossing property, where the roughness penalties are added for $\underline{s}(x)$ and $\bar{s}(x)$ to avoid overfitting and reduce the influence of splitting by bins.

Although the values of $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$ in (15) change depending on the bin width h_j as well as $\hat{y}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{y}^{\langle j\rangle}_{\mathrm{up}}$, it can be considered that $\hat{p}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j\rangle}_{\mathrm{up}}$ are more robust to the bin width than $\hat{y}^{\langle j\rangle}_{\mathrm{low}}$ and $\hat{y}^{\langle j\rangle}_{\mathrm{up}}$. The reason is $p_{\mathrm{low}}(x)$ and $p_{\mathrm{low}}(x)$ slowly change along the x-axis in most cases, regardless of how fast $m^{\mathrm{low}}_{\alpha,Y}(x)$ and $m^{\mathrm{up}}_{\alpha,Y}(x)$ change. In fact, in [58], [62], by implicitly assuming $p_{\mathrm{mode}}(x) := F_{Y|x}(m_Y(x))$ $(x \in (x_{\inf}, x_{\sup}))$ slowly changes along the x-axis, a modal regression method *based on linear quantile regression* is proposed. As a result, we obtain more reliable results than (14) by the simultaneous quantile regression using all data $\{(x_n, y_n)\}_{n=1}^N$ and the functions $\mathcal{J}_{\hat{p}^{\langle j\rangle}_{\mathrm{low}}}(t)$ and $\mathcal{J}_{\hat{p}^{\langle j\rangle}_{\mathrm{up}}}(t)$ under the non-crossing constraint. In the next section, we express the optimization problem on spline functions $\underline{s}(x)$ and $\bar{s}(x)$ in (16) as a *convex* optimization problem on the coefficient vectors of $\underline{s}(x)$ and $\bar{s}(x)$.

3.3 Convex Optimization Problem on Coefficients

We express each spline function $s \in \mathcal{S}^{\rho}_d(\sqcup_b)$ in the following interval normalization form

$$s(x) = u_j(x) = c^{\langle j\rangle}_0 + \sum_{k=1}^{d} c^{\langle j\rangle}_k \Big(\frac{x - \xi_{j-1}}{h_j}\Big)^k \quad \text{for } x \in I_j, \tag{19}$$

where $c^{\langle j\rangle}_k \in \mathbb{R}$ $(k = 0, 1, \ldots, d)$ is the kth coefficient of the jth polynomial piece $u_j \in \mathbb{P}_d$ $(j = 1, 2, \ldots, b)$. Define $\boldsymbol{c} := (c^{\langle 1\rangle}_0, c^{\langle 1\rangle}_1, \ldots, c^{\langle 1\rangle}_d, c^{\langle 2\rangle}_0, c^{\langle 2\rangle}_1, \ldots, c^{\langle 2\rangle}_d, \ldots, c^{\langle b\rangle}_d)^{\mathrm{T}} \in \mathbb{R}^{b(d+1)}$ as the coefficient vector of the spline function $s(x)$. Since the spline function $s(x)$ is ρ-times continuously differentiable on $I = (\xi_0, \xi_b)$, i.e., $s \in C^{\rho}(\xi_0, \xi_b)$, every pair of coefficients of adjacent polynomial pieces $u_j \in \mathbb{P}_d$ and $u_{j+1} \in \mathbb{P}_d$ $(j = 1, 2, \ldots, b-1)$ has to satisfy the following linear equation

$$\begin{aligned}&\frac{1}{h^l_j}\sum_{k=l}^{d}\frac{k!}{(k-l)!}c^{\langle j\rangle}_k - \frac{l!}{h^l_{j+1}}c^{\langle j+1\rangle}_l = 0 \quad (l = 0, 1, \ldots, \rho) \\ &\Leftrightarrow \quad u^{(l)}_j(\xi_j) = u^{(l)}_{j+1}(\xi_j) \quad (l = 0, 1, \ldots, \rho),\end{aligned} \tag{20}$$

where $u^{(l)}_j(x)$ is the lth derivative of $u_j(x)$. From (20), there exists a matrix $\boldsymbol{H} \in \mathbb{R}^{(b-1)(\rho+1)\times b(d+1)}$ s.t. $\boldsymbol{Hc} = \boldsymbol{0} \Leftrightarrow s \in \mathcal{S}^{\rho}_d(\sqcup_b)$, where $\boldsymbol{0}$ denotes a zero vector of appropriate size.

First, we express the data fidelity term in (16) by using the coefficient vector $\boldsymbol{c}$. The value of $s(x_n)$ s.t. $x_n \in I_j$ can be expressed as $\boldsymbol{a}^{\mathrm{T}}_n\boldsymbol{c}$ with a vector $\boldsymbol{a}_n \in \mathbb{R}^{b(d+1)}$ defined by

†If $y_{\inf} = -\infty$ and/or $y_{\sup} = \infty$, these options are not needed.

$$\boldsymbol{a}_n := \Big(\underbrace{0,0,\ldots,0}_{(j-1)(d+1)}, 1, \tfrac{x_n-\xi_{j-1}}{h_j}, \ldots, \tfrac{(x_n-\xi_{j-1})^d}{h_j^d}, \underbrace{0,0,\ldots,0}_{(b-j)(d+1)}\Big)^{\mathrm{T}}.$$

As a result, we have $\mathcal{J}_p(y_n - s(x_n)) = \mathcal{J}_p(y_n - \boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{c})$.

Second, we express the roughness penalty term in (16) with the coefficient vector $\boldsymbol{c}$. The roughness penalty over the interval I is decomposed into those over the subintervals I_j:

$$\int_I |s''(x)|^2\,\mathrm{d}x = \sum_{j=1}^{b}\int_{I_j} |u_j''(x)|^2\,\mathrm{d}x. \tag{21}$$

From (19), each penalty over I_j is written by a quadratic form

$$\int_{I_j} |u_j''(x)|^2\,\mathrm{d}x = \sum_{k=2}^{d}\sum_{l=2}^{d} \frac{k(k-1)l(l-1)}{h_j^3(k+l-3)}\, c_k^{\langle j\rangle} c_l^{\langle j\rangle}. \tag{22}$$

From (21) and (22), there is a symmetric positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{b(d+1)\times b(d+1)}$ s.t. $\int_I |s''(x)|^2\,\mathrm{d}x = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{c}$.

Third, we express the non-crossing constraint and the two optional domain constraints in (16) with the coefficient vector $\boldsymbol{c}$. In fact, it is very difficult to give a useful necessary and sufficient condition for each shape constraint. Instead, we derive a useful sufficient condition for each shape constraint in a manner similar to [42]. In [42], the second author of the current paper derived sufficient conditions for several shape (non-decreasing, non-increasing, convex and concave) constraints by extending the sufficient condition in [63]–[65] for the nonnegativity of the spline function $s(x)$ over I_j:

$$\sum_{k=0}^{l} \frac{(d-k)!}{(l-k)!(d-l)!}\, c_k^{\langle j\rangle} \ge 0 \quad (l = 0,1,\ldots,d)$$
$$\Rightarrow \quad s(x) \ge 0 \text{ for all } x \in I_j. \tag{23}$$

From (23), we can express a sufficient condition for the nonnegativity over I, with a certain matrix $\boldsymbol{G} \in \mathbb{R}^{b(d+1)\times b(d+1)}$, as a linear inequality $\boldsymbol{G}\boldsymbol{c} \ge \boldsymbol{0} \Rightarrow s(x) \ge 0$ for all $x \in I$.

Let $\underline{c}_k^{\langle j\rangle} \in \mathbb{R}$ and $\bar{c}_k^{\langle j\rangle} \in \mathbb{R}$ be coefficients of $\underline{s}(x)$ and $\bar{s}(x)$, respectively. The non-crossing constraint is expressed as $\forall x\ \bar{s}(x) - \underline{s}(x) - \delta \ge 0$, which is also seen as the nonnegativity of a spline function $\bar{s}(x) - \underline{s}(x) - \delta$. Coefficients of the jth polynomial piece of $\bar{s}(x) - \underline{s}(x) - \delta$ are given by $(\bar{c}_0^{\langle j\rangle} - \underline{c}_0^{\langle j\rangle} - \delta, \bar{c}_1^{\langle j\rangle} - \underline{c}_1^{\langle j\rangle}, \bar{c}_2^{\langle j\rangle} - \underline{c}_2^{\langle j\rangle}, \ldots, \bar{c}_d^{\langle j\rangle} - \underline{c}_d^{\langle j\rangle})^{\mathrm{T}}$. Thus, from (23), we have the following sufficient condition

$$\sum_{k=0}^{l} \frac{(d-k)!\,\big(\bar{c}_k^{\langle j\rangle} - \underline{c}_k^{\langle j\rangle}\big)}{(l-k)!(d-l)!} \ge \delta\,\frac{d!}{l!(d-l)!} \quad (l = 0,1,\ldots,d)$$
$$\Rightarrow \quad \bar{s}(x) \ge \underline{s}(x) + \delta \text{ for all } x \in I_j. \tag{24}$$

The optional constraints are also seen as the nonnegativity of spline functions $\underline{s}(x) - y_{\inf}$ and $-\bar{s}(x) + y_{\sup}$, and we have

$$\sum_{k=0}^{l} \frac{(d-k)!}{(l-k)!(d-l)!}\, \underline{c}_k^{\langle j\rangle} \ge y_{\inf}\,\frac{d!}{l!(d-l)!} \quad (l = 0,1,\ldots,d)$$
$$\Rightarrow \quad \underline{s}(x) \ge y_{\inf} \text{ for all } x \in I_j \tag{25}$$

and

$$\sum_{k=0}^{l} \frac{(d-k)!}{(l-k)!(d-l)!}\, \bar{c}_k^{\langle j\rangle} \le y_{\sup}\,\frac{d!}{l!(d-l)!} \quad (l = 0,1,\ldots,d)$$
$$\Rightarrow \quad \bar{s}(x) \le y_{\sup} \text{ for all } x \in I_j. \tag{26}$$

By defining $\bar{\underline{\boldsymbol{c}}} := (\underline{\boldsymbol{c}}^{\mathrm{T}}, \bar{\boldsymbol{c}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{2b(d+1)}$ from both coefficient vectors $\underline{\boldsymbol{c}}, \bar{\boldsymbol{c}} \in \mathbb{R}^{b(d+1)}$ of $\underline{s}, \bar{s} \in \mathcal{S}_d^{\rho}(\sqcup_b)$, the sufficient conditions in (24), (25) and (26) are expressed as $\boldsymbol{G}(\bar{\boldsymbol{c}} - \underline{\boldsymbol{c}}) \ge \delta\boldsymbol{\zeta}$, $\boldsymbol{G}\underline{\boldsymbol{c}} \ge y_{\inf}\boldsymbol{\zeta}$ and $\boldsymbol{G}\bar{\boldsymbol{c}} \le y_{\sup}\boldsymbol{\zeta}$ with some vector $\boldsymbol{\zeta} \in \mathbb{R}^{b(d+1)}$.

To summarize, the problem in (16) is expressed as the following *finite-dimensional convex optimization problem*

$$\begin{aligned}
&\underset{\bar{\underline{\boldsymbol{c}}}\in\mathbb{R}^{2b(d+1)}}{\text{minimize}} \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n \mathcal{J}_{\hat{p}_{\text{low}}^{\langle j\rangle}}\big(y_n - \boldsymbol{a}_n^{\mathrm{T}}\underline{\boldsymbol{c}}\big) + \lambda\underline{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{Q}\underline{\boldsymbol{c}} \\
&\qquad + \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n \mathcal{J}_{\hat{p}_{\text{up}}^{\langle j\rangle}}\big(y_n - \boldsymbol{a}_n^{\mathrm{T}}\bar{\boldsymbol{c}}\big) + \lambda\bar{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{Q}\bar{\boldsymbol{c}} \\
&\text{subject to} \quad \boldsymbol{H}\underline{\boldsymbol{c}} = \boldsymbol{0},\ \boldsymbol{H}\bar{\boldsymbol{c}} = \boldsymbol{0}, \text{ and } \boldsymbol{G}(\bar{\boldsymbol{c}} - \underline{\boldsymbol{c}}) \ge \delta\boldsymbol{\zeta}, \\
&(\text{and optionally} \quad \boldsymbol{G}\underline{\boldsymbol{c}} \ge y_{\inf}\boldsymbol{\zeta} \text{ and/or } \boldsymbol{G}\bar{\boldsymbol{c}} \le y_{\sup}\boldsymbol{\zeta}).
\end{aligned} \tag{27}$$

In the next section, for the convex optimization problem in (27), we give two solvers with different formulations based on the quadratic programming [43], [44] and the alternating direction method of multipliers (ADMM) [45]–[49].

3.4 Optimization Algorithms

First of all, in addition to $\boldsymbol{y} := (y_1, y_2, \ldots, y_N)^{\mathrm{T}} \in \mathbb{R}^N$ and $\boldsymbol{A} := (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N)^{\mathrm{T}} \in \mathbb{R}^{N\times b(d+1)}$, we define

$$\bar{\boldsymbol{G}} := \begin{pmatrix} -\boldsymbol{G} & \boldsymbol{G} \\ \boldsymbol{G} & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{G} \end{pmatrix} \quad \text{and} \quad \bar{\underline{\boldsymbol{\zeta}}} := \begin{pmatrix} \delta\boldsymbol{\zeta} \\ y_{\inf}\boldsymbol{\zeta} \\ -y_{\sup}\boldsymbol{\zeta} \end{pmatrix},$$

where $\boldsymbol{O}$ denotes a zero matrix of appropriate size, and the definitions of $\bar{\boldsymbol{G}}$ and $\bar{\underline{\boldsymbol{\zeta}}}$ changes depending on whether the optional constrains are used or not. Furthermore, we define the upper-side error vector $\underline{\boldsymbol{\epsilon}}_+ := (\underline{\epsilon}_1^+, \underline{\epsilon}_2^+, \ldots, \underline{\epsilon}_N^+)^{\mathrm{T}} \in \mathbb{R}^N$ and the lower-side error vector $\underline{\boldsymbol{\epsilon}}_- := (\underline{\epsilon}_1^-, \underline{\epsilon}_2^-, \ldots, \underline{\epsilon}_N^-)^{\mathrm{T}} \in \mathbb{R}^N$ for $\underline{s}(x)$, where $\underline{\epsilon}_n^+ \ge 0$ and $\underline{\epsilon}_n^- \le 0$ for all n. Similarly, we define the upper-side and lower-side error vectors for $\bar{s}(x)$ as $\bar{\boldsymbol{\epsilon}}_+ := (\bar{\epsilon}_1^+, \bar{\epsilon}_2^+, \ldots, \bar{\epsilon}_N^+)^{\mathrm{T}} \in \mathbb{R}^N$ and $\bar{\boldsymbol{\epsilon}}_- := (\bar{\epsilon}_1^-, \bar{\epsilon}_2^-, \ldots, \bar{\epsilon}_N^-)^{\mathrm{T}} \in \mathbb{R}^N$, where $\bar{\epsilon}_n^+ \ge 0$ and $\bar{\epsilon}_n^- \le 0$. Then, the proposed problem in (27) is expressed as a quadratic programming problem

$$\begin{aligned}
&\underset{\bar{\underline{\boldsymbol{c}}},\underline{\boldsymbol{\epsilon}}_+,\underline{\boldsymbol{\epsilon}}_-,\bar{\boldsymbol{\epsilon}}_+,\bar{\boldsymbol{\epsilon}}_-}{\text{minimize}} \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n\Big(\hat{p}_{\text{low}}^{\langle j\rangle}\underline{\epsilon}_n^+ - \big(1-\hat{p}_{\text{low}}^{\langle j\rangle}\big)\underline{\epsilon}_n^-\Big) + \lambda\underline{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{Q}\underline{\boldsymbol{c}} \\
&\qquad + \sum_{j=1}^{b}\sum_{n\in\mathcal{B}_j} w_n\Big(\hat{p}_{\text{up}}^{\langle j\rangle}\bar{\epsilon}_n^+ - \big(1-\hat{p}_{\text{up}}^{\langle j\rangle}\big)\bar{\epsilon}_n^-\Big) + \lambda\bar{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{Q}\bar{\boldsymbol{c}} \\
&\text{subject to} \quad \boldsymbol{H}\underline{\boldsymbol{c}} = \boldsymbol{0},\ \boldsymbol{H}\bar{\boldsymbol{c}} = \boldsymbol{0},\ \bar{\boldsymbol{G}}\bar{\underline{\boldsymbol{c}}} \ge \bar{\underline{\boldsymbol{\zeta}}},\ \underline{\boldsymbol{\epsilon}}_+ \ge \boldsymbol{0},\ \underline{\boldsymbol{\epsilon}}_- \le \boldsymbol{0}, \\
&\bar{\boldsymbol{\epsilon}}_+ \ge \boldsymbol{0},\ \bar{\boldsymbol{\epsilon}}_- \le \boldsymbol{0},\ \boldsymbol{y} - \boldsymbol{A}\underline{\boldsymbol{c}} = \underline{\boldsymbol{\epsilon}}_+ + \underline{\boldsymbol{\epsilon}}_-, \text{ and } \boldsymbol{y} - \boldsymbol{A}\bar{\boldsymbol{c}} = \bar{\boldsymbol{\epsilon}}_+ + \bar{\boldsymbol{\epsilon}}_-.
\end{aligned} \tag{28}$$

The problem in (28) is solved by a quadratic programming solver, e.g., by one of interior-point methods [43], [44].

Next, we explain an ADMM formulation for the problem in (27). We define three block-diagonal matrices

$$\bar{\underline{A}} := \begin{pmatrix} \underline{A} & \underline{O} \\ \underline{O} & \underline{A} \end{pmatrix}, \quad \bar{\underline{H}} := \begin{pmatrix} \underline{H} & \underline{O} \\ \underline{O} & \underline{H} \end{pmatrix}, \quad \text{and} \quad \bar{\underline{Q}} := \begin{pmatrix} \underline{Q} & \underline{O} \\ \underline{O} & \underline{Q} \end{pmatrix}.$$

Moreover, we add two auxiliary vectors $\boldsymbol{z}_1 \in \mathbb{R}^{2N}$ and $\boldsymbol{z}_2 \in \mathbb{R}^{3b(d+1)}$. Then, we have the following ADMM formulation

$$\begin{aligned} &\underset{\bar{\underline{\boldsymbol{c}}},\boldsymbol{z}_1,\boldsymbol{z}_2}{\text{minimize}} \ \lambda \bar{\underline{\boldsymbol{c}}}^{\mathrm{T}} \bar{\underline{\boldsymbol{Q}}} \bar{\underline{\boldsymbol{c}}} + \iota_{=\boldsymbol{0}}(\bar{\underline{\boldsymbol{H}}} \bar{\underline{\boldsymbol{c}}}) + \bar{\mathcal{J}}^{\boldsymbol{w}}_{\bar{\underline{\boldsymbol{p}}},\boldsymbol{y}}(\boldsymbol{z}_1) + \iota_{\geq \bar{\underline{\boldsymbol{\zeta}}}}(\boldsymbol{z}_2) \\ &\text{subject to} \qquad \boldsymbol{z}_1 = \bar{\underline{\boldsymbol{A}}} \bar{\underline{\boldsymbol{c}}} \text{ and } \boldsymbol{z}_2 = \bar{\underline{\boldsymbol{G}}} \bar{\underline{\boldsymbol{c}}}. \end{aligned} \tag{29}$$

Here the two indicator functions are defined as $\iota_{=\boldsymbol{0}}(\bar{\underline{\boldsymbol{H}}} \bar{\underline{\boldsymbol{c}}}) := 0$ if $\bar{\underline{\boldsymbol{H}}} \bar{\underline{\boldsymbol{c}}} = \boldsymbol{0}$, and $\iota_{=\boldsymbol{0}}(\bar{\underline{\boldsymbol{H}}} \bar{\underline{\boldsymbol{c}}}) := \infty$ otherwise, and $\iota_{\geq \bar{\underline{\boldsymbol{\zeta}}}}(\boldsymbol{z}_2) := 0$ if $\boldsymbol{z}_2 \geq \bar{\underline{\boldsymbol{\zeta}}}$, and $\iota_{\geq \bar{\underline{\boldsymbol{\zeta}}}}(\boldsymbol{z}_2) := \infty$ otherwise. The proximable convex function $\bar{\mathcal{J}}^{\boldsymbol{w}}_{\bar{\underline{\boldsymbol{p}}},\boldsymbol{y}} : \mathbb{R}^{2N} \to [0, \infty)$ is defined by

$$\bar{\mathcal{J}}^{\boldsymbol{w}}_{\bar{\underline{\boldsymbol{p}}},\boldsymbol{y}}(\boldsymbol{z}) := \sum_{j=1}^{b} \sum_{n \in \mathcal{B}_j} w_n \Big(\mathcal{J}_{\hat{p}^{\langle j \rangle}_{\mathrm{low}}}(y_n - z_n) + \mathcal{J}_{\hat{p}^{\langle j \rangle}_{\mathrm{up}}}(y_n - z_{N+n}) \Big)$$

for every $\boldsymbol{z} = (z_1, z_2, \ldots, z_N, z_{N+1}, z_{N+2}, \ldots, z_{2N})^{\mathrm{T}} \in \mathbb{R}^{2N}$, where its *proximity operator* $\operatorname{prox}_{\bar{\mathcal{J}}^{\boldsymbol{w}}_{\bar{\underline{\boldsymbol{p}}},\boldsymbol{y}}} : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ can be computed for each component with the following equation

$$\operatorname{prox}_{w \mathcal{J}_p(y - \cdot)}(z) = \begin{cases} z + pw & \text{if } y - z \geq pw, \\ z - (1-p)w & \text{if } y - z \leq -(1-p)w, \\ y & \text{otherwise.} \end{cases}$$

The proximity operator $\operatorname{prox}_{\iota_{\geq \bar{\underline{\boldsymbol{\zeta}}}}} : \mathbb{R}^{3b(d+1)} \to \mathbb{R}^{3b(d+1)}$ of the indicator function can also be computed for each component as the *projection operator*

$$\operatorname{prox}_{\iota_{\geq \zeta}}(z) = \operatorname{projection}_{\geq \zeta}(z) = \begin{cases} z & \text{if } z \geq \zeta, \\ \zeta & \text{if } z < \zeta. \end{cases}$$

By defining a matrix $\boldsymbol{\Gamma}$ with the ADMM parameter $\gamma > 0$ as

$$\boldsymbol{\Gamma} := 2\gamma\lambda \bar{\underline{\boldsymbol{Q}}} + \bar{\underline{\boldsymbol{A}}}^{\mathrm{T}} \bar{\underline{\boldsymbol{A}}} + \bar{\underline{\boldsymbol{G}}}^{\mathrm{T}} \bar{\underline{\boldsymbol{G}}},$$

we can summarize the ADMM-based spline modal interval regression as Algorithm 1, where i is the index for iterations, and we input the number of the iterations as an integer Iter.

Note that whether we solve the problem in (28) or (29), we will obtain almost the same regression results $\underline{s}^*(x)$ and $\bar{s}^*(x)$ since the original proposed problem in (16) is convex. Comparing the problems in (28) and (29), the dimensions of the primal variables $(\bar{\underline{\boldsymbol{c}}}, \underline{\boldsymbol{\epsilon}}_+, \underline{\boldsymbol{\epsilon}}_-, \bar{\boldsymbol{\epsilon}}_+, \bar{\boldsymbol{\epsilon}}_-)$ in (28) and $(\bar{\underline{\boldsymbol{c}}}, \boldsymbol{z}_1, \boldsymbol{z}_2)$ in (29) are $2b(d+1)+4N$ and $5b(d+1)+2N$, respectively. In addition, in the interior-point method for (28), the dimension of the dual variables containing slack variables [66] to be updated is $(2(b-1)(\rho+1)+2N)+2(3b(d+1)+4N) = 2b(3d+\rho+4)+10N-2(\rho+1)$, while in Algorithm 1 for (29), the dimension of the dual variables[†] $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is $3b(d+1)+2N$.

[†]The vector $\boldsymbol{\eta} \in \mathbb{R}^{2b(d+1)}$ in Algorithm 1 is not a dual variable but an intermediate variable to write the equation in line 8 shortly.

Algorithm 1 Spline Modal Interval Regression with ADMM

Input: data $\{(x_n, y_n)\}_{n=1}^N$, probability $\alpha \in (0,1)$, partition (bins) $\sqcup_b = \{I_j\}_{j=1}^b$, degree d, smoothness ρ $(2 \leq \rho \leq d)$, positive weights $\boldsymbol{w} = (w_n)_{n=1}^N$, smoothing parameter $\lambda > 0$, number of updates Iter, ADMM parameter $\gamma > 0$, lower bound gap $\delta \geq 0$, and options $y_{\inf}$ and/or $y_{\sup}$.
Output: coefficients of lower and upper spline functions $\underline{s}(x)$ and $\bar{s}(x)$.

1: Divide the data $\{(x_n, y_n)\}_{n=1}^N$ into b bins according to values of x_n.
2: Sort values of y_n $(n \in \mathcal{B}_j)$ as $y_1^{\langle j \rangle} < y_2^{\langle j \rangle} < \cdots < y_{N_j}^{\langle j \rangle}$ for each j.
3: Find the optimal index n_j^* in (12) and compute $\hat{p}^{\langle j \rangle}_{\mathrm{low}}$ and $\hat{p}^{\langle j \rangle}_{\mathrm{up}}$ in (15).
4: Construct $\bar{\underline{\boldsymbol{A}}}, \bar{\underline{\boldsymbol{G}}}, \bar{\underline{\boldsymbol{H}}}, \bar{\underline{\boldsymbol{Q}}}, \boldsymbol{\Gamma}$, and $\bar{\underline{\boldsymbol{\zeta}}}$.
5: Set $\boldsymbol{z}_1, \boldsymbol{v}_1 \in \mathbb{R}^{2N}$ and $\boldsymbol{z}_2, \boldsymbol{v}_2 \in \mathbb{R}^{3b(d+1)}$ to any initial values.
6: **for** $i = 1, 2, \ldots,$ Iter **do**
7: $\quad \boldsymbol{\eta} \leftarrow \bar{\underline{\boldsymbol{A}}}^{\mathrm{T}}(\boldsymbol{z}_1 - \boldsymbol{v}_1) + \bar{\underline{\boldsymbol{G}}}^{\mathrm{T}}(\boldsymbol{z}_2 - \boldsymbol{v}_2)$
8: $\quad \bar{\underline{\boldsymbol{c}}} \leftarrow \boldsymbol{\Gamma}^{-1}\big(\boldsymbol{\eta} - \bar{\underline{\boldsymbol{H}}}^{\mathrm{T}}(\bar{\underline{\boldsymbol{H}}} \boldsymbol{\Gamma}^{-1} \bar{\underline{\boldsymbol{H}}}^{\mathrm{T}})^{-1} \bar{\underline{\boldsymbol{H}}} \boldsymbol{\Gamma}^{-1} \boldsymbol{\eta}\big)$
9: $\quad \boldsymbol{z}_1 \leftarrow \operatorname{prox}_{\bar{\mathcal{J}}^{\gamma \boldsymbol{w}}_{\bar{\underline{\boldsymbol{p}}},\boldsymbol{y}}}\big(\bar{\underline{\boldsymbol{A}}} \bar{\underline{\boldsymbol{c}}} + \boldsymbol{v}_1\big)$
10: $\quad \boldsymbol{z}_2 \leftarrow \operatorname{projection}_{\geq \bar{\underline{\boldsymbol{\zeta}}}}\big(\bar{\underline{\boldsymbol{G}}} \bar{\underline{\boldsymbol{c}}} + \boldsymbol{v}_2\big)$
11: $\quad \boldsymbol{v}_1 \leftarrow \boldsymbol{v}_1 + \bar{\underline{\boldsymbol{A}}} \bar{\underline{\boldsymbol{c}}} - \boldsymbol{z}_1$
12: $\quad \boldsymbol{v}_2 \leftarrow \boldsymbol{v}_2 + \bar{\underline{\boldsymbol{G}}} \bar{\underline{\boldsymbol{c}}} - \boldsymbol{z}_2$
13: **end for**
14: Return the optimal solution $\bar{\boldsymbol{c}} = (\underline{\boldsymbol{c}}^{\mathrm{T}}, \bar{\boldsymbol{c}}^{\mathrm{T}})^{\mathrm{T}}$.

Therefore, if $b(d+1) \ll N$, it is expected that the problem in (29) requires less memory than the problem in (28). Actually, sophisticated quadratic programming packages such as [66], that can adaptively reduce the dimension of variables, have been developed in many programming languages. We can try those packages first, and if there are problems like memory shortage, then use the iterations in Algorithm 1.

4. Numerical Experiments

In this section, we conduct numerical experiments to show the effectiveness of the proposed modal interval regression compared to the straightforward least squares method in (14) and two kernel density estimation methods introduced below.

4.1 Existing Methods Based on Kernel Density Estimation

Based on the methods [36], [37] originally used for analysis of time series data, we introduce two different kernel density estimation (KDE) methods that estimate the modal interval. One KDE method sorts $\{(x_n, y_n)\}_{n=1}^N$ as $\{(x_n^{\langle \mathrm{all} \rangle}, y_n^{\langle \mathrm{all} \rangle})\}_{n=1}^N$ s.t. $y_1^{\langle \mathrm{all} \rangle} < y_2^{\langle \mathrm{all} \rangle} < \cdots < y_N^{\langle \mathrm{all} \rangle}$ and reconstructs the conditional cumulative distribution function of Y given $X = x$ by

$$\hat{F}^{(\mathrm{KDE})}_{Y|x}(y) := \begin{cases} 0 & \text{if } y_{\inf} < y < y_1^{\langle \mathrm{all} \rangle}, \\ \dfrac{\sum_{l=1}^n \mathcal{K}\big(\frac{x - x_l^{\langle \mathrm{all} \rangle}}{\sigma}\big)}{\sum_{l=1}^N \mathcal{K}\big(\frac{x - x_l}{\sigma}\big)} & \text{if } y_n^{\langle \mathrm{all} \rangle} \leq y < y_{n+1}^{\langle \mathrm{all} \rangle} \\ & (n = 1, 2, \ldots, N-1), \\ 1 & \text{if } y_N^{\langle \mathrm{all} \rangle} \leq y < y_{\sup}, \end{cases} \tag{30}$$

where $\mathcal{K} : \mathbb{R} \to [0,1]$ is a nonnegative symmetric unimodal kernel satisfying the same conditions as in (5), and $\sigma > 0$ is its bandwidth.Then, in a manner similar to (12) and (13), the

$100\alpha\%$ conditional modal interval $\mathrm{MI}_{\alpha,Y}(x)$ is estimated as

$$\underset{[y_{n_1}^{\langle\mathrm{all}\rangle}, y_{n_2}^{\langle\mathrm{all}\rangle}]}{\operatorname{argmin}} \; y_{n_2}^{\langle\mathrm{all}\rangle} - y_{n_1}^{\langle\mathrm{all}\rangle} \;\text{ s.t. }\; \hat{F}_{Y|x}^{(\mathrm{KDE})}(y_{n_2}^{\langle\mathrm{all}\rangle}) - \hat{F}_{Y|x}^{(\mathrm{KDE})}(y_{n_1}^{\langle\mathrm{all}\rangle}) \geq \alpha \tag{31}$$

for given $X = x$. Since $\hat{F}_{Y|x}^{(\mathrm{KDE})}(y)$ in (30) is piecewise constant, the estimated modal interval by (31) is not continuous, i.e., both lower and upper points change like step functions. Algorithm 2 shows an implementation of this method.

On the other hand, another KDE method reconstructs the conditional probability density function of Y given X by

$$\hat{f}_{Y|X}^{(\mathrm{KDE})}(y|x) := \frac{1}{\sigma_Y} \sum_{n=1}^{N} \frac{\mathcal{K}\left(\frac{x-x_n}{\sigma_X}\right)}{\sum_{l=1}^{N} \mathcal{K}\left(\frac{x-x_l}{\sigma_X}\right)} \mathcal{K}\left(\frac{y-y_n}{\sigma_Y}\right), \tag{32}$$

where $\sigma_X > 0$ and $\sigma_Y > 0$ denote the bandwidths in the x and y directions [67], respectively. Then, based on (10), the $100\alpha\%$ conditional modal interval $\mathrm{MI}_{\alpha,Y}(x)$ is estimated as

$$\underset{[y_{\mathrm{low}}, y_{\mathrm{up}}] \subseteq \mathrm{cl}(y_{\mathrm{inf}}, y_{\mathrm{sup}})}{\operatorname{argmin}} \; y_{\mathrm{up}} - y_{\mathrm{low}} \;\text{ s.t. }\; \int_{y_{\mathrm{low}}}^{y_{\mathrm{up}}} \hat{f}_{Y|X}^{(\mathrm{KDE})}(y|x)\,\mathrm{d}y \geq \alpha \tag{33}$$

for given $X = x$. If we use the Gaussian kernel defined in (6) as $\mathcal{K}(t)$ in (32), the integral in (33) can be calculated by

$$\begin{aligned}\int_{y_{\mathrm{low}}}^{y_{\mathrm{up}}} \hat{f}_{Y|X}^{(\mathrm{KDE})}(y|x)\,\mathrm{d}y &= \sum_{n=1}^{N} \frac{\mathcal{K}\left(\frac{x-x_n}{\sigma_X}\right)}{\sum_{l=1}^{N} \mathcal{K}\left(\frac{x-x_l}{\sigma_X}\right)} \int_{\frac{y_{\mathrm{low}}-y_n}{\sigma_Y}}^{\frac{y_{\mathrm{up}}-y_n}{\sigma_Y}} \mathcal{K}(t)\,\mathrm{d}t \\ &= \frac{1}{2} \sum_{n=1}^{N} \frac{\mathcal{K}\left(\frac{x-x_n}{\sigma_X}\right)}{\sum_{l=1}^{N} \mathcal{K}\left(\frac{x-x_l}{\sigma_X}\right)} \left(\mathrm{erf}\left(\frac{y_{\mathrm{up}}-y_n}{\sqrt{2}\sigma_Y}\right) - \mathrm{erf}\left(\frac{y_{\mathrm{low}}-y_n}{\sqrt{2}\sigma_Y}\right)\right),\end{aligned} \tag{34}$$

where $\mathrm{erf}(y) := \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2}\,\mathrm{d}t$ is the *error function*, but even in this case, it is difficult to exactly solve the problem in (33). To find an approximate solution, we prepare candidates for y_{low} and y_{up} as $y_{\mathrm{cand},i} := y_{\mathrm{cand},0} + i\Delta y \in \mathrm{cl}(y_{\mathrm{inf}}, y_{\mathrm{sup}})$ (s.t. $i \in \mathbb{N}$ and $\Delta y > 0$) and compute $\int_{y_{\mathrm{cand},i_1}}^{y_{\mathrm{cand},i_2}} \hat{f}_{Y|X}^{(\mathrm{KDE})}(y|x)\,\mathrm{d}y$ by (34). Theoretically the estimated modal interval by (33) is continuous since $\hat{f}_{Y|X}^{(\mathrm{KDE})}(y|x)$ in (32) is continuous with respect to x and y, but in fact the discretization of the value of y into $y_{\mathrm{cand},i}$ is performed. If Δy is small, this discretization effect will be smaller, while the computation time will be longer. Algorithm 3 shows an implementation of this method.

4.2 Experiments for Synthetic Data

To validate the accuracy of each method for various bivariate data, we generate 100 datasets for each of six different distributions and estimate their 50% conditional modal intervals $\mathrm{MI}_{0.5,Y}(x)$ by the proposed method in (16), the least squares method in (14), and the two KDE methods in (31) and (33).

4.2.1 Simulated Distributions

Distribution 1 (Normal Y with uniform X): In the first distribution, we use a normal (Gaussian) distribution

Algorithm 2 KDE-Based Modal Interval Estimation in (31)

Input: data $\{(x_n, y_n)\}_{n=1}^N$, probability $\alpha \in (0,1)$, and bandwidth σ.
Output: lower and upper points y_{low} and y_{up} of $\mathrm{MI}_{\alpha,Y}(x)$ for some x.

Sort the data as $\{(x_n^{\langle\mathrm{all}\rangle}, y_n^{\langle\mathrm{all}\rangle})\}_{n=1}^N$ s.t. $y_1^{\langle\mathrm{all}\rangle} < y_2^{\langle\mathrm{all}\rangle} < \cdots < y_N^{\langle\mathrm{all}\rangle}$.
$w_n \leftarrow \mathcal{K}\left(\frac{x-x_n^{\langle\mathrm{all}\rangle}}{\sigma}\right)$ $(n = 1, 2, \ldots, N)$
$w_{\mathrm{total}} \leftarrow \sum_{n=1}^N w_n$, $w_{\mathrm{accum}} \leftarrow 0$
for $n = 1, 2, \ldots, N$ **do**
  $w_{\mathrm{accum}} \leftarrow w_{\mathrm{accum}} + w_n$
  $F_n \leftarrow w_{\mathrm{accum}} / w_{\mathrm{total}}$
end for
$L_{\mathrm{MI}} \leftarrow y_N^{\langle\mathrm{all}\rangle} - y_1^{\langle\mathrm{all}\rangle}$, $y_{\mathrm{low}} \leftarrow y_1^{\langle\mathrm{all}\rangle}$, $y_{\mathrm{up}} \leftarrow y_N^{\langle\mathrm{all}\rangle}$, $n_2 \leftarrow 2$, $n_1 \leftarrow 1$
while $F_{n_1} \leq 1 - \alpha$ **do**
  while $F_{n_2} - F_{n_1} < \alpha$ **do**
    $n_2 \leftarrow n_2 + 1$
  end while
  if $L_{\mathrm{MI}} > y_{n_2}^{\langle\mathrm{all}\rangle} - y_{n_1}^{\langle\mathrm{all}\rangle}$ **then**
    $L_{\mathrm{MI}} \leftarrow y_{n_2}^{\langle\mathrm{all}\rangle} - y_{n_1}^{\langle\mathrm{all}\rangle}$, $y_{\mathrm{low}} \leftarrow y_{n_1}^{\langle\mathrm{all}\rangle}$, $y_{\mathrm{up}} \leftarrow y_{n_2}^{\langle\mathrm{all}\rangle}$
  end if
  $n_2 \leftarrow n_1 + 2$, $n_1 \leftarrow n_1 + 1$
end while
Return the lower and upper points y_{low} and y_{up}.

Algorithm 3 KDE-Based Modal Interval Estimation in (33)

Input: data $\{(x_n, y_n)\}_{n=1}^N$, probability $\alpha \in (0,1)$, bandwidths σ_X and σ_Y, and candidates $\{y_{\mathrm{cand},i}\}_{i=0}^{I_{\max}}$ ($y_{\mathrm{cand},i} = y_{\mathrm{cand},0} + i\Delta y$ and $\Delta y > 0$).
Output: lower and upper points y_{low} and y_{up} of $\mathrm{MI}_{\alpha,Y}(x)$ for some x.

$w_n \leftarrow \mathcal{K}\left(\frac{x-x_n}{\sigma_X}\right)$ $(n = 1, 2, \ldots, N)$
$w_{\mathrm{total}} \leftarrow \sum_{n=1}^N w_n$
$F_i \leftarrow \frac{1}{2w_{\mathrm{total}}} \sum_{n=1}^N w_n \left(1 + \mathrm{erf}\left(\frac{y_{\mathrm{cand},i} - y_n}{\sqrt{2}\sigma_Y}\right)\right)$ $(i = 0, 1, \ldots, I_{\max})$
$L_{\mathrm{MI}} \leftarrow I_{\max}\Delta y$, $y_{\mathrm{low}} \leftarrow y_{\mathrm{cand},0}$, $y_{\mathrm{up}} \leftarrow y_{\mathrm{cand},I_{\max}}$, $i_2 \leftarrow 1$, $i_1 \leftarrow 0$
while $F_{i_1} \leq 1 - \alpha$ and $i_1 \leq I_{\max} - 1$ **do**
  while $F_{i_2} - F_{i_1} < \alpha$ and $i_2 \leq I_{\max}$ **do**
    $i_2 \leftarrow i_2 + 1$
  end while
  if $i_2 \leq I_{\max}$ and $L_{\mathrm{MI}} > (i_2 - i_1)\Delta y$ **then**
    $L_{\mathrm{MI}} \leftarrow (i_2 - i_1)\Delta y$, $y_{\mathrm{low}} \leftarrow y_{\mathrm{cand},i_1}$, $y_{\mathrm{up}} \leftarrow y_{\mathrm{cand},i_2}$
  end if
  $i_2 \leftarrow i_1 + 2$, $i_1 \leftarrow i_1 + 1$
end while
Return the lower and upper points y_{low} and y_{up}.

$$f_{Y|X}(y|x) := \frac{1}{\sqrt{2\pi v_Y(x)}} e^{-\frac{(y-\mu_Y(x))^2}{2v_Y(x)}} \quad \text{for } y \in \mathbb{R}$$

as a *symmetric* conditional probability density function of Y given X, where the marginal probability density of X is a uniform distribution $f_X(x) := \frac{1}{10}$ for $x \in (0, 10)$ and we define $\mu_Y(x) := 5 - 2\sin\frac{\pi x}{5}$ and $\sqrt{v_Y(x)} := 2 - \frac{x}{10}$ for $x \in (0, 10)$.

Distribution 2 (Normal Y with normal X): The conditional probability density function is the same as in Distribution 1, but the marginal probability density is a normal distribution

$$f_X(x) := \frac{1}{\sqrt{2\pi v_X}} e^{-\frac{(x-\mu_X)^2}{2v_X}} \quad \text{for } x \in \mathbb{R},$$

where we define $\mu_X := 5$ and $\sqrt{v_X} := \sqrt{6}$.

Distribution 3 (Lognormal Y with uniform X): In the third distribution, we use a lognormal distribution

$$f_{Y|X}(y|x) := \frac{1}{\sqrt{2\pi\hat{v}_Y(x)}y} e^{-\frac{(\log y - \hat{\mu}_Y(x))^2}{2\hat{v}_Y(x)}} \quad \text{for } y \in (0, \infty)$$

as an *asymmetric* conditional probability density function of Y given X, where the marginal probability density of X is the same uniform distribution as in Distribution 1 and we define $\hat{\mu}_Y(x) := 1 - \sin\frac{\pi x}{5}$ and $\sqrt{\hat{v}_Y(x)} := 1 - \frac{x}{20}$ for $x \in (0, 10)$.

Distribution 4 (Lognormal Y with normal X): The conditional probability density function is the same as in Distribution 3, while the marginal probability density is the same normal distribution as in Distribution 2.

Distribution 5 (Another lognormal Y with uniform X): In the fifth distribution, we also use a lognormal distribution as the conditional probability density function but its parameters $\hat{\mu}_Y(x) := \sqrt{\frac{x}{10}}$ and $\sqrt{\hat{v}_Y(x)} := 1$ for $x \in (0, 10)$ are different from Distributions 3 and 4, while the marginal probability density is the same uniform distribution as in Distribution 1.

Distribution 6 (Another lognormal Y with lognormal X): The conditional probability density function is the same as in Distribution 5, but the marginal probability density is a lognormal distribution

$$f_X(x) := \frac{1}{\sqrt{2\pi\hat{v}_X}x} e^{-\frac{(\log x - \hat{\mu}_X)^2}{2\hat{v}_X}} \quad \text{for } x \in (0, \infty),$$

where we define $\hat{\mu}_X := 1$ and $\sqrt{\hat{v}_X} := 0.5$.

Figures 1(a), 2(a), and 3(a) show the true lower and upper curves $m^{\text{low}}_{0.5,Y}(x)$ and $m^{\text{up}}_{0.5,Y}(x)$ of $\text{MI}_{0.5,Y}(x)$ for Distributions 1 & 2, 3 & 4, and 5 & 6, respectively, by black lines. The middle curve $m^{\text{mid}}_{0.5,Y}(x) := (m^{\text{low}}_{0.5,Y}(x) + m^{\text{up}}_{0.5,Y}(x))/2$ of $\text{MI}_{0.5,Y}(x)$ is also shown by a black dashed line. For each distribution, we randomly generate $N = 1000$ bivariate observations $\{(x_n, y_n)\}_{n=1}^{1000}$ s.t. $x_n \in (0, 10)$. First, each x_n is generated from the marginal probability density $f_X(x)$. Note that if $x_n \le 0$ or $x_n \ge 10$ is generated, we reject it and generate x_n again. Second, each y_n is generated from the conditional probability density $f_{Y|X}(y|x)$. Then, each method estimates $\text{MI}_{0.5,Y}(x)$ from $\{(x_n, y_n)\}_{n=1}^{1000}$. We compare the estimated results in terms of both visual features and average accuracy.

4.2.2 Parameter Settings

For the proposed method in (16), we set knots of spline functions, i.e., bins in the first step, to $\xi_0 = 0$ and $\xi_b = 10$ with the constant bin width $h_j = \frac{10}{b}$ $(j = 1, 2, \ldots, b)$, where we prepare three different values $b = 5$, $b = 10$, and $b = 20$ as the number of the bins. We set the smoothness to $\rho = 2$ that is sufficient unless there is a special reason, e.g., the third or higher derivatives of estimated curves are used in the analysis. On the degree d, we found a difference between the optimal solution for $d = 3$ and that for $d \ge 4$ when $b = 5$, but the difference almost disappeared as b increased to 10 and 20 in our preliminary experiments. Hence, we set $d = 3$ to keep the computational complexity as low as possible. We set the smoothing parameter to $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ and the lower bound gap to $\delta = 0$, and use the lower option $y_{\text{inf}} = 0$ for datasets generated from Distributions 3, 4, 5 and 6. We use two types of weights $w_n > 0$. One is setting $w_n = 1$ for all n, which is equivalent to not using weights. The other is

$$w_n = \hat{f}_X^{(\text{KDE})}(x_n) := \frac{1}{\sigma N} \sum_{l=1}^{N} \mathcal{K}\left(\frac{x_n - x_l}{\sigma}\right) \tag{35}$$

as in [15], where $\mathcal{K}(t)$ is the Gaussian kernel and its bandwidth σ is automatically adjusted from $\{x_n\}_{n=1}^{1000}$ by using the program of the kernel density estimation in MATLAB [68].

For the least squares method in (14), we use the same values $b = 5$, $b = 10$, and $b = 20$ as the number of the bins, and obtain the optimal natural cubic spline functions by the program in MATLAB [69] with $\lambda \in \{0.01, 0.1, 1, 10, 100\}$. For the two KDE methods in (31) and (33), we use the Gaussian kernel in (6) as $\mathcal{K}(t)$. In (31), we set the bandwidth to $\sigma \in \{0.1, 0.2, \ldots, 2.9, 3\}$. In (33), we set the bandwidths to $\sigma_X \in \{0.1, 0.2, \ldots, 2.9, 3\}$ and $\sigma_Y \in \{0.05, 0.1, \ldots, 0.95, 1\}$ with candidates $y_{\text{cand},i} = 0.05i$ $(i = 0, 1, \ldots, 400)$, and use the error function program in MATLAB to compute $\text{erf}(y)$.

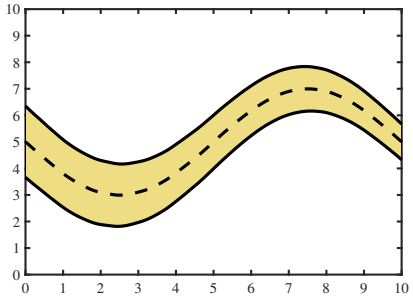
4.2.3 Comparison of Visual Features

At first, we compare the estimated results *visually* because it is easier to understand the effectiveness and the characteristics of the proposed method than to compare the estimation accuracy numerically. Figures 1(b)–1(d), 2(b)–2(d), and 3(b)–3(d) show typical results[†] of the conventional methods and the proposed method for Distributions 1, 3, and 6, respectively, where bivariate observations $\{(x_n, y_n)\}_{n=1}^{1000}$ are shown by blue points except for Fig. 3(b), and blue/red circles in Fig. 3(b) depict the lower/upper points $\hat{y}^{\langle j\rangle}_{\text{low}}$ and $\hat{y}^{\langle j\rangle}_{\text{up}}$ of the local 50% empirical modal interval of Y in each bin.

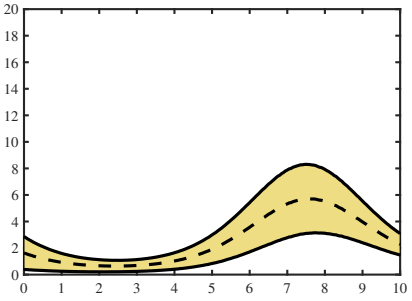
Figures 1(b), 1(c), and 1(d) show examples of results for Distribution 1 by the KDE method in (31) with $\sigma = 0.5$, the KDE method in (33) with $\sigma_X = \sigma_Y = 0.5$, and the proposed method in (16) with $b = 20$, $\lambda = 0.1$ and $w_n = \hat{f}_X^{(\text{KDE})}(x_n)$, respectively. In each figure, the estimated interval contains many bivariate observations and has a similar trend to the true modal interval in Fig. 1(a). From Fig. 1(b), the result by the KDE method in (31) has a stepped shape with very large fluctuations, which looks very unnatural. From Fig. 1(c), the result by the KDE method in (33) has a shape with smaller fluctuations than that in Fig. 1(b), but it is slightly discontinuous due to the discretization of y_{low} and y_{up} into $y_{\text{cand},i}$ in (34). On the other hand, from Fig. 1(d), the result by the proposed method has a smooth shape similar to the true one in Fig. 1(a) and well reproduces the gradual shortening of the interval width based on the decreasing property of $\sqrt{v_Y(x)}$.

Figures 2(b), 2(c), and 2(d) show examples of results for Distribution 3 by the KDE method in (31) with $\sigma = 0.5$, the KDE method in (33) with $\sigma_X = 0.6$ and $\sigma_Y = 0.5$, and the proposed method with $b = 10$, $\lambda = 0.1$ and $w_n = \hat{f}_X^{(\text{KDE})}(x_n)$,

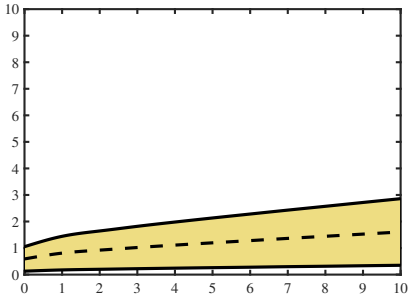
[†]The bandwidths σ, σ_X and σ_Y in the KDE methods are set to values when the best accuracy is achieved in Table 1 of Sect. 4.2.4.



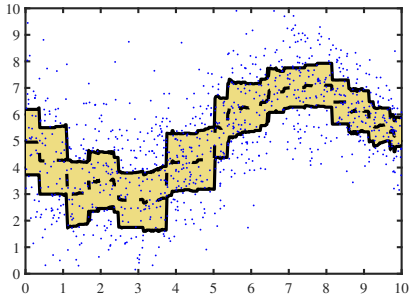
(a) True 50% conditional modal interval $\mathrm{MI}_{0.5,Y}(x)$ for Distributions 1 and 2.



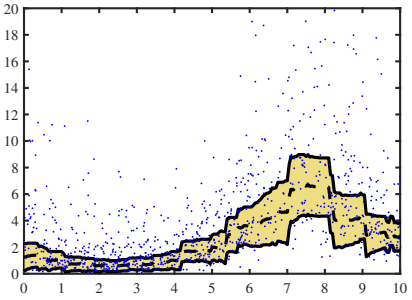
(a) True 50% conditional modal interval $\mathrm{MI}_{0.5,Y}(x)$ for Distributions 3 and 4.



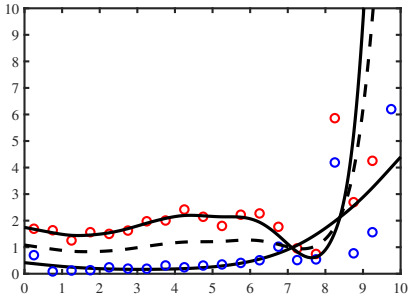
(a) True 50% conditional modal interval $\mathrm{MI}_{0.5,Y}(x)$ for Distributions 5 and 6.



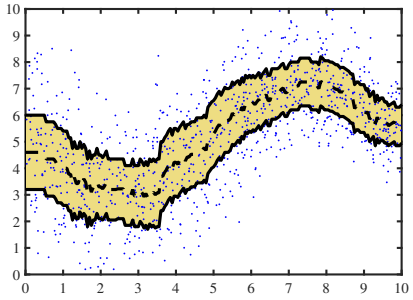
(b) Example of the estimated result by the KDE method in (31) with $\sigma = 0.5$.



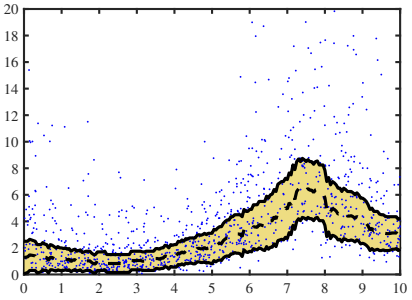
(b) Example of the estimated result by the KDE method in (31) with $\sigma = 0.5$.



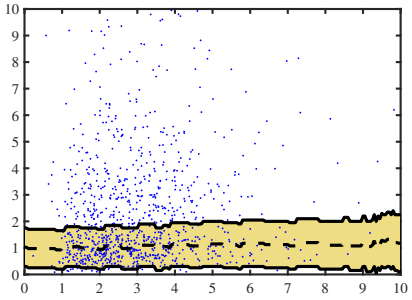
(b) Example of the estimated result by the least squares method in (14) with $b = 20$ and $\lambda = 1$.



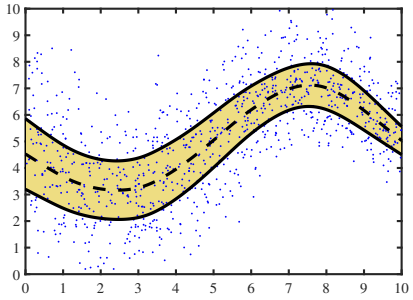
(c) Example of the estimated result by the KDE method in (33) with $\sigma_X = \sigma_Y = 0.5$.



(c) Example of the estimated result by the KDE method in (33) with $\sigma_X = 0.6$ and $\sigma_Y = 0.5$.

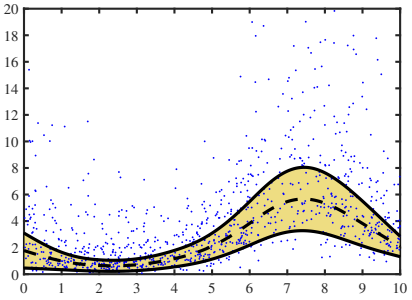


(c) Example of the estimated result by the KDE method in (33) with $\sigma_X = 2$ and $\sigma_Y = 0.1$.



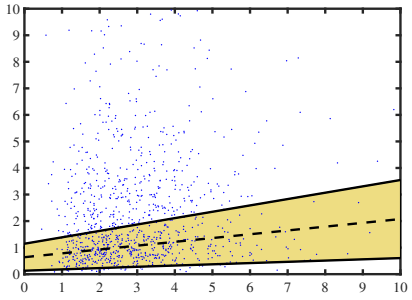
(d) Example of the estimated result by the proposed method in (16) with $b = 20$, $\lambda = 0.1$ and $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$.

Fig. 1 Example of the estimated result by each method from $\{(x_n, y_n)\}_{n=1}^{1000}$ of Distribution 1.



(d) Example of the estimated result by the proposed method in (16) with $b = 10$, $\lambda = 0.1$ and $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$.

Fig. 2 Example of the estimated result by each method from $\{(x_n, y_n)\}_{n=1}^{1000}$ of Distribution 3.



(d) Example of the estimated result by the proposed method in (16) with $b = 5$, $\lambda = 100$ and $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$.

Fig. 3 Example of the estimated result by each method from $\{(x_n, y_n)\}_{n=1}^{1000}$ of Distribution 6.

Table 1 Average RMSEs between the true modal intervals and the estimated intervals in 100 trials.

Method	Distribution 1	Distribution 2	Distribution 3	Distribution 4	Distribution 5	Distribution 6
KDE Method in (31)	0.8096 ($\sigma = 0.5$)	0.8800 ($\sigma = 0.6$)	0.8395 ($\sigma = 0.5$)	0.9604 ($\sigma = 0.5$)	0.2835 ($\sigma = 1.8$)	**0.4539** ($\sigma = 2$)
KDE Method in (33)	0.7034 ($\sigma_X = 0.5$, $\sigma_Y = 0.5$)	0.7729 ($\sigma_X = 0.5$, $\sigma_Y = 0.6$)	0.8704 ($\sigma_X = 0.6$, $\sigma_Y = 0.5$)	0.9869 ($\sigma_X = 0.5$, $\sigma_Y = 0.5$)	0.3222 ($\sigma_X = 1.5$, $\sigma_Y = 0.05$)	0.4589 ($\sigma_X = 2$, $\sigma_Y = 0.1$)
Least Squares Method in (14)	0.8095 ($b = 5$, $\lambda = 0.01$)	1.0003 ($b = 5$, $\lambda = 0.01$)	0.9319 ($b = 5$, $\lambda = 0.01$)	1.1034 ($b = 5$, $\lambda = 0.01$)	**0.2448** ($b = 5$, $\lambda = 100$)	0.9263 ($b = 5$, $\lambda = 100$)
Proposed Method with $w_n = 1$	0.6022 ($b = 20$, $\lambda = 1$)	**0.7162** ($b = 20$, $\lambda = 1$)	**0.6873** ($b = 10$, $\lambda = 1$)	**0.8279** ($b = 5$, $\lambda = 1$)	0.2591 ($b = 5$, $\lambda = 100$)	0.6827 ($b = 5$, $\lambda = 100$)
Proposed Method with $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$	**0.5980** ($b = 20$, $\lambda = 0.1$)	0.7709 ($b = 20$, $\lambda = 0.1$)	0.6924 ($b = 10$, $\lambda = 0.1$)	0.9913 ($b = 5$, $\lambda = 0.01$)	0.2499 ($b = 5$, $\lambda = 100$)	0.5063 ($b = 5$, $\lambda = 100$)

respectively. Even though the conditional probability density function $f_{Y|X}(y|x)$ is asymmetric differently from Distribution 1, each estimated result has a similar trend to the true modal interval in Fig. 2(a). From Fig. 2(b), the result by the KDE method in (31) has a unnatural stepped shape especially for $x \geq 7$. From Fig. 2(c), the result by the KDE method in (33) is closer to the modal interval in Fig. 2(a), but it is still discontinuous. On the other hand, from Fig. 2(d), the result by the proposed method has a smooth and accurate shape.

Figures 3(b), 3(c), and 3(d) show examples of results for Distribution 6 by the least squares method in (14) with $b = 20$ and $\lambda = 1$, the KDE method in (33) with $\sigma_X = 2$ and $\sigma_Y = 0.1$, and the proposed method in (16) with $b = 5$, $\lambda = 100$ and $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$, respectively. Since the marginal probability density function $f_X(x)$ is a lognormal distribution, the number of observations s.t. $x_n \geq 6$ is extremely small differently from Distributions 1 and 3. As a result, in Fig. 3(b), both lower and upper points $\hat{y}_{\mathrm{low}}^{\langle j \rangle}$ and $\hat{y}_{\mathrm{up}}^{\langle j \rangle}$ in the 13th to 20th bins are far from the true modal interval, and the accuracy of the least squares method is poor for $x \geq 6$ because the lower and upper curves $\underline{s}^*(x)$ and $\overline{s}^*(x)$ cross each other. This shows that the least squares method in (14) fails to construct the interval in some cases. From Figs. 3(c) and 3(d), the result by the KDE method in (33) does not express the strictly increasing properties of the lower and upper curves $m_{0.5,Y}^{\mathrm{low}}(x)$ and $m_{0.5,Y}^{\mathrm{up}}(x)$ and the width $m_{0.5,Y}^{\mathrm{up}}(x) - m_{0.5,Y}^{\mathrm{low}}(x)$ of the true modal interval $\mathrm{MI}_{0.5,Y}(x)$ in Fig. 3(a), while the proposed method reconstructs a smooth modal interval possessing these three strictly increasing properties. From Figs. 1, 2, and 3, we see that, for various distributions, the proposed method can reproduce a smooth modal interval and express the characteristics of the spread of the distribution more accurately compared to the conventional methods.

4.2.4 Comparison of Estimation Accuracy

Next, we compare the average accuracy of each method in 100 datasets for each distribution. We evaluate the accuracy of the proposed estimate $\underline{s}^*(x)$ for the lower curve $m_{0.5,Y}^{\mathrm{low}}(x)$ by the *root mean square error (RMSE)* in the 100 trials and 51 sampling points $x_i = 0.2i$ ($i = 0, 1, \ldots, 50$) as follows:

$$\mathrm{RMSE}_{\mathrm{low}} := \sqrt{\frac{1}{5100} \sum_{\mathrm{trial}=1}^{100} \sum_{i=0}^{50} |m_{0.5,Y}^{\mathrm{low}}(0.2i) - \underline{s}_{\mathrm{trial}}^*(0.2i)|^2}.$$

Similarly, $\mathrm{RMSE}_{\mathrm{up}}$ of $\overline{s}^*(x)$ for $m_{0.5,Y}^{\mathrm{up}}(x)$ is computed, and the total RMSE is given by $\mathrm{RMSE} := \mathrm{RMSE}_{\mathrm{low}} + \mathrm{RMSE}_{\mathrm{up}}$. We also compute the total RMSEs for the KDE methods in (31) and (33), and the least squares method in (14). Table 1 shows the average RMSE in 100 trials for each method with parameters when the best accuracy was achieved, where the proposed method is divided into two cases, one with constant weights and the other with the non-constant weights in (35).

From Table 1, for Distributions 1, 3, and 5, the proposed method achieved higher estimation accuracy than the two KDE methods in (31) and (33), and there was not much difference between the two types of weights in the proposed method since the weights in (35) were nearly constant due to the uniform marginal distribution of X. In addition, each bin $\mathcal{B}_j$ contained a similar number of observations, and thus the least squares method in (14) could estimate the modal interval with reasonable accuracy, especially with the highest accuracy for Distribution 5, without the lower and upper curves crossing each other, differently from Fig. 3(b).

For Distributions 2 and 4, since the number of observations close to both ends of the x direction was reduced due to the normal marginal distribution of X, the estimation accuracy of each method was lower than that for Distributions 1 and 3. In particular, the least squares method in (14) had the lowest estimation accuracy. Both KDE methods in (31) and (33) could relatively suppress the accuracy degradation from Distribution 1 to Distribution 2, which is probably because the kernels $\mathcal{K}(t)$ in the x and y directions and the true conditional and marginal probability densities are all Gaussian. On the other hand, even though the proposed method with the constant weights does not use shape information such as a normal or lognormal distribution, it still achieved higher estimation accuracy. The use of the weights in (35) decreased the accuracy, which implies that, rather than the weights have some negative impact, there is less need to use weights since the smoothness parameter λ has a greater impact than w_n.

For Distribution 6, the number of observations for $x \geq 6$ was very small due to the lognormal marginal distribution of

Table 2 Average RMSEs of the least squares method and the proposed method for each b and λ.

Method	b	λ	Distribution 1	Distribution 2	Distribution 3	Distribution 4	Distribution 5	Distribution 6
Least Squares Method in (14)	5	0.01	**0.8095**	**1.0003**	**0.9319**	**1.1034**	0.3625	1.2282
		0.1	0.8219	1.0302	0.9639	1.1827	0.3426	1.2021
		1	1.1886	1.3706	1.4866	1.7028	0.2922	1.1092
		10	1.7339	1.8277	2.1332	2.2420	0.2543	0.9868
		100	1.8884	1.9562	2.3174	2.3875	**0.2448**	**0.9263**
	10	0.01	2.8741	2.9159	2.2667	2.2925	0.9356	1.9102
		0.1	2.8315	2.8690	2.1945	2.2158	0.8964	1.7543
		1	2.8319	2.8623	2.2174	2.2529	0.8634	1.5949
		10	3.0609	3.0811	2.5571	2.5965	0.8472	1.4809
		100	3.2347	3.2489	2.7789	2.8240	0.8396	1.3979
	20	0.01	3.0579	3.0966	2.3802	2.4266	1.0515	3.0172
		0.1	3.0125	3.0476	2.2878	2.3168	0.9951	2.5975
		1	2.9832	3.0146	2.2424	2.2670	0.9608	2.3089
		10	3.0942	3.1194	2.4312	2.4699	0.9444	2.1130
		100	3.3139	3.3257	2.7517	2.8110	0.9357	1.9565
Proposed Method with $w_n = 1$	5	0.01	0.7008	0.8743	0.7453	0.9114	0.4278	1.5139
		0.1	0.6711	0.8177	0.6932	0.8491	0.4015	1.3507
		1	0.6359	0.7933	0.7076	**0.8279**	0.3601	1.1722
		10	0.7711	1.0874	1.1878	1.6589	0.3034	0.9366
		100	1.6027	2.0099	2.1428	2.4998	**0.2591**	**0.6827**
	10	0.01	0.8202	1.3248	0.9073	1.1206	0.5367	1.9843
		0.1	0.7683	0.9026	0.7937	0.9926	0.4841	1.5939
		1	0.6426	0.7719	**0.6873**	0.9076	0.4098	1.3379
		10	0.7262	1.0218	1.1133	1.7103	0.3375	1.0405
		100	1.5700	1.9705	2.0811	2.4959	0.3012	0.7651
	20	0.01	0.9258	1.0316	1.1012	1.3804	0.7185	2.7599
		0.1	0.7490	0.8549	0.9044	1.1629	0.6101	2.2405
		1	**0.6022**	**0.7162**	0.7510	1.0639	0.5216	1.8002
		10	0.6898	0.9649	1.1028	1.8712	0.4588	1.4732
		100	1.5430	1.9264	2.0272	2.5617	0.4290	1.0762
Proposed Method with $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$	5	0.01	0.6708	0.8937	0.6965	**0.9913**	0.3991	1.1359
		0.1	0.6466	0.8898	0.7162	1.1237	0.3613	0.9291
		1	0.8454	1.5960	1.2652	2.5365	0.3071	0.7041
		10	1.7000	2.6361	2.1933	3.0928	0.2641	0.5074
		100	1.9630	2.8681	2.3961	3.2037	**0.2499**	**0.5063**
	10	0.01	0.7680	0.8759	0.7937	0.9995	0.4822	1.3668
		0.1	0.6350	0.8305	**0.6924**	1.2501	0.4076	1.0566
		1	0.8045	1.5063	1.2067	2.6068	0.3421	0.7996
		10	1.6675	2.6280	2.1473	3.1361	0.3064	0.5686
		100	1.9392	2.8719	2.3627	3.2372	0.3011	0.5410
	20	0.01	0.7411	0.8358	0.8989	1.1690	0.6055	1.8140
		0.1	**0.5980**	**0.7709**	0.7524	1.5490	0.5176	1.4403
		1	0.7647	1.4552	1.2189	2.8786	0.4599	1.0904
		10	1.6429	2.5632	2.1214	3.2071	0.4354	0.7400
		100	2.0152	2.8163	2.3310	3.2759	0.4287	0.6588

X. As a result, the estimation accuracy of each method decreased relatively largely from Distribution 5. In this case, the two KDE methods in (31) and (33) achieved higher estimation accuracy than the proposed method, though the KDE methods cannot fully reproduce the increasing properties of the true interval as shown in Fig. 3(c). The accuracy of the proposed method with the weights in (35) was close to that of the KDE methods, while the use of the constant weights decreased the accuracy. This implies that when the marginal probability density of X is a long-tailed distribution, it is better to use weights and set them to small values in regions where the number of observations is extremely small.

We also verify the influence of parameter settings on the proposed method. Table 2 summarizes the average RMSEs of the least squares method in (14) and the proposed method with the constant or non-constant weights for each b and λ. From Table 2, on the least squares method, the estimation accuracy was higher when $b = 5$ than when $b = 10$ and $b = 20$ for all the distributions. This is because the accuracy of the estimated lower and upper points in (13) decreased as the number b of the bins increased. When b is small, the least squares method can only roughly capture the modal interval, and thus it cannot achieve high accuracy unless the shape of the true modal interval is simple as in Fig. 3(a). On the other hand, on the proposed method, there was at least one distribution which was best for each b. When the value of λ was fixed, changing the value of b had a relatively small impact on the accuracy. Thus, compared to the least squares method,

Table 3 Average computation times in 100 trials to estimate the modal interval at 51 sampling points.

Distribution	Proposed Method						KDE method in (31)	KDE method in (33)
	Quadratic Programming in (28)			ADMM in (29)				
	$b = 5$	$b = 10$	$b = 20$	$b = 5$	$b = 10$	$b = 20$		
Distribution 1	0.4 sec	0.5 sec	0.8 sec	0.2 sec	0.3 sec	0.5 sec	1.4 sec	4.6 sec
Distribution 2	0.4 sec	0.5 sec	0.8 sec	0.2 sec	0.3 sec	0.5 sec	1.4 sec	4.7 sec
Distribution 3	0.4 sec	0.5 sec	0.8 sec	0.2 sec	0.3 sec	0.5 sec	1.4 sec	4.5 sec
Distribution 4	0.4 sec	0.5 sec	0.8 sec	0.2 sec	0.3 sec	0.5 sec	1.4 sec	4.5 sec
Distribution 5	0.4 sec	0.5 sec	0.8 sec	0.2 sec	0.3 sec	0.5 sec	1.4 sec	4.5 sec
Distribution 6	0.4 sec	0.5 sec	0.9 sec	0.2 sec	0.3 sec	0.7 sec	1.4 sec	4.5 sec

the proposed method in (16) is relatively robust for the value of b. In the proposed method, if b is large, the flexibility of spline functions is increased, while overfitting tends to occur in regions where the number of observations is small. Since the modal interval for Distributions 1 and 2 in Fig. 1(a) and that for Distributions 3 and 4 in Fig. 2(a) are smooth but vary to some extent, high accuracy tended to be achieved when $b = 10$ or $b = 20$ and the smoothing parameter λ was not too large. Meanwhile, both lower and upper curves of the modal interval for Distributions 5 and 6 in Fig. 3(a) are similar to straight lines in shape. Therefore, the highest accuracy was achieved when $b = 5$ with the largest smoothing parameter $\lambda = 100$. In particular, for Distribution 6, using the weights in (35) implicitly increased the value of λ in regions where the number of observations is small, and higher estimation accuracy was achieved than using the constant weights. To summarize the above discussion, for various distributions, the proposed method can estimate the modal interval as well as or more accurately than the least squares method and the KDE methods, with the robustness for the number of bins.

Differently from the least squares method, the proposed method could achieve good estimation accuracy for all the cases of $b = 5, 10, 20$ by tuning the smoothing parameter λ. However, automatic tuning of λ from bivariate observations is a difficult problem in practical applications, as with tuning of the bandwidths in the KDE methods. Since the proposed method can visualize the modal interval quickly as explained in the next section, we can attempt several parameter values and then select a good result by visual judgment as follows. To keep the computational complexity as low as possible, we use $\rho = 2$ and $d = 3$. We set δ and the options $y_{\inf}$ and $y_{\sup}$ according to requirements. We set w_n as in (35) if the distribution of x_n is long-tailed, and use $w_n = 1$ otherwise. We set b so that the number of observations in each bin $\mathcal{B}_j$ is not too small. Then, we only need to adjust the smoothing parameter λ while looking at estimated modal intervals by the proposed method. If overfitting occurs, i.e., lower and upper curves rapidly fluctuate, then we increase λ. If underfitting occurs, i.e., both curves are almost linear and the estimated interval contains only a few observations, then we decrease λ. After several attempts, we select one visually good result.

4.2.5 Comparison of Computation Time

Finally, we compare the computation times of the proposed method with the two KDE methods on MATLAB R2022a 64-bit (macOS Monterey, SoC Apple M1 Pro, and memory 16 GB). We implement the proposed method by both quadratic programming in (28) using a package [66] and ADMM in (29) using Algorithm 1 with $\gamma = 1$ and Iter = 1000. We implement the two KDE methods as in Algorithm 2 and Algorithm 3. Table 3 shows the average computation times in 100 trials of each distribution to estimate the modal interval $\mathrm{MI}_{0.5,Y}(x)$ at 51 sampling points $x_i = 0.2i$ ($i = 0, 1, \ldots, 50$).

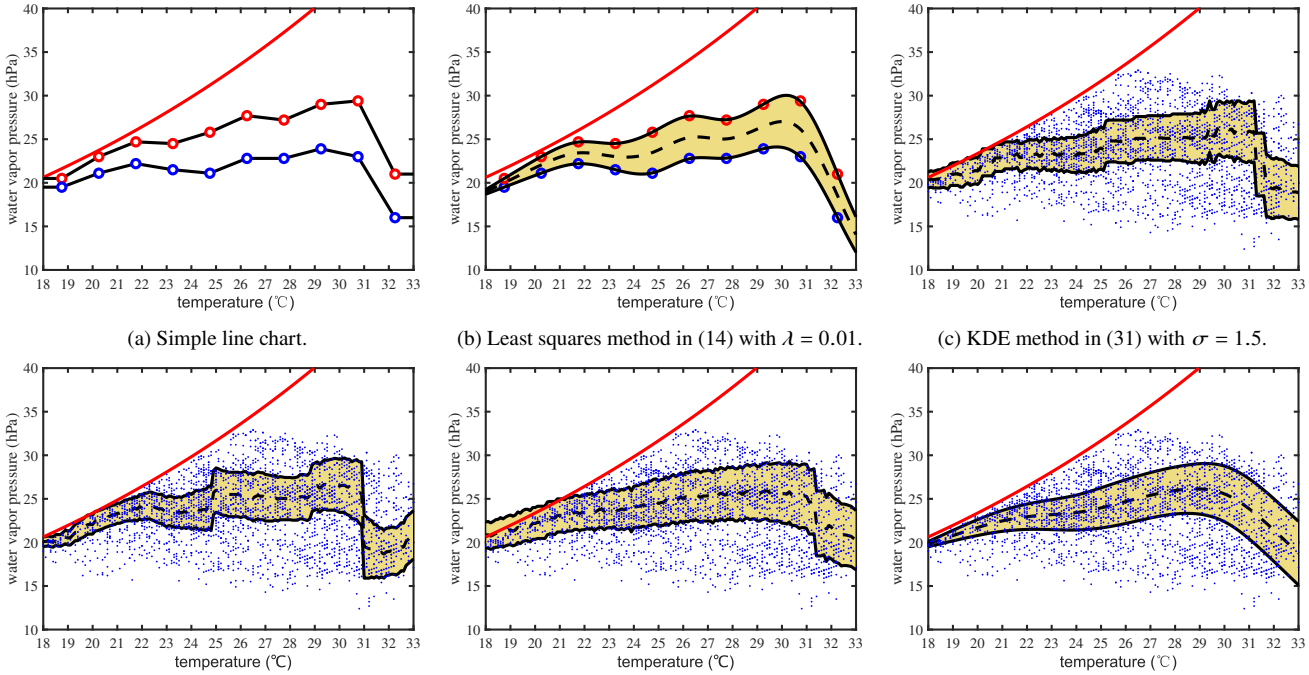
From Table 3, we see that whether using the quadratic programming or ADMM, the proposed method estimated the modal interval quickly, i.e., in less than 1 second per dataset. The computation time became longer when b increased, and ADMM was a little faster than the quadratic programming using the sophisticated package [66]. Note that both results were almost the same due to the convexity of the problem in (16), and ADMM's computation time can be further reduced if a recently proposed overrelaxation technique [49] is used.

On the other hand, both KDE methods required longer computation time than the proposed method. This is because the KDE methods have to find the modal interval by (31) or (33) as many times as the number of the sampling points x_i, while the proposed method solves the problem in (16) only once independently of the number of the sampling points x_i. The KDE method in (33) took more than three times longer to estimate the modal interval than the KDE method in (31) since the cumulative distribution in line 3 of Algorithm 3 takes longer to compute than that in line 6 of Algorithm 2.

4.3 Experiments on Real-World Data

We apply the above methods to estimate the 50% conditional modal interval $\mathrm{MI}_{0.5,Y}(x)$ of real-world meteorological data provided by China Meteorological Data Service Centre [70], where we analyze hourly data observed at 18 surface meteorological stations in Beijing, China, from August 1 to 7, 2021, and select temperature X [°C] and water vapor pressure Y [hPa] as a pair of random variable (X, Y). The number of observations is $N = 2706$, i.e., we have $\{(x_n, y_n)\}_{n=1}^{2706}$.

Since we cannot know the true modal interval for real-world data, we investigate the effectiveness of each method from features of the estimated results in Fig. 4. We set knots of spline functions to $\xi_0 = 18$ and $\xi_{10} = 33$ with the constant bin width $h_j = 1.5$ ($j = 1, 2, \ldots, 10 =: b$), and the smoothness and degree to $\rho = 2$ and $d = 3$ for the proposed method in (16). Figures 4(a), 4(b) ,4(c), 4(d), 4(e), and 4(f) show the results by the simple line chart, the least squares method in



(a) Simple line chart. (b) Least squares method in (14) with $\lambda = 0.01$. (c) KDE method in (31) with $\sigma = 1.5$.

(d) KDE method in (33) with $\sigma_X = \sigma_Y = 0.5$. (e) KDE method in (33) with $\sigma_X = \sigma_Y = 1.5$. (f) Proposed method in (16) with $\lambda = 1$.

Fig. 4 Estimated modal intervals for real-world temperature and water vapor pressure data in [70].

(14) with $\lambda = 0.01$, the KDE method in (31) with $\sigma = 1.5$, the KDE method in (33) with $\sigma_X = \sigma_Y = 0.5$, the KDE method in (33) with $\sigma_X = \sigma_Y = 1.5$, and the proposed method in (16) with $\lambda = 1$ and $w_n = \hat{f}_X^{(\mathrm{KDE})}(x_n)$, respectively. In these figures, the red line denotes the *saturated vapor pressure* [71]

$$y_{\sup}(x) := 6.112\, e^{\frac{17.67x}{x+243.5}}, \tag{36}$$

which means the maximum possible value of the water vapor pressure Y given temperature $X = x$. In Figs. 4(a) and 4(b), blue/red circles depict the lower/upper points $\hat{y}_{\mathrm{low}}^{\langle j \rangle}$ and $\hat{y}_{\mathrm{up}}^{\langle j \rangle}$ of the local 50% empirical modal interval of Y in each bin, and in Figs. 4(c)–4(f), blue points depict the bivariate observation $\{(x_n, y_n)\}_{n=1}^{2706}$. All the estimated 50% modal intervals have similar trends and contain many bivariate observations.

From Figs. 4(a) and 4(b), the results by the simple line chart and the least squares method in (14) with small λ can capture the trend of the modal interval, but the fluctuations of the lower and upper curves obscure the characteristics of the modal interval. From Fig. 4(c), the result by the KDE method in (31) not only has a unnatural stepped shape especially for $x \geq 31$, but also overshoots the saturated vapor pressure in (36) for $x \leq 21$, which means that the estimation is inaccurate. From Figs. 4(d) and 4(e), the result by the KDE method in (33) with relatively small σ_X and σ_Y also has a unnatural stepped shape especially around $x = 25$ and $x = 31$, while that with relatively large σ_X and σ_Y has a shape with smaller fluctuations but clearly overshoots the saturated vapor pressure for $x \leq 21$. On the other hand, from Fig. 4(f), the result by the proposed method with the appropriate parameters has a natural smooth shape without overshooting the saturated vapor pressure. Moreover, we can see that the result by the proposed method noticeably expresses the increasing properties of the lower and upper curves of the the modal interval and its width for $x \leq 29$ and the decreasing property of the lower and upper curves for $x \geq 30$. Therefore, we find that the proposed method also performs well for real-world data and provides good visualization of the shape features of the region where bivariate observations appear most frequently.

5. Conclusion

In this paper, for analysis and visualization of bivariate data, we proposed a two-step modal interval regression method using flexible spline regression models. In the local empirical modal interval detection step, after dividing bivariate observations into bins for one random variable, we expressed the empirical modal interval for the other random variable as the lower and upper quantiles in each bin. In the simultaneous spline quantile regression step, both lower and upper curves of the conditional modal interval were simultaneously estimated as spline functions having the non-crossing property. By using the asymmetric absolute errors based on the quantile regression, this step can be formulated as a convex optimization problem on the concatenated coefficient vector of the lower and upper spline functions and solved very quickly with the quadratic programming or ADMM. Extensive numerical experiments for synthetic data generated from various distributions demonstrated the effectiveness of the proposed method visually and numerically, in comparison with the straightforward least squares method and the two kernel density estimation methods. The numerical experiments also

showed that the proposed method is robust for the number of bins and requires less computation time than the two kernel density estimation methods. Furthermore, we also applied the proposed method to real-world meteorological data and confirmed that the proposed method can construct a smooth modal interval that noticeably expresses the characteristics of the region where observations appear most frequently.

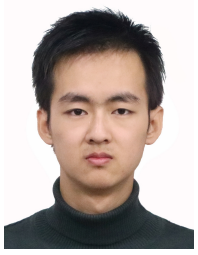
Acknowledgments

This work was supported by JSPS KAKENHI (19K20361). The authors would like to thank the anonymous reviewers for their many constructive comments which were very helpful to improve the contents and the readability of this paper. The authors are also grateful to Dr. Atsunori Kashiwagi, Dr. Yuji Tezuka, and Dr. Osamu Sekine in Omi Medical Center for their valuable advice in fruitful discussions on data analysis.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, NY, 2009.
[2] G.A.F. Seber and A.J. Lee, Linear Regression Analysis, 2nd ed., Wiley, New York, NY, 2003.
[3] J. Shao, Mathematical Statistics, 2nd ed., Springer, New York, NY, 2003.
[4] B. Abraham and J. Ledolter, Introduction to Regression Modeling, Duxbury, Belmont, CA, 2005.
[5] J.A. Rice, Mathematical Statistics and Data Analysis, 3rd ed., Duxbury, Belmont, CA, 2006.
[6] R.J. Freund, W.J. Wilson, P. Sa, Regression Analysis, 2nd ed., Academic Press, London, UK, 2006.
[7] D.C. Montgomery, E.A. Peck, and G.G. Vining, Introduction to Linear Regression Analysis, 5th ed., Wiley, New York, NY, 2013.
[8] G. Bassett Jr. and R. Koenker, "Asymptotic theory of least absolute error regression," J. Am. Stat. Assoc., vol.73, no.363, pp.618–622, Sept. 1978.
[9] P. Bloomfield and W.L. Steiger, Least Absolute Deviations: Theory, Applications, and Algorithms, Birkhäuser, Boston, MA, 1983.
[10] P.J. Huber and E.M. Ronchetti, Robust Statistics, 2nd ed., Wiley, New York, NY, 2009.
[11] R.R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, 4th ed., Academic Press, London, UK, 2016.
[12] R.A. Maronna, R.D. Martin, V.J. Yohai, and M. Salibián-Barrera, Robust Statistics: Theory and Methods (with R), 2nd ed., Wiley, New York, NY, 2019.
[13] R. Koenker and G. Bassett Jr., "Regression quantiles," Econometrica, vol.46, no.1, pp.33–50, Jan. 1978.
[14] R. Koenker and K.F. Hallock, "Quantile regression," J. Econ. Perspect., vol.15, no.4, pp.143–156, 2001.
[15] R. Koenker, Quantile regression, Cambridge University Press, New York, NY, 2005.
[16] R. Koenker, V. Chernozhukov, X. He, and L. Peng, Handbook of Quantile Regression, Chapman & Hall, New York, NY, 2017.
[17] M.J. Lee, "Mode regression," J. Econom., vol.42, no.3, pp.337–349, Nov. 1989.
[18] W. Yao, B.G. Lindsay, and R. Li, "Local modal regression," J. Nonparametr. Stat., vol.24, no.3, pp.647–663, Sept. 2012.
[19] Y. Feng, J. Fan, and J.A.K. Suykens, "A statistical learning approach to modal regression," J. Mach. Learn. Res., vol.21, no.2, pp.1–35, 2020.
[20] H. Sasaki, T. Sakai, and T. Kanamori, "Robust modal regression with direct gradient approximation of modal regression risk," Proc. Conf. Uncert. Artif. Intell. (UAI), Online, pp.380–389, Aug. 2020.
[21] H. Chen, Y. Wang, F. Zheng, C. Deng, and H. Huang, "Sparse modal additive model," IEEE Trans. Neural Netw. Learn. Syst., vol.32, no.6, pp.2373–2387, June 2021.
[22] A. Eldeeb, S. Desoky, M. Ahmed, "A new robust algorithm for penalized regression splines based on mode-estimation," Int. J. Nonlinear Anal. Appl., vol.12, no.1, pp.1037–1055, 2021.
[23] C. de Boor, "Best approximation properties of spline functions of odd degree," J. Math. Mech., vol.12, no.5, pp.747–749, 1963.
[24] S. Wold, "Spline functions in data analysis," Technomet., vol.16, no.1, pp.87–111, Feb. 1974.
[25] B.W. Silverman, "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," J. Royal Stat. Soc. B, vol.47, no.1, pp.1–52, 1985.
[26] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.
[27] M. Unser, "Splines: A perfect fit for signal and image processing," IEEE Signal Process. Magazine, vol.16, no.6, pp.22–38, Nov. 1999.
[28] J.O. Ramsay and B.W. Silverman, Functional Data Analysis, 2nd ed, Springer, New York, NY, 2005.
[29] C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, Cambridge, MA, 2005.
[30] J. Quiñonero-Candela and C.E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," J. Mach. Learn. Res., vol.6, no.65, pp.1939–1959, Dec. 2005.
[31] J.Q. Shi and T. Choi, Gaussian Process Regression Analysis for Functional Data, Chapman & Hall, New York, NY, 2011.
[32] R.B. Gramacy, Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences, Chapman & Hall, New York, NY, 2021.
[33] R.J. Hyndman, "Computing and graphing highest density regions," Am. Stat., vol.50, no.2, pp.120–126, May 1996.
[34] J. Fan and Q. Yao, Nonlinear Time Series: Nonparametric and Parametric Methods Annals of the Institute of Statistical Mathematics, Springer, New York: NY, 2003.
[35] J.H.J. Einmahl, M. Gantner, and G. Sawitzki, "The shorth plot," J. Comput. Graph. Stat., vol.19, no.1, pp.62–73, 2010.
[36] J.G. de Gooijer and Ali Gannoun, "Nonparametric conditional predictive regions for time series," Comput. Stat. Data Anal., vol.33, no.3, pp.259–275, May 2000.
[37] W. Polonik and Q. Yao, "Conditional minimum volume predictive regions for stochastic processes," J. Am. Stat. Assoc. vol.95, no.450, pp.509–519, Jun. 2000.
[38] J. Demongeot, A. Laksaci, M. Rachdi, and S. Rahmani, "On the local linear modelization of the conditional distribution for functional data," Sankhyā: Indian J. Stat., vol.76, no.2, pp.328–355, Mar. 2014.
[39] M. Rachdi, A. Laksaci, I.M. Almanjahie, and Z. Chikr-Elmezouar, "FDA: Theoretical and practical efficiency of the local linear estimation based on the kNN smoothing of the conditional distribution when there are missing data," vol.90, no.8, 2020.
[40] L. Zhu, J. Lu, and Y. Chen, "HDI-Forest: Highest density interval regression forest," Proc. Int. Joint Conf. Artif. Intell. (IJCAI), Macao, China, pp.4468–4474, Aug. 2019.
[41] M.H. Roy and D. Larocque, "Prediction intervals with random forests," Stat. Methods Med. Res., vol.29, no.1, pp.205–229, 2020.
[42] D. Kitahara, K. Leng, Y. Tezuka, and A. Hirabayashi, "Simultaneous spline quantile regression under shape constraints," Proc. Eur. Signal Process. Conf. (EUSIPCO), Amsterdam, The Netherlands, pp.2423–2427, Oct. 2020.
[43] R.J. Vanderbei and T.J. Carpenter, "Symmetric indefinite systems for interior point methods," Math. Program., vol.58, no.1–3, pp.1–32, Jan. 1993.
[44] A. Altman and J. Gondzio, "Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization," Optim. Methods Softw., vol.11, no.1–4, pp.275–302, 1999.
[45] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite elements approximations,"

Comput. Math. Appl., vol.2, no.1, pp.17–40, 1976.
[46] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn., vol.3, no.1, pp.1–122, Jan. 2011.
[47] R. Glowinski, "On alternating direction methods of multipliers: A historical perspective," Modeling, Simulation and Optimization for Science and Technology, eds. W. Fitzgibbon, Y.A. Kuznetsov, P. Neittaanmäki, and O. Pironneau, pp.59–82, Springer Netherlands, Dordrecht, The Netherlands, 2014.
[48] J. Eckstein and W. Yao, "Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives," Pacific J. Optim., vol.11, no.4, pp.619–644, Oct. 2015.
[49] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, "Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists," SIAM Review, to appear.
[50] G.R. Shorack, Probability for Statistics, 2nd ed., Springer, New York, NY, 2017.
[51] R. J. Casady and J. D. Cryer, "Monotone percentile regression," Ann. Stat., vol.4, no.3, pp.532–541, May 1976.
[52] D. Griffiths and M. Willcox, "Percentile regression: A parametric approach," J. Am. Stat. Assoc., vol.73, no.363, pp.496–498, Sep. 1978.
[53] T. J. Cole, "Fitting smoothed centile curves to reference data," J. Royal Stat. Soc. A, vol.151, no.3, pp.385–418, 1988.
[54] Y.C. Chen, C.R. Genovese, R.J. Tibshirani, and L. Wasserman, "Nonparametric modal regression," Ann. Stat., vol.44, no.2, pp.489–514, Apr. 2016.
[55] Y.C. Chen, "Modal regression using kernel density estimation: A review," WIREs Comput. Stat., vol.10, no.4, 14 pages, July/Aug. 2018.
[56] J.E. Chacón, "The modal age of statistics," Int. Stat. Review, vol.88, no.1, pp.122–141, Apr. 2020.
[57] P. Čížek and S. Sadıkoğlu, "Robust nonparametric regression: A review," WIREs Comput. Stat., vol.12, no.3, 16 pages, May/June 2020.
[58] H. Ota, K. Kato, and S. Hara, "Quantile regression approach to conditional mode estimation," Electron. J. Statist., vol.13, no.2, pp.3120–3160, Sept. 2019.
[59] W. Yao and L. Li, "A new regression model: Modal linear regression," Scand. J. Stat. Theory Appl., vol.41, no.3, pp.656–671, Sept. 2014.
[60] A. Pensia, V. Jog, and P.L. Loh, "Estimating location parameters in entangled single-sample distributions," preprint arXiv:1907.03087, 70 pages, 2019.
[61] K. Dearborn and R. Frongillo, "On the indirect elicitability of the mode and modal interval," Ann. Inst. Stat. Math., vol.72, pp.1095–1108, 2020.
[62] H. Ota and S. Hara, "On estimation of conditional modes using multiple quantile regressions," preprint arXiv:1712.08754, 26 pages, 2017.
[63] W. Heß and J.W. Schmidt, "Positive quartic, monotone quintic C^2-spline interpolation in one and two dimensions," J. Comput. Appl. Math., vol.55, no.1, pp.51–67, Oct. 1994.
[64] D. Kitahara and I. Yamada, "Probability density function estimation by positive quartic C^2-spline functions," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Brisbane, Australia, pp.3556–3560, Apr. 2015.
[65] D. Kitahara and I. Yamada, "Two-dimensional positive spline smoothing and its application to probability density estimation," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Shanghai, China, pp.4219–4223, Mar. 2016.
[66] MATLAB, Quadratic Programming Algorithms, https://www.mathworks.com/help/optim/ug/quadratic-programming-algorithms.html, accessed Feb. 15, 2022.
[67] R.J. Hyndman, D.M. Bashtannyk, and G.K. Grunwald, "Estimating and visualizing conditional densities," J. Comput. Graph. Stat., vol.5, no.4, pp.315–336, Dec. 1996.
[68] MATLAB, Kernel Smoothing Function Estimate for Univariate and Bivariate Data: ksdensity, https://www.mathworks.com/help/stats/ksdensity.html, accessed Feb. 15, 2022.
[69] MATLAB, Smoothing Splines, https://www.mathworks.com/help/curvefit/smoothing-splines.html, accessed Feb. 15, 2022.
[70] China Meteorological Data Service Centre, http://data.cma.cn/en, accessed Feb. 15, 2022.
[71] D. Bolton, "The computation of equivalent potential temperature," Mon. Weather Rev., vol.108, no.7, pp.1046–1053, July 1980.



Sai Yao received his B.E. degree in software engineering from Dalian University of Technology, Dalian, China in 2018, and his B.E. and M.E. degrees in information science and engineering from Ritsumeikan University, Shiga, Japan in 2018 and 2020, respectively. His research interest includes data analysis and spline smoothing.

Daichi Kitahara received his B.E. degree in computer science in 2012, and his M.E. and Ph.D. degrees in communications and computer engineering in 2014 and 2017 from the Tokyo Institute of Technology, Tokyo, Japan, respectively. From 2017 to 2022, he was an Assistant Professor with the College of Information Science and Engineering, Ritsumeikan University. Currently he is a Researcher with the Graduate School of Engineering, Osaka University. His research interest includes signal processing and its applications, inverse problem, optimization theory, and multivariate spline theory.

Hiroki Kuroda received his B.E. degree in computer science, and his M.E. and Ph.D degrees in information and communications engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2013, 2015, and 2019, respectively. Currently, he is an Assistant Professor with the Department of Information and Management Systems Engineering, Nagaoka University of Technology. His current research interests include signal processing and its applications, nonlinear inverse problem, and sparse optimization.

Akira Hirabayashi received his D.E. degree in computer science in 1999, from the Tokyo Institute of Technology, Tokyo, Japan. From 2013, he is a Professor of the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. His research interests include compressed sensing, sampling theory, and signal and image processing. Prof. Hirabayashi is also a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Acoustical Society of Japan (ASJ).