Bandwidth selection for nonparametric modal regression

Haiming Zhou

Division of Statistics, Northern Illinois University, DeKalb, Illinois 60115, U.S.A.

Xianzheng Huang

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.

Abstract

In the context of estimating local modes of a conditional density based on kernel density estimators, we show that existing bandwidth selection methods developed for kernel density estimation are unsuitable for mode estimation. We propose two methods to select bandwidths tailored for mode estimation in the regression setting. Numerical studies using synthetic data and a real-life data set are carried out to demonstrate the performance of the proposed methods in comparison with several well received bandwidth selection methods for density estimation.

Keywords: Bootstrap; Cross validation; Hausdorff distance; Mean shift algorithm.

1 Introduction

In a regression problem, it is often of interest to infer some typical value(s) of a response, Y, given a covariate value, $X = x$. One may view the mean or median associated with the conditional probability density function (pdf), $p(y|x)$, as a typical value. But when $p(y|x)$ is skewed or multimodal, modes of the conditional distribution can be better representations of the response. In such scenarios, the conditional modes can reveal important information regarding the association of Y and X that the conditional mean or quantiles cannot provide; and a mode can yield more precise prediction than the mean and median. These advantages

of conditional modes have been especially appreciated among researchers in traffic engineering (Einbeck and Tutz, 2006), meteorology (Hyndman et al., 1996), astronomy (Bamford et al., 2008), and economics (Huang and Yao, 2012), for instance.

Existing methods for nonparametric estimation of conditional modes are based on kernel density estimators of $p(y|x)$. Given a kernel density estimate, $\hat{p}(y|x)$, the mean shift algorithm is employed to find local modes of the estimated conditional density (Comaniciu and Meer, 2002; Einbeck and Tutz, 2006; Chen et al., 2016), leading to a mode (set) estimate. With this dependence of mode estimation on a kernel density estimator, one may naturally adopt well justified bandwidth selection method for density estimation in order to estimate modes. Natural as this idea is, we show in this article that bandwidths desirable, or even optimal (in some sense), for density estimation are usually not suitable for mode estimation.

To overcome the drawback of existing bandwidth selection approaches, we propose two methods to choose bandwidths in the context of modal regression. We review the methodology of nonparametric modal regression in Section 2. Then we relate four existing bandwidth selection methods in Section 3. These serve as the competing methods with which we compare our proposed strategies, which are described in Section 4. Section 5 presents simulation studies to compare the six methods, and apply them to the Old Faithful geyser data set. Section 6 gives a recap of our findings in this study, where we also point out limitations of the proposed methods, and suggest future research directions for developing improved bandwidth selection methods.

2 Modal regression

Denote by $\mathscr{X}$ the support of X and by $\mathscr{Y}$ the support of Y. Given $x \in \mathscr{X}$, the mode set of $p(y|x)$ is $M(x) = \{y \in \mathscr{Y} : p_y(y|x) = 0 \text{ and } p_{yy}(y|x) < 0\}$, where $p_y(y|x) = (\partial/\partial y)p(y|x)$ and $p_{yy}(y|x) = (\partial^2/\partial y^2)p(y|x)$. The notational convention of using subscripts attaching to a function to refer to partial derivatives of the function is used throughout the article. We are interested in estimating $M(x)$ in this study. Let $\{(X_i, Y_i)\}_{i=1}^n$ be a random sample from the joint distribution of (X, Y), specified by the joint pdf $p(x, y)$. A simple nonparametric

estimator of $p(y|x)$ is

$$\hat{p}(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y}{h_2}\right)}{\frac{1}{nh_1}\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right)},$$

where h_1 and h_2 are bandwidths, and $K_1(t)$ and $K_2(t)$ are kernel functions. It follows that an estimator of $M(x)$ is given by $\hat{M}(x) = \{y \in \mathscr{Y} : \hat{p}_y(y|x) = 0 \text{ and } \hat{p}_{yy}(y|x) < 0\}$.

A computationally efficient algorithm to find $\hat{M}(x)$ is the so-called mean shift algorithm. The algorithm is developed when $K_2(t)$ is radially symmetric (Comaniciu and Meer, 2002) and its derivative satisfies $K_2'(t) = cK_3(t)t$, where c is some negative constant and $K_3(t)$ is a kernel. The standard normal kernel and the Epanechnikov kernel are instances where a kernel possesses these features. For illustration purposes, we set $K_2(t)$ as the standard normal kernel in the sequel. With this choice of $K_2(t)$, the equation one solves for y to find the local modes of $\hat{p}(y|x)$, originating from $\hat{p}_y(y|x) = 0$, reduces to

$$\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y}{h_2}\right)(Y_i - y) = 0.$$

Starting from multiple initial values, the mean shift algorithm entails repeatedly evaluating the following updating formula until convergence,

$$y^{(k+1)} = \frac{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y^{(k)}}{h_2}\right) Y_i}{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y^{(k)}}{h_2}\right)}, \tag{2.1}$$

where $y^{(k)}$ and $y^{(k+1)}$ denote generically two adjacent updated values of y.

As in most kernel-based estimation, the so-obtained $\hat{M}(x)$ is sensitive to the choice of bandwidths, $\mathbf{h} = (h_1, h_2)$. One may conjecture that a good choice of $\mathbf{h}$ for estimating $p(y|x)$ is also good for estimating $M(x)$. Next we review four bandwidth selection methods that have different rationales and have been shown to perform well in existing literature on density estimation.

3 Bandwidth selection for density estimation

Most well received approaches for choosing bandwidths in conditional density estimators aim at finding $\mathbf{h}$ that minimizes a loss function defined as the weighted integrated squared error,

$$\mathrm{ISE}_D(\mathbf{h}) = \int\int \{\hat{p}(y|x) - p(y|x)\}^2 p(x) w(x) dx dy, \tag{3.1}$$

or the corresponding risk function referred to as the weighted integrated mean squared error,

$$\mathrm{IMSE}_D(\mathbf{h}) = \int\int E\{\hat{p}(y|x) - p(y|x)\}^2 p(x) w(x) dx dy,$$

where $p(x)$ is the pdf of X, and $w(x)$ is a nonnegative weight function with bounded support used to avoid estimating $p(y|x)$ at an x around which data are too sparse. Given the observed data $\mathbf{X} = (X_1, \ldots, X_n)$, one may let $w(x) = I(x \in [x_L, x_U])$, where $I(t)$ is the indicator function, and x_L and x_U are, for example, the 2.5th and 97.5th percentile of $\mathbf{X}$, respectively. All integrals in this article integrate over the support of the corresponding variable.

Bashtannyk and Hyndman (2001) compared a variety of bandwidth selection methods based on parametric estimation of IMSE_D. We revisit three of these strategies in this section that are motivated by different viewpoints. The fourth strategy revisited in this section is proposed by Fan and Yim (2004) and Hall et al. (2004), who derived a cross validation (CV) criterion by approximating ISE_D. All four methods have been shown to possess good performance in estimating conditional densities in certain scenarios.

3.1 Reference rules

Letting $K_1(t) = K_2(t) = K(t)$ and defining $\mu_k = \int t^k K(t) dt$, $\nu_k = \int t^k K^2(t) dt$, for $k = 0, 1, \ldots$, Hyndman et al. (1996) showed that

$$\mathrm{IMSE}_D(\mathbf{h}) \approx \frac{c_1}{n h_1 h_2} - \frac{c_2}{n h_1} + c_3 h_1^4 + c_4 h_2^4 + c_5 h_1^2 h_2^2, \tag{3.2}$$

where

$$
\begin{aligned}
c_1 &= \int \nu_0^2 w(x) dx, \\
c_2 &= \iint \nu_0 p^2(y|x) w(x) dy dx, \\
c_3 &= \iint \frac{\mu_2^2 p(x)}{4} \left\{ 2\frac{p_x(x)}{p(x)} p_x(y|x) + p_{xx}(y|x) \right\}^2 w(x) dy dx, \\
c_4 &= \iint \frac{\mu_2^2 p(x)}{4} \left\{ p_{yy}(y|x) \right\}^2 w(x) dy dx, \\
c_5 &= \iint \frac{\mu_2^2 p(x)}{2} \left\{ 2\frac{p_x(x)}{p(x)} p_x(y|x) + p_{xx}(y|x) \right\} p_{yy}(y|x) w(x) dy dx.
\end{aligned}
$$

By minimizing (3.2) with respect to (w.r.t.) $\mathbf{h}$, Hyndman et al. (1996) showed that the approximated optimal bandwidths are given by

$$
\hat{h}_1 = c_1^{1/6} \left\{ 4 \left(\frac{c_3^5}{c_4} \right)^{1/4} + 2c_5 \left(\frac{c_3}{c_4} \right)^{3/4} \right\} n^{-1/6}, \qquad \hat{h}_2 = \hat{h}_1 \left(\frac{c_3}{c_4} \right)^{1/4}. \tag{3.3}
$$

Further assuming that $p(x)$ is a normal pdf, and $p(y|x)$ is a normal pdf with mean and standard deviation both being linear functions of x, Bashtannyk and Hyndman (2001) elaborated (3.3), resulting in the reference rule denoted by $\mathbf{h}_N = (\hat{h}_{1,N}, \hat{h}_{2,N})$.

3.2 Regression-based bandwidth selection

Motivated by the property of the scaled kernel that $E\{K_{h_2}(Y - y)|X = x\} \approx p(y|x)$ as $h_2 \to 0$, where $K_h(t) = K(t/h)/h$, Fan et al. (1996) developed estimators of $p(y|x)$ following local polynomial estimation of the mean function (Fan and Gijbels, 1996), viewing $p(y|x)$ as the conditional mean when regressing $K_{h_2}(Y - y)$ on X. Following this view, Bashtannyk and Hyndman (2001) proposed to first fix h_2 at a reference rule, then select h_1 by minimizing the penalized mean squared prediction error defined by

$$
Q_{h_2}(h_1) = \frac{\Delta}{n} \sum_{i=1}^{n} \sum_{k=1}^{\mathcal{M}} \left\{ \hat{p}(y_k|X_i) - K_{h_2}(Y_i - y_k) \right\}^2 w(X_i) \Xi\{W_{h_1,i}(X_i)\}, \tag{3.4}
$$

where $\{y_k\}_{k=1}^{\mathcal{M}}$ is a sequence of equally spaced grid points over $\mathscr{Y}$, Δ is the distance between two adjecent grid points, $\Xi(u)$ is a penalty function that has the first order Taylor expansion of the form $\Xi(u) = 1 + 2u + O(u^2)$, and $W_{h_1,i}(x) = K_{h_1}(X_i - x) / \sum_{j=1}^{n} K_{h_1}(X_j - x)$, for

$i = 1, \ldots, n$. Köhler et al. (2014) provided a review of popular choices of the penalty function. We use the penalty $\Xi(u) = (1+u)/(1-u)$ as in Akaike (1970) in the simulation study in Section 5. Denote by $\mathbf{h}_R = (\hat{h}_{1,R}, \hat{h}_{2,R})$ the bandwidths resulting from this method.

3.3 Bootstrap bandwidth selection

Yet another approach considered in Bashtannyk and Hyndman (2001) involves parametric bootstrap, mimicking the idea in Hall et al. (1999). This method targets at finding $\mathbf{h}$ that minimizes the following empirical version of $\mathrm{ISE}_D(\mathbf{h})$,

$$A(\mathbf{h}; \mathbf{X}, \mathbf{Y}, \hat{p}, p) = \frac{\Delta}{n} \sum_{i=1}^{n} \sum_{k=1}^{\mathcal{M}} \left\{ \hat{p}(y_k|X_i) - p(y_k|X_i) \right\}^2 w(X_i), \tag{3.5}$$

where $\hat{p}(y|x)$ is an estimate of $p(y|x)$ based on the observed data $(\mathbf{X}, \mathbf{Y})$ given $\mathbf{h}$. Since (3.5) depends on the unknown $p(y|x)$, Bashtannyk and Hyndman (2001) constructed the following estimate of (3.5) based on $\mathcal{L}$ bootstrap samples, $\{(\mathbf{X}, \mathbf{Y}^{(\ell)})\}_{\ell=1}^{\mathcal{L}}$, simulated from an estimate of $p(y|x)$ denoted by $\hat{p}^*(y|x)$,

$$\hat{A}(\mathbf{h}) = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} A(\mathbf{h}; \mathbf{X}, \mathbf{Y}^{(\ell)}, \hat{p}^{(\ell)}, \hat{p}^*), \tag{3.6}$$

where $\hat{p}^{(\ell)}$, referring to $\hat{p}^{(\ell)}(y|x)$, is the same type of estimate as $\hat{p}(y|x)$ but computed based on the ℓth bootstrap sample, $(\mathbf{X}, \mathbf{Y}^{(\ell)})$, for $\ell = 1, \ldots, \mathcal{L}$; and $\hat{p}^*$, referring to $\hat{p}^*(y|x)$, is a parametric estimate of $p(y|x)$. More specifically, $\hat{p}^*(y|x)$ results from fitting a parametric model (using data $(\mathbf{X}, \mathbf{Y})$) of Y given X specified by

$$Y_i = \beta_0 + \beta_1 X_i + \ldots + \beta_k X_i^k + \sigma \epsilon_i, \text{ for } i = 1, \ldots, n,$$

assuming $\{\epsilon_i\}_{i=1}^n$ independent model errors from $N(0,1)$, where k is determined by the Akaike information criterion (AIC). Once $\hat{p}^*(y|x)$ is obtained, one uses this model to generate the bootstrap responses, $\mathbf{Y}^{(\ell)}$, given $\mathbf{X}$, for $\ell = 1, \ldots, \mathcal{L}$. Denote by $\mathbf{h}_B = (\hat{h}_{1,B}, \hat{h}_{2,B})$ the bandwidths selected by this approach.

3.4 A cross validation method

Fan and Yim (2004) and Hall et al. (2004) proposed a cross-validation criterion based on an elaboration of $\mathrm{ISE}_{D}(\mathbf{h})$ in (3.1) as follows,

$$\begin{aligned}\mathrm{ISE}_{D}(\mathbf{h}) = &\iint \hat{p}(y|x)^2 p(x) w(x) dx dy - 2 \iint \hat{p}(y|x) p(x,y) w(x) dx dy \\ &+ \int p(y|x)^2 p(x) w(x) dx dy.\end{aligned} \tag{3.7}$$

Note that the third term does not depend on $\mathbf{h}$ and thus can be ignored when one minimizes $\mathrm{ISE}_{D}(\mathbf{h})$ w.r.t. $\mathbf{h}$. Therefore, they proposed the following estimator of the first two terms in (3.7) as a CV criterion,

$$\mathrm{CV}_{D}(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} w(X_i) \int \hat{p}_{-i}(y|X_i)^2 dy - \frac{2}{n}\sum_{i=1}^{n} w(X_i) \hat{p}_{-i}(Y_i|X_i), \tag{3.8}$$

where $\hat{p}_{-i}(y|X_i)$ is an estimate of $p(y|x)$ based on data $\{(X_j, Y_j), j \neq i\}$ given $\mathbf{h}$. With the kernel associated with Y being a standard normal pdf, the integral in (3.8) can be derived explicitly. Denoted by $\mathbf{h}_{D} = (\hat{h}_{1,D}, \hat{h}_{2,D})$ the value of $\mathbf{h}$ that minimizes (3.8).

4 Bandwidth selection for mode estimation

4.1 Preliminary

In contrast to mean and quantile regression, the unknown quantity to be estimated in modal regression is not a one-to-one function, but a set, or a one-to-many function, of which the size is also unknown. This makes constructing a sensible proxy of a loss function less straightforward. For a mode estimate $\hat{M}(x)$, a reasonable loss function is the weighted integrated squared error defined by

$$\mathrm{ISE}_{M}(\mathbf{h}) = \int \{\mathrm{Haus}(\hat{M}(x),\, M(x))\}^2 p(x) w(x) dx, \tag{4.1}$$

where $\mathrm{Haus}(A, B) = \inf\{r :\, A \subset B \oplus r,\, B \subset A \oplus r\}$ is the Hausdorff distance between two sets, A and B, in which $A \oplus r = \{b : \inf_{a \in A} d(a, b) \leq r\}$, and $B \oplus r$ is similarly defined, in which $d(a, b)$ denotes the Euclidean distance between two points, a and b. In mean and quantile regression, a proxy of a loss function associated with a mean/quantile estimate is easily

obtained using residuals. For example, the weighted mean squared residuals, $n^{-1}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 w(X_i)$, is a proxy of the weighted integrated squared error of a mean estimate $\hat{m}(x)$, $\int\{\hat{m}(x) - m(x)\}^2 p(x)w(x)dx$, where $m(x) = E(Y|X = x)$ and $\hat{Y}_i = \hat{m}(X_i)$. In quantile regression, the check function evaluated at the residual is used to construct a CV criterion (Koenker, 2005). When both X and Y are continuous, given an X_i, there is typically only one corresponding Y_i observed, and thus it is not clear how to construct a residual associated with a set estimate $\hat{M}(X_i)$. It certainly should not be $d(\hat{M}(X_i),\, Y_i,\,)$, where $d(A, a)$ denotes the distance between a set A and a point a, defined as the minimum Euclidean distance between a and $b \in A$. This is because $d(\hat{M}(X_i),\, Y_i)$ is not guaranteed to represent well $\text{Haus}(\hat{M}(X_i),\, M(X_i))$ when $p(y|X_i)$ is multimodal, even if $Y_i \in M(X_i)$.

If one considers $\hat{M}(x)$ as a byproduct of density estimation, one may want to use the bandwidths chosen for estimating $p(y|x)$ in modal regression. But, since a smaller ISE_D or IMSE_D does not necessarily imply a smaller ISE_M, we conjecture that bandwidth selection methods designed for density estimation are not adequate for modal regression. The intuition is that, in order to infer the modes of a conditional density, one does not have to estimate well features like the tails of the distribution or the height of a mode. Yet most well accepted density estimation methods do strive to capture these features, and the accompanying bandwidth selection methods are often distracted by, for instance, the tail behavior of $p(y|x)$ (Hall, 1992).

Chen et al. (2016) assumed $h_1 = h_2 = h$ and proposed to use the volume of an estimated prediction set to choose h. Following this idea, a loss function of $\hat{M}(x)$ is defined by $\text{Vol}(h) = \hat{\epsilon}_{1-\alpha,h}\int N(x)dx$, where $\epsilon_{1-\alpha,h}$ is the $(1-\alpha)$ quantile of $\{d(\hat{M}(X_i),\, Y_i)\}_{i=1}^{n}$ and $N(x)$ is the size of $\hat{M}(x)$, i.e., the number of points in $\hat{M}(x)$. This loss function is constructed to balance between the number of estimated local modes and the distance between the estimated modes and $\mathbf{Y}$. This method has two pitfalls. First, it relies on an extra tuning parameter α; and, second, setting $h_1 = h_2$ is not well justified, especially in a regression setting where X and Y play very different roles.

4.2 Two proposed methods

Also hoping to account for the size of a mode set estimate while striving for accurate prediction as in Chen et al. (2016), we propose the following CV criterion,

$$\mathrm{CV}_{M}(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} d^2(\hat{M}_{-i}(X_i),\, Y_i) N_{-i}^2(X_i) w(X_i), \tag{4.2}$$

where $\hat{M}_{-i}(X_i)$ is an estimate of $M(X_i)$ based on data $\{(X_j, Y_j), j \neq i\}$ given $\mathbf{h}$, and $N_{-i}(X_i)$ is the size of $\hat{M}_{-i}(X_i)$, for $i = 1, \ldots, n$. Denote by $\mathbf{h}_{M} = (\hat{h}_{1,M},\, \hat{h}_{2,M})$ the bandwidths that minimize (4.2).

Our second proposal is motivated by estimating $\mathrm{ISE}_{M}(\mathbf{h})$ in (4.1). Given $M(X_i)$, an empirical version of $\mathrm{ISE}_{M}(\mathbf{h})$ is

$$A_{M}(\mathbf{h}; \mathbf{X}, \mathbf{Y},\, \hat{M},\, M) = \frac{1}{n}\sum_{i=1}^{n} \Big\{\mathrm{Haus}(\hat{M}(X_i),\, M(X_i))\Big\}^2 w(X_i),$$

where $\hat{M}(x)$ is the nonparametric estimate of $M(x)$ based on data $(\mathbf{X}, \mathbf{Y})$ given $\mathbf{h}$. Since $A_{M}(\mathbf{h}; \mathbf{X}, \mathbf{Y},\, \hat{M},\, M)$ is not available in practice due to its dependence on the unknown $M(X_i)$, we use $\mathcal{L}$ bootstrap samples, $\{(\mathbf{X},\, \mathbf{Y}^{(\ell)})\}_{\ell=1}^{\mathcal{L}}$, to estimate it via

$$\hat{A}_{M}(\mathbf{h}) = \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}} A_{M}(\mathbf{h}; \mathbf{X}, \mathbf{Y}^{(\ell)},\, \hat{M}^{(\ell)},\, \hat{M}^{*}), \tag{4.3}$$

where $\hat{M}^{(\ell)}$ refers to the mode estimate, of the same type of estimate as $\hat{M}(x)$, based on the ℓth bootstrap sample, for $\ell = 1, \ldots, \mathcal{L}$, and $\hat{M}^{*}$ refers to a mode estimate obtained from a parametric estimate of $p(y|x)$, denoted by $\tilde{p}^{*}(y|x)$. In particular, $\tilde{p}^{*}(y|x)$ results from fitting a finite mixture model, implemented by R package `flexmix`,

$$Y|X \sim \sum_{k=1}^{K} \pi_k \phi\left(\frac{y - \beta_{k0} - \beta_{k1} b_1(X) - \ldots - \beta_{kJ} b_J(X)}{\sigma_k}\right),$$

where $\beta_{k0}, \ldots, \beta_{kJ}$ and σ_k are estimated using data $(\mathbf{X}, \mathbf{Y})$, $\{b_1(\cdot), \ldots, b_J(\cdot)\}$ form a B-spline basis, and J is determined by AIC. Once a parametric estimate $\tilde{p}^{*}(y|x)$ is obtained, we generate bootstrap sample $\mathbf{Y}^{(\ell)}$ given $\mathbf{X}$, for $\ell = 1, \ldots, \mathcal{L}$. Denote by $\mathbf{h}_{B}^{*} = (\hat{h}_{1,B}^{*},\, \hat{h}_{2,B}^{*})$ the resultant bandwidths.

5 Empirical study

5.1 Simulation design

We design simulation experiments aiming to, first, demonstrate the bandwidths chosen by different methods, second, compare performance of the density estimator and mode estimator when different bandwidths are used. In the simulation experiment, we consider the following five true model configurations:

(C1) $Y = m(X) - 1 + \epsilon$, where $m(x) = x + x^2$, $\epsilon \sim \Gamma(3, 2)$, and $X \sim N(0, 1)$. Here $\Gamma(a, b)$ is the gamma distribution with mean a/b. In this case, $p(y|x)$ is unimodal with $M(x) \approx \{m(x)\}$.

(C2) $[Y|X = x] \sim 0.5N\left(m_1(x),\, 1\right) + 0.5N\left(m_2(x),\, 1\right)$, where $m_1(x) = x + x^2$, $m_2(x) = m_1(x) - 6$, and $X \sim N(0, 1)$. In this case, $p(y|x)$ is bimodal with $M(x) \approx \{m_1(x),\, m_2(x)\}$.

(C3) For $x \leq 0$, the model of $[Y|X = x]$ is the same as that under (C1); for $x > 0$, the model of $[Y|X = x]$ is the same as that under (C2), where $X \sim N(0, 1)$. In this case, $p(y|x)$ is unimodal if $x \leq 0$, and it is bimodal if $x > 0$.

(C4) $[Y|X = x] \sim 0.5N\left(m_1(x),\, 0.5^2\right) + 0.3N\left(m_2(x),\, 0.5^2\right) + 0.2N\left(m_3(x),\, 0.5^2\right)$, where $m_1(x) = x + x^2$, $m_2(x) = m_1(x) - 3$, $m_3(x) = m_1(x) - 6$, and $X \sim N(0, 1)$. In this case, $p(y|x)$ is trimodal with $M(x) \approx \{m_1(x),\, m_2(x),\, m_3(x)\}$.

(C5) $[Y|X = x] \sim 0.2\sum_{j=1}^{5} N\left(m_j(x),\, 0.2^2\right)$, where $m_j(x) = x + x^2 - 1.5(j - 1)$, and $X \sim N(0, 1)$. In this case, $p(y|x)$ has five modes in $M(x) \approx \{m_j(x),\, j = 1, \ldots, 5\}$.

Under each true model configuration, we generate 500 Monte Carlo (MC) replicates, each of sample size $n = 500$, from the true model of (X, Y). Based on each MC replicate, we use the R package `hdrcde` to obtain the bandwidths $\mathbf{h}_N$, $\mathbf{h}_R$, and $\mathbf{h}_B$, and use the R package `lpme`, created by the first author, to obtain the bandwidths $\mathbf{h}_D$, $\mathbf{h}_M$, and $\mathbf{h}_B^*$. We then use each of the six pairs of bandwidths to estimate the conditional density, $p(y|x)$, followed by estimating the mode set, $M(x)$. To address the aforementioned second aim, we compute two metrics associated with density estimation and mode estimation, respectively, one is the

density-based empirical integrated squared error (EISE),

$$\text{EISE}_D = \sum_{j=1}^{\mathcal{M}'} \sum_{k=0}^{\mathcal{M}} \left\{\hat{p}(y_j|x_k) - p(y_j|x_k)\right\}^2 p(x_k)\Delta\Delta',$$

the other is the mode-based EISE,

$$\text{EISE}_M = \sum_{k=0}^{\mathcal{M}} \left\{\text{Haus}(\hat{M}(x_k), M(x_k))\right\}^2 p(x_k)\Delta,$$

where $\{x_k = x_L + k\Delta\}_{k=0}^{\mathcal{M}}$, Δ is the partition resolution, $\mathcal{M}$ is the largest integer no greater than $(x_U - x_L)/\Delta$, and $\{y_j\}_{j=1}^{\mathcal{M}'}$ is a sequence of grid points equally spaced over the observed sample range of $\mathbf{Y}$ with $y_{j+1} - y_j = \Delta'$.

Finally, to formulate some benchmarks with which we compare the six pairs of bandwidths, we also find the bandwidths that minimize EISE_D and the ones that minimize EISE_M, denoted by $\tilde{\mathbf{h}}_D = (\tilde{h}_{1,D}, \tilde{h}_{2,D})$ and $\tilde{\mathbf{h}}_M = (\tilde{h}_{1,M}, \tilde{h}_{2,M})$, respectively. Naturally, $\tilde{\mathbf{h}}_D$ can be viewed as the optimal choice of $\mathbf{h}$ for the purpose of density estimation, which is practically unattainable due to the dependence of EISE_D on the unknown true density. Similarly, $\tilde{\mathbf{h}}_M$ can be viewed as the (unrealistic) optimal choice of $\mathbf{h}$ for the purpose of mode estimation. Using these two pairs of optimal (in different senses) bandwidths, we also obtain the corresponding density/mode estimates, which are the benchmark estimates with which we compare the other density/mode estimates.

5.2 Simulation results

Figures 1–5 show the estimated mode curves under the five true model configurations with $[x_L, x_U] = [-2, 2]$. Under each true model configuration, the comparison between the mode estimates resulting from the two optimal (in different senses) bandwidths, $\tilde{\mathbf{h}}_D$ and $\tilde{\mathbf{h}}_M$ (see panels (a) and (e)), support our earlier conjecture that bandwidths suitable for density estimation are poor choices for mode estimation. In particular, using $\tilde{\mathbf{h}}_D$ results in overfitted mode curves, which are much more noisy than the estimated mode curves when $\tilde{\mathbf{h}}_M$ is used. The phenomenon of overfitting is also evident in the estimated mode curves resulting from setting $\mathbf{h} = \mathbf{h}_D$, and also often seen when letting $\mathbf{h} = \mathbf{h}_B$, both choices of $\mathbf{h}$ lead to mode estimates clearly outperformed by mode estimates when our proposed bandwidths, $\mathbf{h}_M$ or

$\mathbf{h}_B^*$, are used. The overfitting trend observed when using the two density-based bandwidths, $\tilde{\mathbf{h}}_D$ and $\mathbf{h}_D$, does less harm when $p(y|x)$ has many modes, as seen under (C5), where mode estimates resulting from using $\tilde{\mathbf{h}}_D$ and $\mathbf{h}_D$ are relatively comparable with estimates resulting from using $\tilde{\mathbf{h}}_M$ and $\mathbf{h}_M$, respectively (see panels (a), (b), (e) and (f) in Figure 5). When the normal reference $\mathbf{h}_N$ or the regression-based bandwidths $\mathbf{h}_R$ are used, although less noisy, the resultant estimated mode curves exhibit underfitting and fail to capture key features of the true conditional mode curves around the boundary or at the valley. In fact, they fail miserably at other regions of $\mathscr{X}$ too when there are many modes as in (C5) (see panels (d) and (h) in Figure 5).

To address the first aim of the simulation experiment, Figures 6–10 present the scatter plots of chosen bandwidths across 500 MC replicates. The depicted optimal (in different senses) bandwidths (see panels (a) and (e)) reveal that $\tilde{\mathbf{h}}_D$ and $\tilde{\mathbf{h}}_M$ are indeed very different, especially in the bandwidth associated with y, except for (C5) that has the largest number of modes among the fives cases. In particular, $\hat{h}_{2,D}$ is substantially smaller than $\hat{h}_{2,M}$ under (C1)–(C4), which explains the severe overfitting when setting $\mathbf{h} = \tilde{\mathbf{h}}_D$ observed in Figures 1–4. If the interest lies in density estimation, among the four density-based methods, the one yielding $\mathbf{h}_D$ (see panel (f)) is a clear winner since $\mathbf{h}_D$ is closer to $\tilde{\mathbf{h}}_D$ than $\mathbf{h}_N$, $\mathbf{h}_R$, and $\mathbf{h}_B$. But if one is interested in inferring conditional modes, all four existing methods miss the mark since none of them yield bandwidths approximating $\tilde{\mathbf{h}}_M$ well. In contrast, our proposed mode-based bandwidths, $\mathbf{h}_M$ and $\mathbf{h}_B^*$, are much more promising in approximating $\tilde{\mathbf{h}}_M$, especially under (C2). Between these two, $\mathbf{h}_B^*$ tends to be more variable than $\mathbf{h}_M$, which can be due to the mixture model estimation.

Figure 11 addresses the second aim by depicting boxplots of EISE_M and EISE_D evaluated at eight choices of $\mathbf{h}$. One can see (from the top panels) that the two mode-based methods yield much smaller EISE_M than all density-based methods. Between these two methods, the one involving bootstrap (with $\mathbf{h} = \mathbf{h}_B^*$) produces more variable EISE_M than the method involving CV (with $\mathbf{h} = \mathbf{h}_M$). This is expected due to the much higher variability of $\mathbf{h}_B^*$ than that of $\mathbf{h}_M$ noted earlier. When it comes to density estimation (see the bottom panels), among the four density-based methods, the one involving CV (with $\mathbf{h} = \mathbf{h}_D$) gives the lowest EISE_D, which is also expected because of the close agreement between $\mathbf{h}_D$ and $\tilde{\mathbf{h}}_D$.

Finally, Table 1 presents the MC averages and standard errors of EISE_M under the five true model configurations. From there one may gain the perception that using $\mathbf{h}_M$ gives slightly better numerical performance than when $\mathbf{h}_B^*$ is used. When there are many modes as in (C5), the mode-based CV method (leading to $\mathbf{h}_M$) and the density-based CV method (leading to $\mathbf{h}_D$) perform similarly. Additionally, under the most challenging simulation setting, (C3), the two density-based methods that give $\mathbf{h}_R$ and $\mathbf{h}_B$ are surprisingly competitive.

5.3 Application to Old Faithful geyser data

Our empirical comparison of various bandwidth selection methods ends with applying the six considered methods to the Old Faithful geyser data analyzed in Bashtannyk and Hyndman (2001). This data set consists of 299 observations of the waiting time (in minutes) between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, collected from August 1st to August 15th, 1985. Applying the four density-based bandwidth selection methods reviewed in Section 3 and the two mode-based bandwidth selection methods proposed in Section 4 to this data set yield the following choices of $\mathbf{h}$: $\mathbf{h}_N = (4.12, 0.87)$, $\mathbf{h}_R = (2.20, 0.87)$, $\mathbf{h}_B = (3.60, 0.40)$, $\mathbf{h}_D = (4.12, 0.09)$, $\mathbf{h}_M = (2.68, 0.60)$, and $\mathbf{h}_B^* = (1.53, 0.37)$. Figure 12 presents the data and the six sets of estimated mode curves corresponding to these choices of bandwidths.

As a reminiscence of the overfitting pattern observed in the simulation study, setting $\mathbf{h} = \mathbf{h}_D$ leads to estimated mode sets that claim far more local modes than the mode estimates from other methods. The estimates resulting from setting $\mathbf{h} = \mathbf{h}_N$ and $\mathbf{h} = \mathbf{h}_R$ are comparable, both less wiggly compared to those resulting from using the mode-based bandwidths, $\mathbf{h}_M$ and $\mathbf{h}_B^*$. This may be a sign of underfitting, a pattern repeatedly observed for these two density-based methods in the simulation study. Between the two mode-based bandwidths, the one involving bootstrap, i.e., $\mathbf{h}_B^*$, leads to more wiggly estimated mode curves. For this particular data set, the performance of the density-based method involving bootstrap, producing $\mathbf{h}_B$, is similar to that when $\mathbf{h}_M$ is used.

6 Discussion

In this study, we are interested in inferring local modes of Y given $X = x$, and we argue that bandwidth selection methods developed for kernel-based density estimation are not suitable for mode estimation. Even though the four density-based bandwidth selection methods considered in this article only give a small subset of the large collection of existing methods for choosing bandwidths in density estimation, they represent four very different strategies of bandwidth selection in that context; and our numerical studies provide convincing evidence that a bandwidth selection method that performs well in estimating the conditional density typically performs poorly in mode estimation.

We proposed two bandwidth selection methods tailored for mode estimation, and demonstrated their promising improvement over the density-based methods in estimating conditional modes. The first proposed method is a cross validation procedure for finding $\mathbf{h}$ that minimizes a CV criterion accounting for the size of a mode set estimate and the distance between this set and the observed response. Due to the difficulty in formulating a proxy for the true mode set at a given observed covariate value, it is unclear if there exists a consistent estimator of $\text{ISE}_{M}(\mathbf{h})$ given in (4.1). The CV criterion in (4.2), $\text{CV}_{M}(\mathbf{h})$, is constructed with the hope that a large $\text{CV}_{M}(\mathbf{h})$ usually implies a large $\text{ISE}_{M}(\mathbf{h})$. Noticing that in the simulation study, under (C1), $\hat{h}_{2,M}$ tends to be larger than the corresponding optimal choice, $\tilde{h}_{2,M}$, we believe that $\text{CV}_{M}(\mathbf{h})$ is an inconsistent estimator of $\text{ISE}_{M}(\mathbf{h})$. Theoretical properties of this CV criterion deserves more in-depth investigation, which can lead to an improved CV procedure.

Our second proposed method depends on a mode set estimate, $\hat{M}^{*}(\cdot)$, as the byproduct of estimating the conditional density via a finite mixture model. This mode set estimate acts like a proxy for the true mode set in the bootstrap procedure that leads to the selected bandwidths $\mathbf{h}_{B}^{*}$. Because estimating a finite mixture model (with unknown number of components) can be subject to high variability, the performance of this bandwidth selection procedure is less stable than the first proposed method. But, with some precautions when fitting a finite mixture model, $\hat{M}^{*}(\cdot)$ can be a well-behaved proxy, in which case using it to choose bandwidths via bootstrap often yield satisfactory mode estimates. A question one

may raise is why not just use $\hat{M}^*(x)$ as the ultimate mode estimate instead of implementing the extra bootstrap procedure to select $\mathbf{h}$ to be used for finding another mode estimate. Indeed, if one wishes to take a parametric or semiparametric route to estimate $M(x)$, one may first estimate the conditional density parametrically or semiparametrically, with finite mixture models as an example, and this (semi)parametric estimate of the conditional density can lead to estimates of local modes. In this study, we estimate $M(x)$ nonparametrically, and we find that the resulting mode estimates using $\mathbf{h} = \mathbf{h}_B^*$ often improves over $\hat{M}^*(x)$.

Relevant works that may provide hints for developing bandwidth selection methods suitable for conditional mode estimation include Chen et al. (2015); Genovese et al. (2016), and Chen et al. (2016), where the authors considered nonparametric estimation of density ridges of the joint distribution of a multivariate random variable. In these works, the authors used the same bandwidth h for all variables in the joint density and proposed different strategies to choose h. In the regression setting where Y and X are treated differently, we observed (in simulation study omitted here) significantly worse mode estimation when assuming $h_1 = h_2$ than when this assumption is relaxed. It is possible that methods proposed in these existing works can be revised to allow different bandwidths for different variables, leading to new strategies in the context of conditional mode estimation. Lastly, both our proposed methods struggle more under (C3) compared to the other four cases, where the number of conditional modes varies over $\mathscr{X}$. We believe that this is the situation that calls for variable bandwidths, $\mathbf{h}(x)$ or $\mathbf{h}(X_i)$, which is beyond the scope of the current study where all considered methods produce fixed bandwidths $\mathbf{h}$. One potential direction to follow in order to choose variable bandwidths is to modify the strategies proposed in Comaniciu et al. (2001), where the authors also treated all variables symmetrically and chose one variable bandwidth, $h(\mathbf{x})$ or $h(\mathbf{X}_i)$, for all variables.

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22*(1), 203–217.

Bamford, S. P., A. L. Rojas, R. C. Nichol, C. J. Miller, L. Wasserman, C. R. Genovese,

and P. E. Freeman (2008). Revealing components of the galaxy population through non-parametric techniques. *Monthly Notices of the Royal Astronomical Society 391*(2), 607–616.

Bashtannyk, D. M. and R. J. Hyndman (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis 36*(3), 279–298.

Chen, Y.-C., C. R. Genovese, S. Ho, and L. Wasserman (2015). Optimal ridge detection using coverage risk. In *Advances in Neural Information Processing Systems*, pp. 316–324.

Chen, Y.-C., C. R. Genovese, R. J. Tibshirani, L. Wasserman, et al. (2016). Nonparametric modal regression. *The Annals of Statistics 44*(2), 489–514.

Chen, Y.-C., C. R. Genovese, L. Wasserman, et al. (2016). A comprehensive approach to mode clustering. *Electronic Journal of Statistics 10*(1), 210–241.

Comaniciu, D. and P. Meer (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence 24*(5), 603–619.

Comaniciu, D., V. Ramesh, and P. Meer (2001). The variable bandwidth mean shift and data-driven scale selection. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Volume 1, pp. 438–445. IEEE.

Einbeck, J. and G. Tutz (2006). Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(4), 461–475.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, Volume 66. Chapman & Hall/CRC.

Fan, J., Q. Yao, and H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika 83*(1), 189–206.

Fan, J. and T. H. Yim (2004). A crossvalidation method for estimating conditional densities. *Biometrika 91*(4), 819–834.

Genovese, C. R., M. Perone-Pacifico, I. Verdinelli, and L. Wasserman (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 99–126.

Hall, P. (1992). On global properties of variable bandwidth density estimators. *The Annals of Statistics*, 762–778.

Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association 99*(468), 1015–1026.

Hall, P., R. C. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association 94*(445), 154–163.

Huang, M. and W. Yao (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association 107*(498), 711–724.

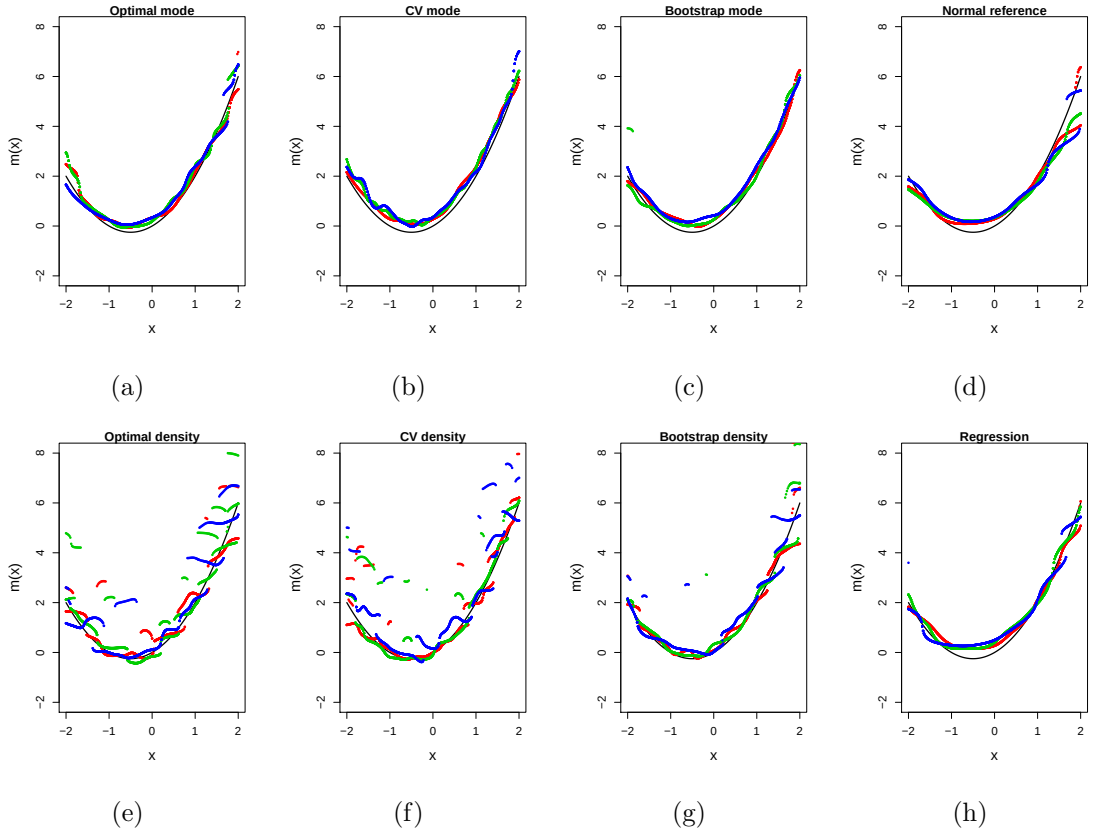
Hyndman, R. J., D. M. Bashtannyk, and G. K. Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics 5*(4), 315–336.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Köhler, M., A. Schindler, and S. Sperlich (2014). A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review 82*(2), 243–274.

Table 1: Monte Carlo averages of EISE_M across 500 MC replicates using eight choices of bandwidths. Numbers in parentheses are (10× standard errors) associated with the averages. Acronym of different methods are the same as those used in Figure 11

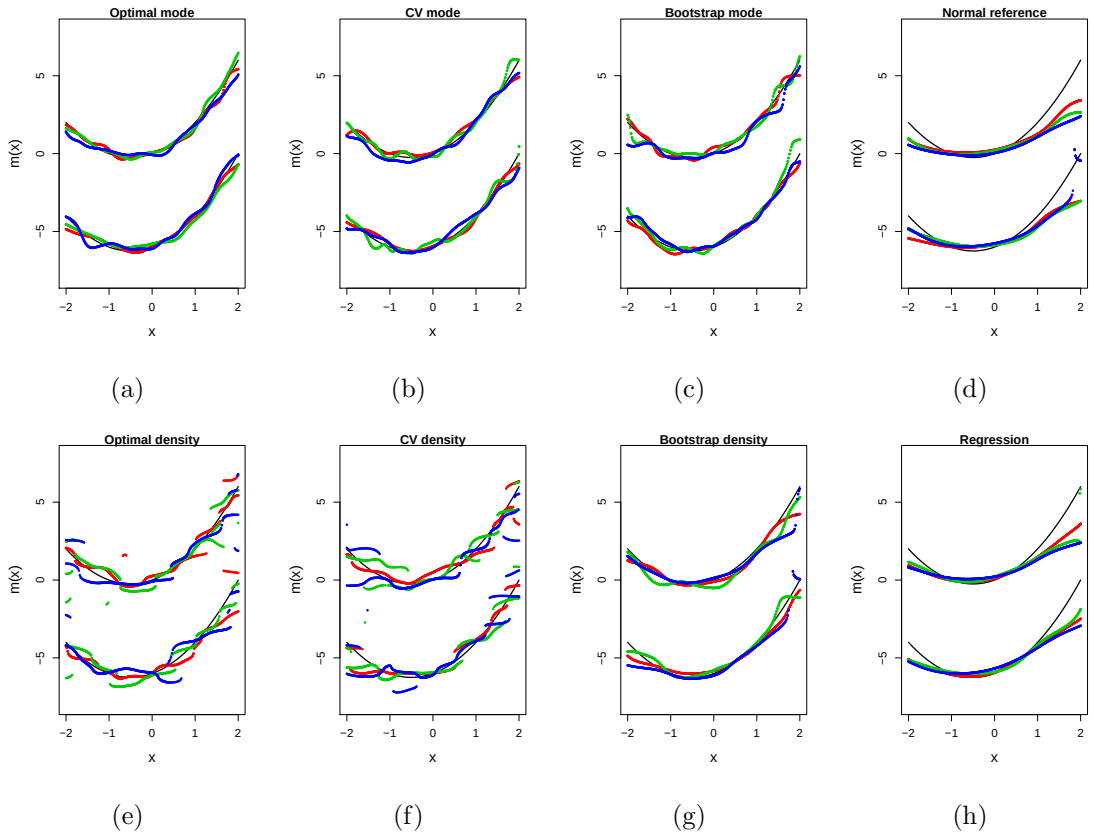
	O-M	CV-M	B-M	O-D	CV-D	B-D	N-D	R-D
(C1)	0.08	0.14	0.14	1.37	1.44	0.41	0.16	0.13
	(0.01)	(0.01)	(0.10)	(0.28)	(0.38)	(0.15)	(0.04)	(0.04)
(C2)	0.08	0.12	0.11	0.47	0.55	0.30	0.55	0.42
	(0.01)	(0.05)	(0.04)	(0.09)	(0.16)	(0.26)	(0.12)	(0.12)
(C3)	0.48	1.62	2.06	3.20	3.25	1.95	2.63	1.93
	(0.18)	(0.29)	(0.30)	(0.32)	(0.37)	(0.16)	(0.18)	(0.15)
(C4)	0.15	0.25	0.21	0.34	0.36	14.8	15.2	12.9
	(0.04)	(0.08)	(0.06)	(0.05)	(0.06)	(3.25)	(2.42)	(2.10)
(C5)	0.46	0.56	0.60	0.53	0.54	10.2	10.2	10.9
	(0.06)	(0.08)	(0.10)	(0.07)	(0.07)	(1.23)	(1.31)	(1.49)



(a) (b) (c) (d)

(e) (f) (g) (h)

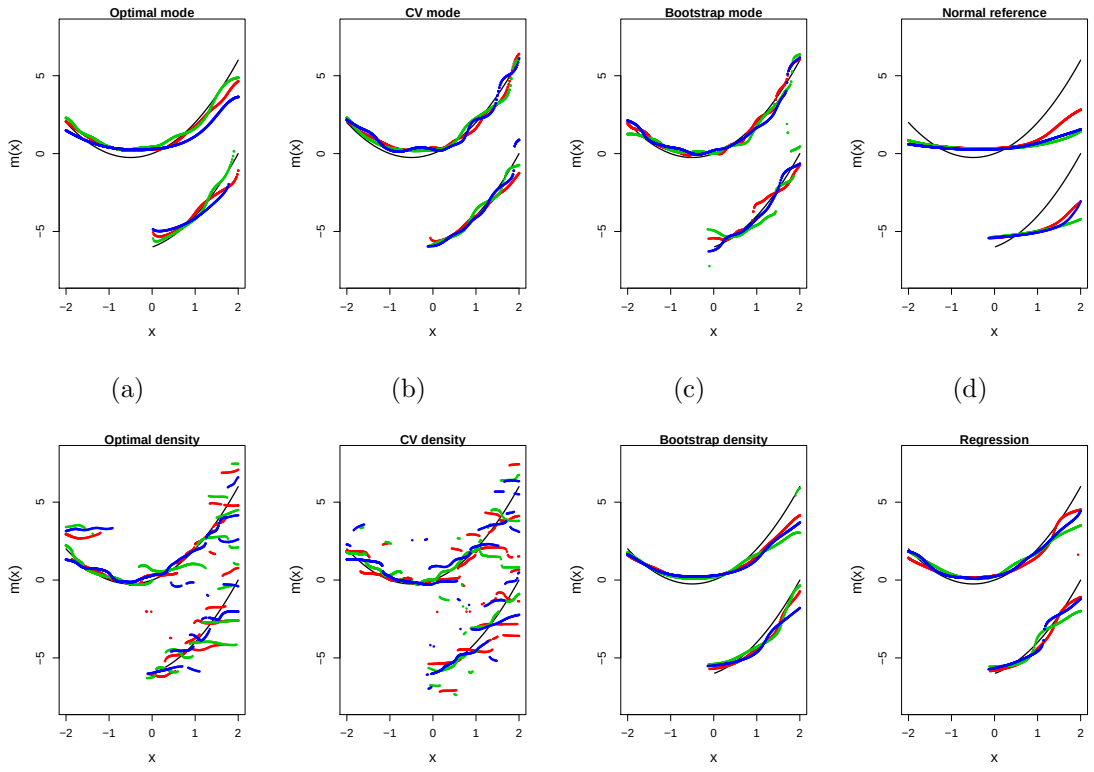
Figure 1: Estimated mode curves resulting from eight choices of bandwidths $\mathbf{h}$ under (C1). These bandwidths are (a) $\tilde{\mathbf{h}}_M$, the optimal bandwidths that minimize EISE_M; (b) $\mathbf{h}_M$, the mode-based bandwidths involving CV; (c) $\mathbf{h}_B^*$, the mode-based bandwidths involving bootstrap; (d) $\mathbf{h}_N$, the normal reference; (e) $\tilde{\mathbf{h}}_D$, the optimal bandwidths that minimize EISE_D; (f) $\mathbf{h}_D$, the density-based bandwidths involving CV; (g) $\mathbf{h}_B$, the density-based bandwidths involving bootstrap; (h) $\mathbf{h}_R$, resulting from the regression-based method. In each panel, the black line depicts the true mode curve, the red, green, and blue lines are three estimated mode curves from the same method that yield EISE_M being the first, second, and third quantiles among the 500 EISE_M's for that method from the simulation, respectively.



(a) (b) (c) (d)

(e) (f) (g) (h)

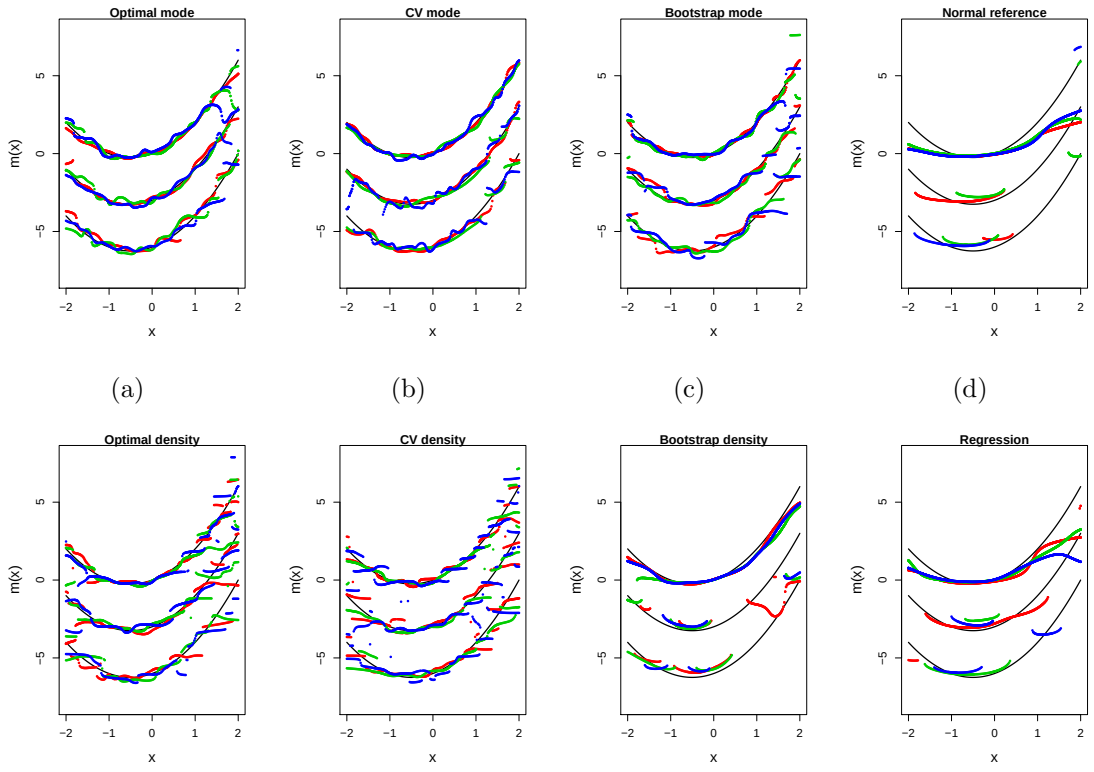
Figure 2: Estimated mode curves resulting from eight choices of bandwidths $\mathbf{h}$ under (C2). The correspondence of the eight panels to eight methods and the correspondence of different colors to different lines are the same as those in Figure 1.



(a) (b) (c) (d)

(e) (f) (g) (h)

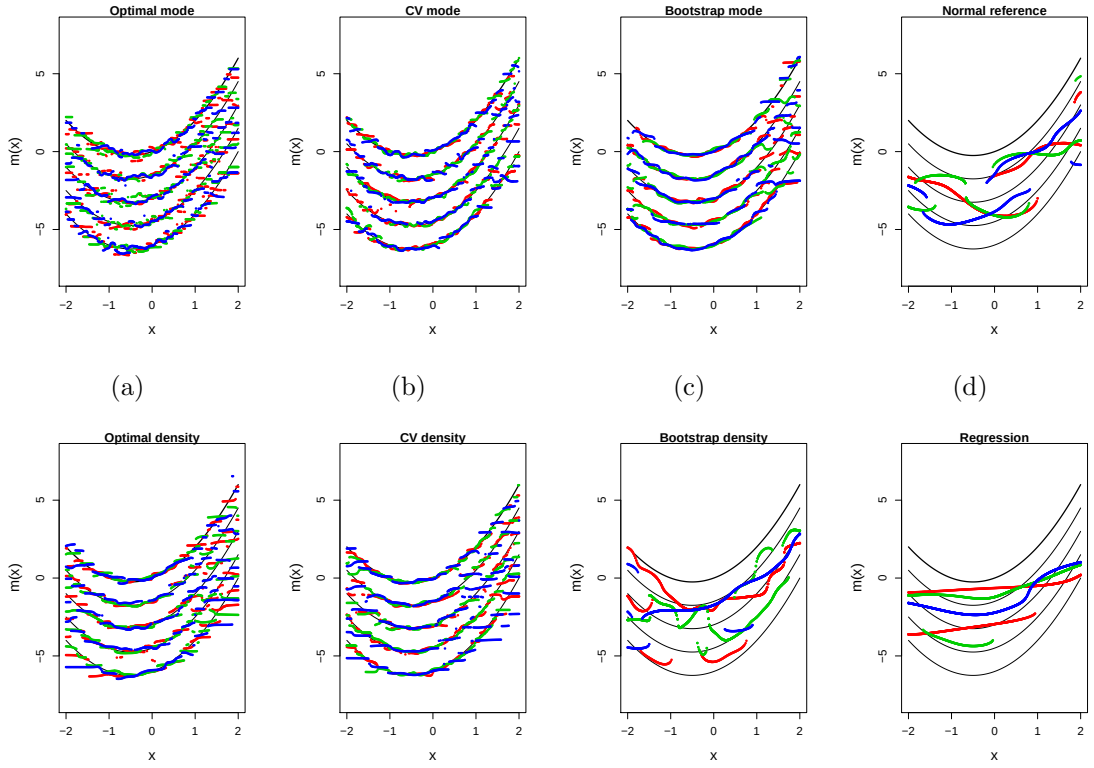
Figure 3: Estimated mode curves resulting from eight choices of bandwidths $\mathbf{h}$ under (C3). The correspondence of the eight panels to eight methods and the correspondence of different colors to different lines are the same as those in Figure 1.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: Estimated mode curves resulting from eight choices of bandwidths $\mathbf{h}$ under (C4). The correspondence of the eight panels to eight methods and the correspondence of different colors to different lines are the same as those in Figure 1.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5: Estimated mode curves resulting from eight choices of bandwidths $\mathbf{h}$ under (C5). The correspondence of the eight panels to eight methods and the correspondence of different colors to different lines are the same as those in Figure 1.

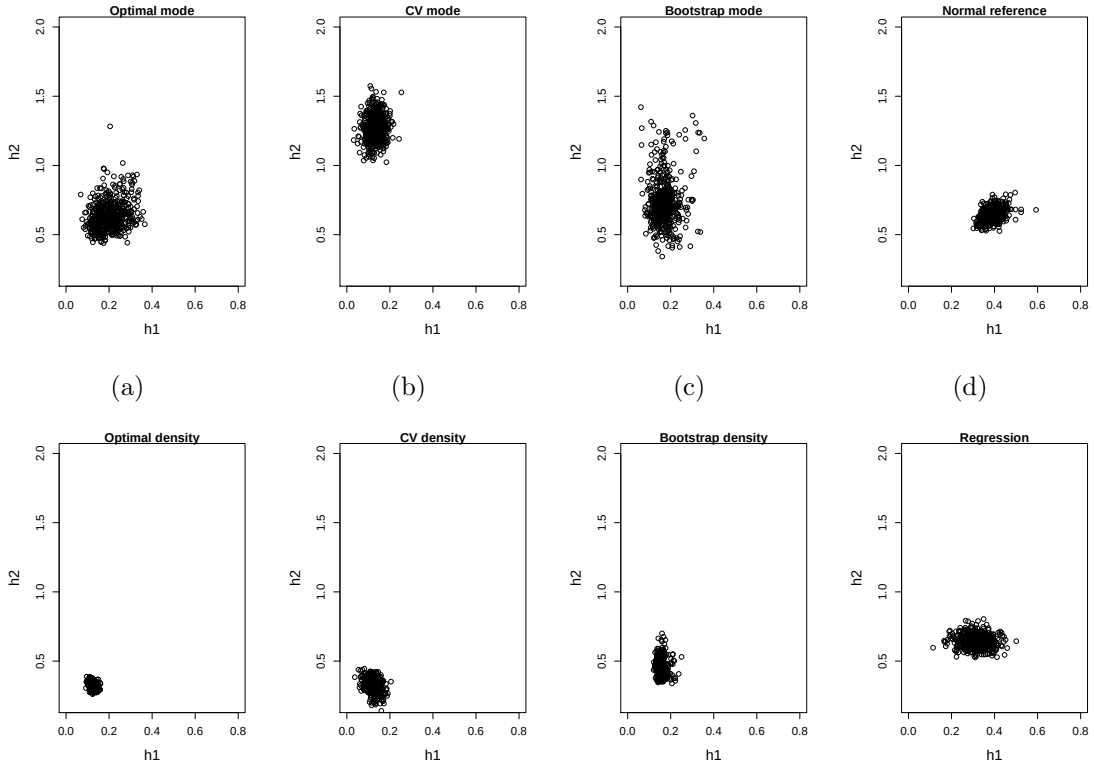


Figure 6: Scatter plots of selected bandwidths across 500 MC replicates under (C1) corresponding to eight ways of bandwidth selection. The correspondence of the eight panels to eight methods is the same as that in Figure 1.

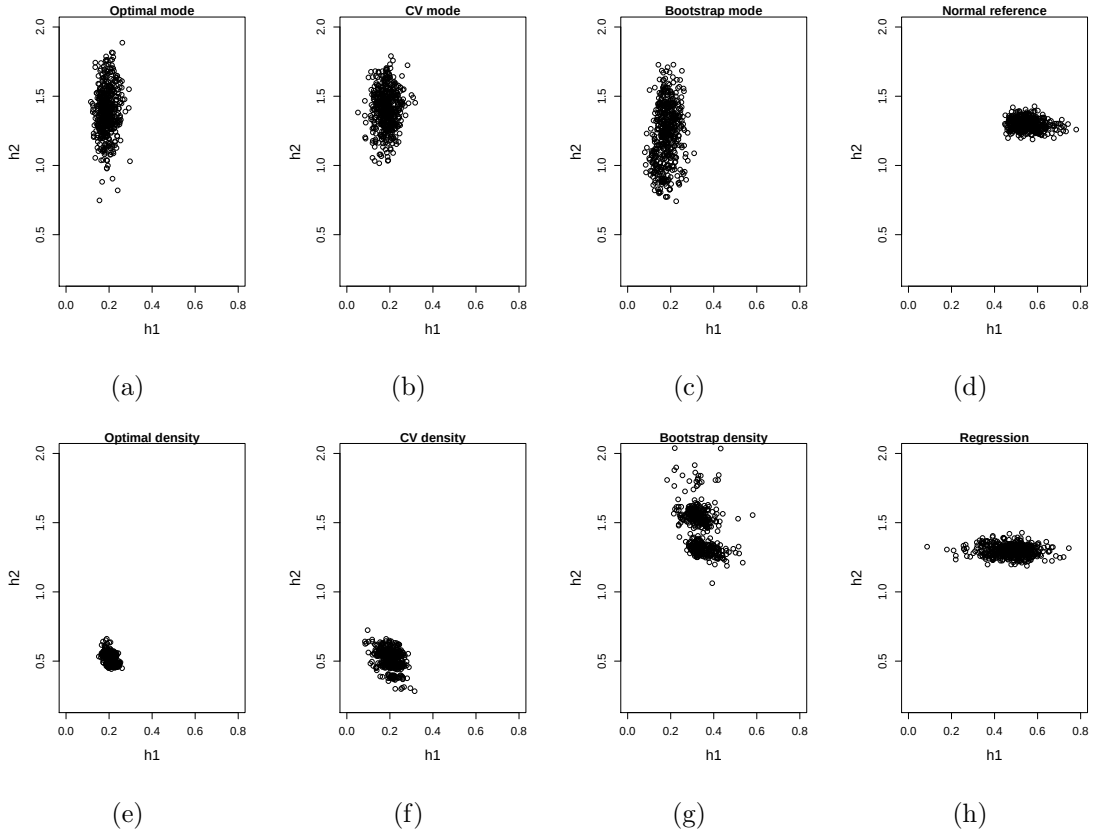


Figure 7: Scatter plots of selected bandwidths across 500 MC replicates under (C2) corresponding to eight ways of bandwidth selection. The correspondence of the eight panels to eight methods is the same as that in Figure 1.

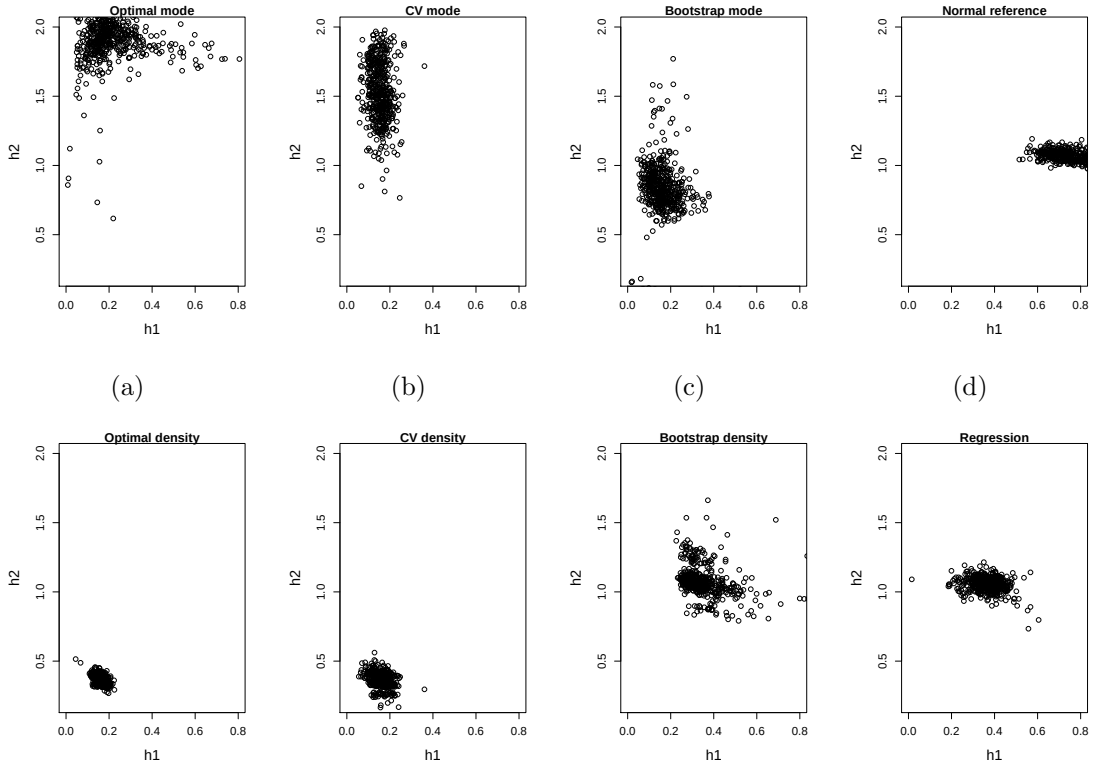


Figure 8: Scatter plots of selected bandwidths across 500 MC replicates under (C3) corresponding to eight ways of bandwidth selection. The correspondence of the eight panels to eight methods is the same as that in Figure 1.

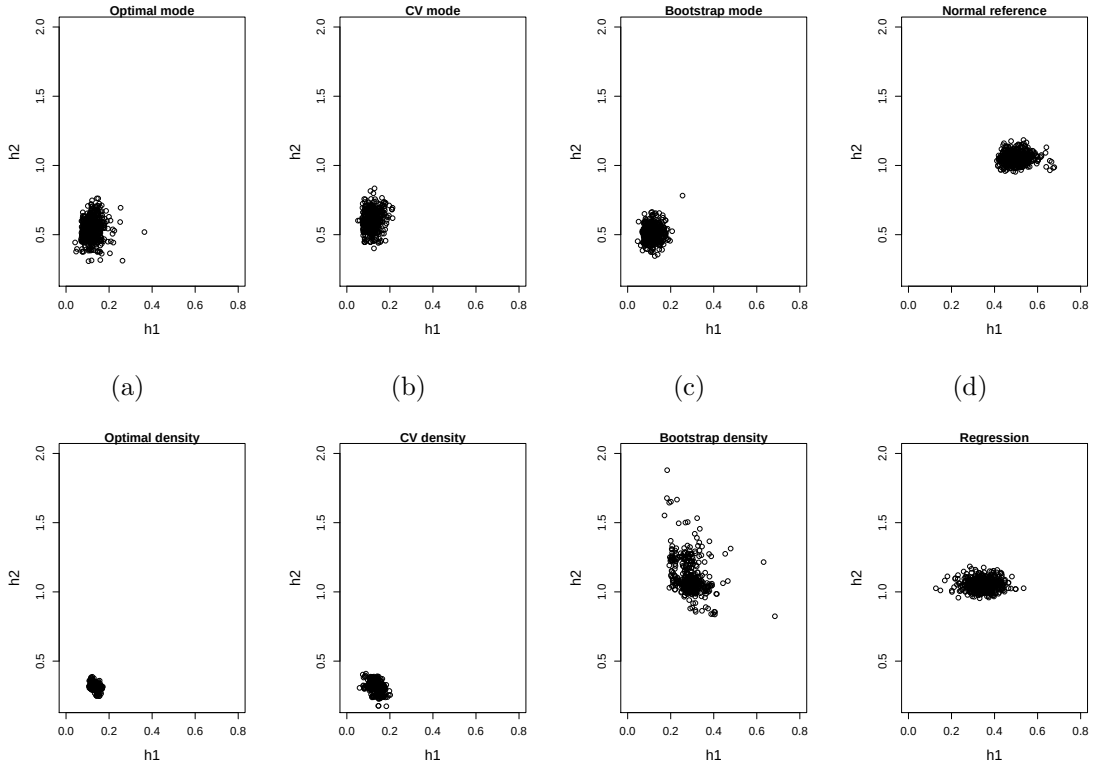


Figure 9: Scatter plots of selected bandwidths across 500 MC replicates under (C4) corresponding to eight ways of bandwidth selection. The correspondence of the eight panels to eight methods is the same as that in Figure 1.

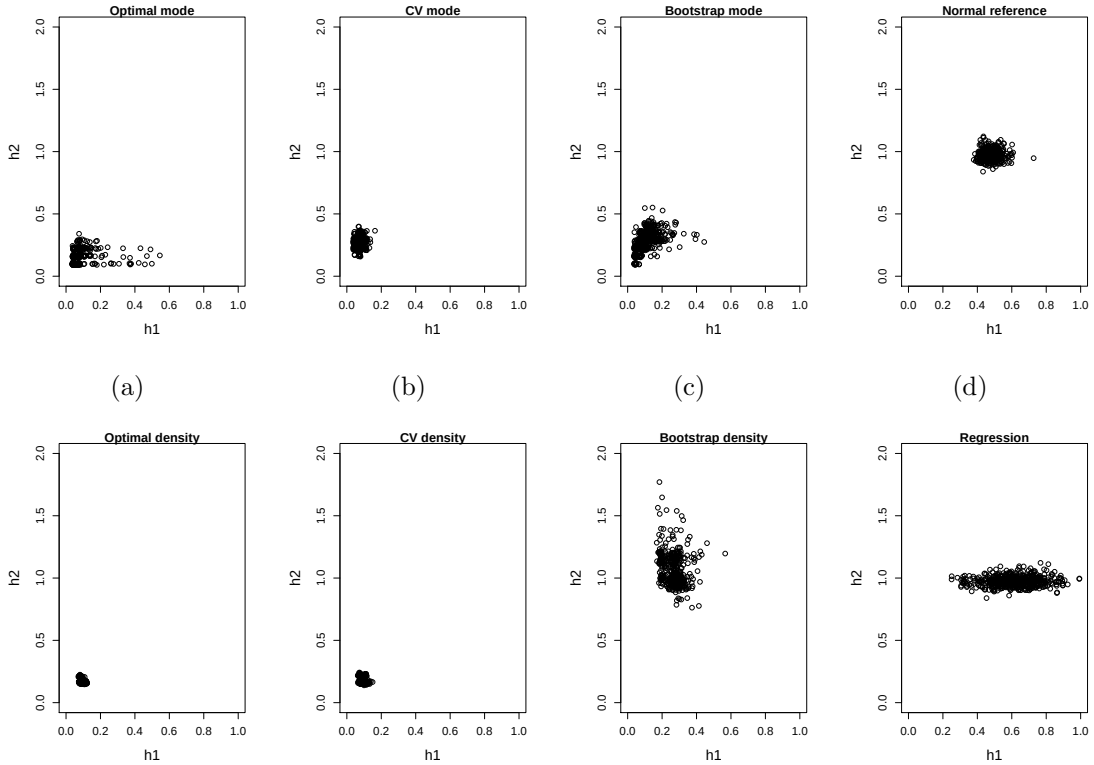


Figure 10: Scatter plots of selected bandwidths across 500 MC replicates under (C5) corresponding to eight ways of bandwidth selection. The correspondence of the eight panels to eight methods is the same as that in Figure 1.

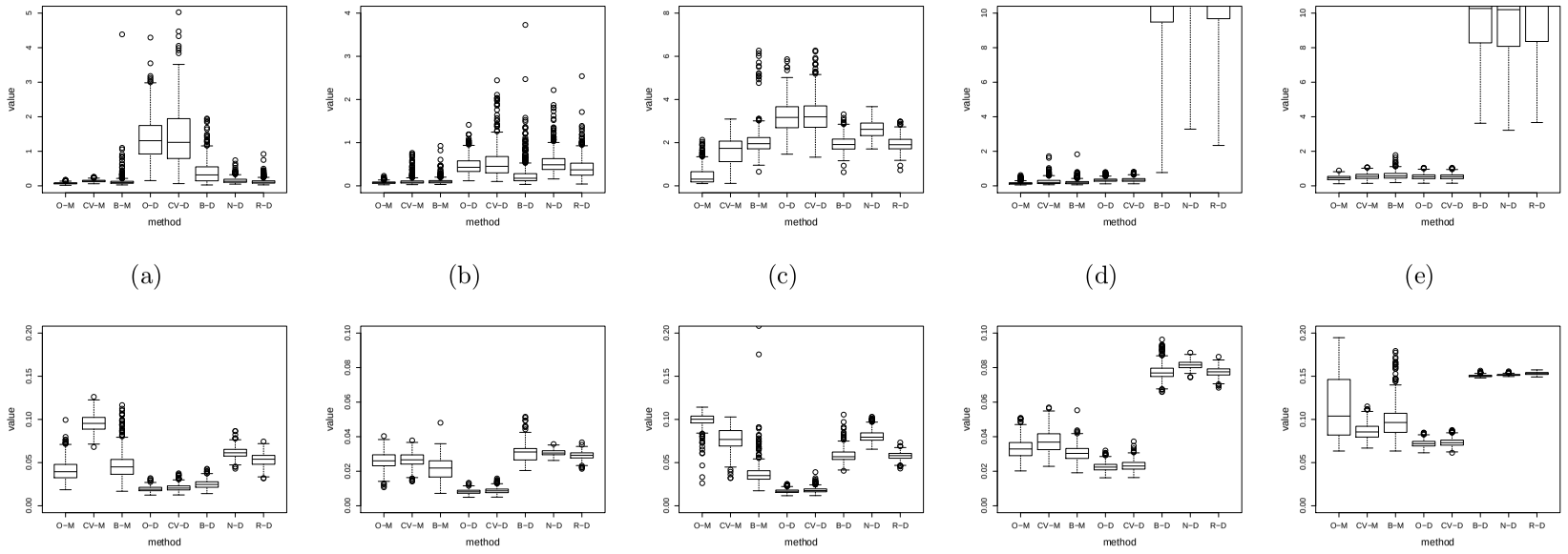


Figure 11: The top row shows boxplots of EISE_M under (C1)–(C5) in panels (a)–(e), respectively. The bottom row shows boxplots of EISE_D under (C1)–(C5) in panels (f)–(j), respectively. In each panel, the eight adopted $\mathbf{h}$ corresponding to the eight boxes are the optimal mode-based bandwidths, $\tilde{\mathbf{h}}_M$ (O-M), the mode-based bandwidths involving CV, $\mathbf{h}_M$ (CV-M), the mode-based bandwidths involving bootstrap, $\mathbf{h}_B^*$ (B-M), the optimal density-based bandwidths, $\tilde{\mathbf{h}}_D$ (O-D), the density-based bandwidths involving CV, $\mathbf{h}_D$ (CV-D), the density-based bandwidths involving bootstrap, $\mathbf{h}_B$ (B-D), the normal reference, $\mathbf{h}_N$ (N-D), and the regression-based bandwidths, $\mathbf{h}_R$ (R-D).

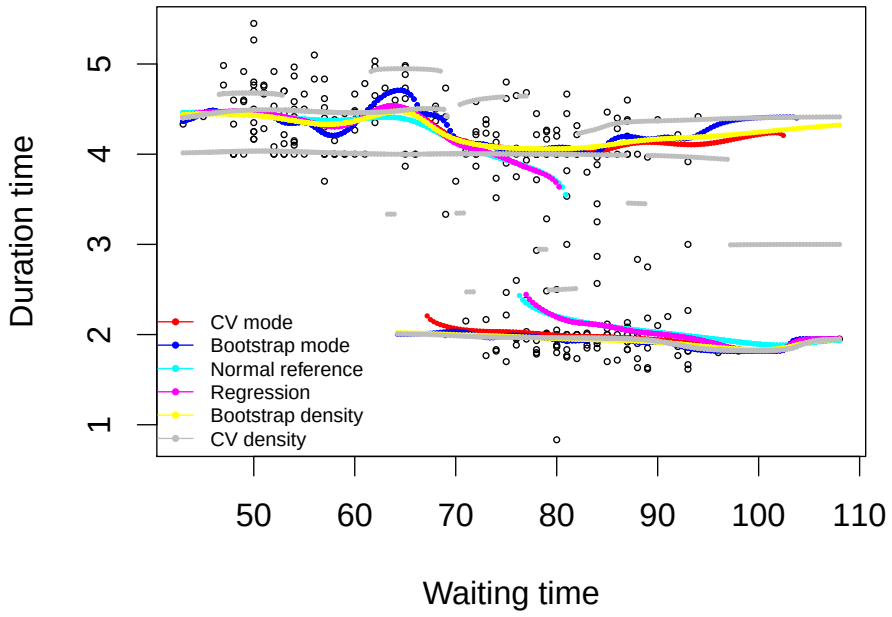


Figure 12: Estimated mode curves resulting from six bandwidth selection methods applied the Old Faithful geyser data (with data points in black circles). The six methods are the mode-based method involving CV (CV mode, red lines), the mode-based method involving bootstrap (Bootstrap mode, blue lines), the normal reference (cyan lines), the regression-based method (pink lines), the density-based method involving bootstrap (Bootstrap density, yellow lines), and the density-based method involving CV (CV density, grey lines).