*Biometrika* (2012), **99**, 1, *pp*. 1–12

*Printed in Great Britain*

Advance Access publication on 31 July 2012

# Jump information criterion for statistical inference in estimating discontinuous curves

BY ZHIMING XIA

*School of Mathematics, Northwest University, Xi'an, Shaanxi, 710127, P. R. China*
statxzm@nwu.edu.cn

AND PEIHUA QIU

*Department of Biostatistics, University of Florida, Gainesville, Florida, 32608, U.S.A.*
pqiu@ufl.edu

## SUMMARY

Nonparametric regression analysis when the regression function is discontinuous has broad applications. Existing methods for estimating a discontinuous regression curve usually assume that the number of jumps in the regression curve is known beforehand, which is unrealistic in certain cases. Although there has been research on estimation of a discontinuous regression curve when the number of jumps is unknown, this problem is still mostly open because such research often requires assumptions on other related quantities such as a known minimum jump size. In this paper, we propose a jump information criterion, which consists of a term measuring the fidelity of the estimated regression curve to the observed data and a penalty related to the number of jumps and jump sizes. Then, the number of jumps can be determined by minimizing our criterion. Theoretical and numerical work shows that our method works well in practice.

*Some key words*: BIC; Jump information criterion; Curve estimation; Jump regression analysis; Kernel smoothing; Penalty.

## 1. INTRODUCTION

Regression curve estimation from noisy data is a fundamental problem with broad applications. In some cases, such curves have jumps and other discontinuities. For instance, the sea-level pressures observed by the Bombay weather station during 1921 and 1990 contain jumps (**?**). The penny thickness data studied in Gijbels & Goderniaux (2004) have a jump as well. In cases when a regression curve has jumps, conventional curve estimation methods such as kernel and spline smoothing methods would not work well at the jump positions. Because jumps are important data structures, proper estimation of their number, positions and sizes is essential in such applications. Estimation of jump regression curves is the focus of this paper.

There has been much discussion in the literature about estimation of jump regression curves (Qiu, 2005), but most existing methods assume that the number of jumps is known (e.g., Antoniadis & Gijbels, 1993; Hall & Titterington, 1992; Müller, 1992; Ma & Yang, 2011; McDonald & Owen, 1986; Müller, 2002; Qiu et al., 1991; Wang, 1995). Several papers discuss cases when the number of jumps is unknown, but require the specification of a minimum jump size (Qiu, 1994) or a significance (Koo, 1997; Qiu & Yandell, 1998; Wu & Chu, 1993; Yin, 1988) beforehand. These methods transfer the difficulty in determining the number of jumps to the difficulty in specifying an alternative parameter.

In the context of change-point detection, some people have considered estimation of the number of change-points using versions of the Bayesian information criterion (Hannart & Naveau, 2012; Zhang & Siegmund, 2007) or other criteria (Yao, 1988; Frick, Munk & Sieling, 2014). However, jump regression is quite different from change-point detection, in that the mean response could be an arbitrary continuous curve between two consecutive jump points in the former, but is constant between two consecutive change-points in the latter.

In this paper, we suggest a jump information criterion for estimating a regression curve with unknown number of jumps. This criterion contains a penalty term for accommodating the unobservable jump structure, different from existing criteria that usually contain penalties related to the complexity of the entire model (Leeb & Pötscher, 2005; Hastie et al., 2008). We demonstrate that the model complexity caused by jumps is asymptotically negligible when the complexity of the entire model is concerned. So, in the jump information criterion, only the model complexity due to jumps is included in the penalty. After the jump structure of the regression function $f$ is estimated, there are two possible ways to estimate $f$, either in design subintervals separated by the estimated jump locations, or by firstly removing the estimated jump part of $f$ from the original data and then estimating the continuous part of $f$ from the resulting new data. Kang et al. (2000) have shown that the second approach can give a more reliable estimate of $f$ because the information contained in the observations in neighboring subintervals can be shared when estimating $f$ around the detected jump positions. This strategy is also used in this paper.

## 2. STATISTICAL METHODS

### 2·1. *Statistical model*

Assume that $n$ observations $\{(Y_i, x_i) : i = 1, \ldots, n\}$ are generated from the model

$$Y_i = f(x_i) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $x_1 < \cdots < x_n$ are fixed design points in the interval $[0, 1]$, and $\{\varepsilon_i\}$ are independent and identically distributed random errors with mean 0 and variance $\sigma^2$. In (1), it is further assumed that the regression function $f$ is discontinuous with a continuous part $f_C$ and a jump part $f_J$. If $\{s_j : j = 1, \ldots, m_0\}$ are $m_0$ jump positions of $f$ with corresponding jump magnitudes $\{d_j : j = 1, \ldots, m_0\}$, then

$$f(x) = f_C(x) + f_J(x) = f_C(x) + \sum_{j=1}^{m_0} d_j I_{\{x > s_j\}}, \quad x \in [0, 1]. \tag{2}$$

In (2), the number of jumps $m_0$, jump locations $s_j$'s and jump sizes $d_j$'s are all assumed unknown. For convenience of theoretical inference, the design points are assumed to be equally-spaced with $x_i = i/n$, for $i = 0, 1, \ldots, n$, and the random errors $\{\varepsilon_i\}$ are assumed to have equal variance. All methodologies developed in this paper can be applied to cases with unequally-spaced design points, after some minor modifications. More specifically, our methods will work well in cases when there is a continuous design density function $q(x)$, such that the unequally-spaced design points can be expressed as $\int_0^{x_i} q(u)\, du = i/n$. They will also work well in cases when the error variances are unequal but $\sum_{i=1}^{n} \mathrm{var}(Y_i)/n \to \sigma^2$. Without loss of generality, we assume that $|d_1| > \cdots > |d_{m_0}|$.

### 2·2. *Model estimation when the number of jumps is assumed known*

As mentioned in Section 1, most existing methods for estimating jump regression curves assume that the number of jumps or its upper bound is known to be $m$, say. Such an assumption can

result in various problems. For instance, the case of $m < m_0$ could lead to a non-consistent estimate of $f$ near the $m_0 - m$ jump positions that are not detected, and the case of $m \geq m_0$ could lead to detection of $m - m_0$ spurious jumps. In this subsection, we develop a methodology to detect the jump structure and to estimate $f$ as well when $m$ is assumed known.

In our methodology, the local linear kernel smoothing procedure is used as a building block, which requires a bandwidth and a kernel function (Loader, 1999). To obtain a reasonably good estimate, the following two assumptions are needed.

*Assumption* 1. The bandwidth $h_n > 0$ satisfies $\log n/(nh_n^3) = o(1)$ and $nh_n^5/\log n = o(1)$.

*Assumption* 2. Let $K_c(x) = K(x)$ be a bounded symmetric density kernel function with support $[-1/2, 1/2]$ which is uniformly Lipschitz-1 continuous. The two one-sided kernel functions are defined to be $K_l(x) = K(x)I_{\{x\in[-1/2,0)\}}$ and $K_r(x) = K(x)I_{\{x\in[0,1/2]\}}$.

Then, for a given $x \in [h_n/2, 1 - h_n/2]$, we estimate $f(x)$ in the neighborhood $[x - h_n, x + h_n]$ and its two one-sided parts $[x - h_n, x)$ and $[x, x + h_n]$, respectively, by

$$\widehat{a}_j(x) = \sum_{i=1}^{n} Y_i K_j^* \left( \frac{x_i - x}{h_n} \right) \quad (j = c, l, r), \tag{3}$$

where $K_j^*\{(x_i - x)/h_n\} = K_j\{(x_i - x)/h_n\}\{w_{j,2} - w_{j,1}(x_i - x)\}/(w_{j,0}w_{j,2} - w_{j,1}^2)$, $w_{j,k} = \sum_{i=1}^{n}(x_i - x)^k K_j\{(x_i - x)/h_n\}$, for $j = c, l, r$ and $k = 0, 1, 2$. Based on $\widehat{a}_l(x)$ and $\widehat{a}_r(x)$, we define $M_n(x) = \widehat{a}_r(x) - \widehat{a}_l(x)$, which contains information about the jump point near $x$. As pointed out by Qiu (1994), $M_n(x) \approx 0$ if there are no jumps in $[x - h_n/2, x + h_n/2]$, and $M_n(x)$ is relatively large if there exists a jump around $x$. More specifically, if $x$ is close to a jump point $s_j$, say $x = s_j + \tau h_n$, for $\tau \in [-1/2, 1/2]$, we shall show in Theorem 1 that $|M_n(x)|$ converges to $|d_j g(\tau)|$, where $|g(\tau)|$ is non-negative in interval $[-1/2, 1/2]$ and reaches its maximum value 1 when $\tau = 0$. The property discussed above is summarized in Theorem 1, in which the following notation is used:

$$D = [h_n/2, 1 - h_n/2], \quad D_{\delta,j} = [s_j - \delta, s_j + \delta],$$
$$D_{\delta,m_1\to m_2} = \bigcup_{j=m_1}^{m_2} D_{\delta,j}, \; \overline{D}_{\delta,m_1\to m_2} = D\backslash D_{\delta,m_1\to m_2}, \qquad 1 \leq m_1 \leq m_2 \leq m_0.$$

Also, $\|f\|_{L^\infty(D)}$ denotes $\max_{x\in D} |f(x)|$.

THEOREM 1. *Assume that $f$ is second-order differentiable and $f''$ is uniformly bounded on $D$ except at the jump points $\{s_j : j = 1, \ldots, m_0\}$ at which $f$ has bounded left and right second-order derivatives. Then, under Assumptions* 1 – 2*, we have*

$$\left(\frac{nh_n}{\log n}\right)^{1/2} \|M_n\|_{L^\infty(\overline{D}_{h_n/2,1\to m_0})} = \mathcal{O}(1), \quad \left(\frac{nh_n}{\log n}\right)^{1/2} \left\|M'_{n,j}\right\|_{L^\infty[-1/2,1/2]} = \mathcal{O}(1),$$

*where both equations hold almost surely, $M'_{n,j}(\tau) = M_n(s_j + \tau h_n) - d_j g(\tau)$ $(j = 1, \ldots, m_0)$, and $g(\tau) = I_{\tau\in[-1/2,0)} \int_{-\tau}^{1/2} K_r^*(u)\mathrm{d}u + I_{\tau\in[0,1/2]} \int_{-1/2}^{-\tau} K_l^*(u)\mathrm{d}u$.*

Methods (Müller, 1992; Antoch, Grégoire & Hušková, 2007; Grégoire & Hamrouni, 2002) for estimating the jump location in cases with one jump point can be extended. From Theorem 1, the jump detection criterion $|M_n(x)|$ is small in continuity regions $\overline{D}_{h_n/2,1\to m_0}$ and large in regions around the true jump points. Therefore, if we assume that $|d_1| > \cdots > |d_{m_0}|$, for convenience of

discussion, then $s_1$ is the asymptotic global maximizer of $|M_n(x)|$ in $D$, and a natural estimator is

$$\widehat{s}_1(m) = \arg\max_{x \in D} |M_n(x)|. \tag{4}$$

Similarly, $s_2$ is the asymptotic global maximizer of $|M_n(x)|$ in $D \backslash \widehat{D}_{\epsilon h_n,1}$, where $\widehat{D}_{\epsilon h_n,1} = [\widehat{s}_1(m) - \epsilon h_n, \widehat{s}_1(m) + \epsilon h_n]$ and $\epsilon > 1/2$ is a positive number, and can be estimated by an estimator defined similarly to (4). Generally, the estimator of $s_j$ can be defined by

$$\widehat{s}_j(m) = \arg \max_{x \in D \backslash \bigcup_{i=1}^{j-1} \widehat{D}_{\epsilon h_n,i}} |M_n(x)| \quad (j = 1, \ldots, m), \tag{5}$$

where $\widehat{D}_{\epsilon h_n,i} = [\widehat{s}_i(m) - \epsilon h_n, \widehat{s}_i(m) + \epsilon h_n]$. The corresponding estimators of $d_j$ can be defined by $\widehat{d}_j(m) = M_n\{\widehat{s}_j(m)\}$ $(j = 1, \ldots, m)$. Furthermore, the jump part $f_J(x)$ of the regression curve $f(x)$ can be estimated by $\widehat{f}_{J,m}(x) = \sum_{j=1}^{m} \widehat{d}_j(m) I_{\{x > \widehat{s}_j(m)\}}$.

After the jump part is estimated, we can define new observations $\{Y_{i,m} = Y_i - \widehat{f}_{J,m}(x_i) : i = 1, \ldots, n\}$, and estimate the continuity part $f_C(x)$ by local linear kernel smoothing in the entire design interval $[0, 1]$ from the new observations. The local linear kernel estimator $\widehat{f}_{C,m}(x)$ is defined in the same way as $\widehat{a}_c(x)$ after replacing $Y_i$ by $Y_{i,m}$. Finally, the regression function $f(x)$ can be estimated by

$$\widehat{f}_m(x) = \widehat{f}_{C,m}(x) + \widehat{f}_{J,m}(x). \tag{6}$$

Let $\widehat{Y} = (\widehat{f}_m(x_1), \ldots, \widehat{f}_m(x_n))^T, \widehat{Y}_C = (\widehat{f}_{C,m}(x_1), \ldots, \widehat{f}_{C,m}(x_n))^T, \widehat{Y}_J = (\widehat{f}_{J,m}(x_1), \ldots, \widehat{f}_{J,m}(x_n))^T$. Then, by (6), we have

$$\widehat{Y} = \widehat{Y}_C + \widehat{Y}_J = H_C (I - H_J) Y + H_J Y = HY, \tag{7}$$

where $H_C$ and $H_J$ are the hat matrices of $\widehat{Y}_C$ and $\widehat{Y}_J$, respectively, in terms of $Y = (Y_1, \ldots, Y_n)^T$, and $H = H_C(I - H_J) + H_J$, $H_J$ depends on the estimators $\widehat{s}_j(m)$ $(j = 1, \ldots, m)$. By Theorem 2, these estimators converge almost surely to the true jump locations $s_j$. Regarding the hat matrix $H$, we have the following result.

PROPOSITION 1. *Under Assumptions* 1 *and* 2*, we have*

$$\text{tr}\,(H) = \frac{K_c^*(0)}{h_n} + m + o\left(\frac{1}{h_n}\right), \tag{8}$$

*where* $\text{tr}(H)$ *denotes trace.*

Stein's unbiased risk estimation theory (Stein, 1981) provides a precise definition of the degrees of freedom of an estimator and is often used to measure the model complexity. For an approximate linear smoother like (7), its degree of freedom is shown to be $\text{tr}\,(H)$ approximately (Efron, 2004). Because the first term on the right-hand-side of (8) tends to infinity, the impact of the second term $m$ on $\text{tr}\,(H)$ is negligible, so the overall complexity of the estimated regression model does not much depend on the number of jumps, as mentioned in Section 1.

The theorem below builds the uniform strong consistency of the estimators defined in (4)–(6).

THEOREM 2. *Under the assumptions of Theorem* 1*, if* $m < m_0$*, then*

$$\left(\frac{n}{h_n \log n}\right)^{1/2} |\widehat{s}_j(m) - s_j| = \mathcal{O}(1) \quad (j = 1, \ldots, m),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left|\widehat{d}_j(m) - d_j\right| = \mathcal{O}(1) \quad (j = 1, \ldots, m),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left\|\widehat{f}_m - f\right\|_{L^\infty(\overline{D}_{\delta_n, 1\to m} \cap \overline{D}_{h_n/2, m+1\to m_0})} = \mathcal{O}(1),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left\|\widehat{f}_m(s_j + \tau h_n) - f(s_j + \tau h_n) - d_j h_c(\tau)\right\|_{L^\infty[-1/2, 1/2]} = \mathcal{O}(1)$$
$$(j = m + 1, \ldots, m_0),$$

*where all equations hold almost surely,* $h_c(\tau) = I_{\{\tau \in [-1/2, 0)\}} \int_{-\tau}^{1/2} K^*_{c,0}(u)\mathrm{d}u - I_{\{\tau \in [0, 1/2]\}} \int_{-1/2}^{-\tau} K^*_{c,0}(u)\mathrm{d}u$*,* $\delta_n = \{(h_n \log n)/n\}^{1/2-\delta}$*, and* $\delta \in (0, 1/2)$*. If* $m \geq m_0$*, then*

$$\left(\frac{n}{h_n \log n}\right)^{1/2} |\widehat{s}_j(m) - s_j| = \mathcal{O}(1) \quad (j = 1, \ldots, m_0),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left|\widehat{d}_j(m) - d_j\right| = \mathcal{O}(1) \quad (j = 1, \ldots, m_0),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left|\widehat{d}_j(m)\right| = \mathcal{O}(1) \quad (j = m_0 + 1, \ldots, m),$$
$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left\|\widehat{f}_m - f\right\|_{L^\infty(\overline{D}_{\delta_n, 1\to m_0})} = \mathcal{O}(1),$$

*where all equations hold almost surely.*

Theorem 2 shows that if the assumed upper bound of the number of jumps is smaller than the true number, our method can estimate the largest $m$ jumps consistently. But, it will miss the remaining $m_0 - m$ jumps, which would result in an asymptotic bias of the size $d_j h_c(\tau)$ near the jump points $s_j$ $(j = m_0 + 1, \ldots, m)$, when estimating $f$. If the assumed number of jumps is larger than or equal to the true number, Theorem 2 says that all of the $m_0$ true jumps can be estimated consistently by our approach. But, we artificially create $m - m_0$ more jumps whose estimated jump sizes converge to zero uniformly in the rate of $\{\log n/(nh_n)\}^{1/2}$. This rate will play a key role in constructing our jump information criterion.

### 2·3. *Jump information criterion*

Theorem 2 shows that there will be problems if the number of jumps or its upper bound is assumed known. To estimate the number of jumps properly based on the observed data, let us check Theorem 2 more carefully. Theorem 2 says that the curve estimator $\widehat{f}_m(x)$ is consistent when $m \geq m_0$, and it has relatively large biases at certain jump points when $m < m_0$. These properties of $\widehat{f}_m(x)$ are reflected in the sum of squares of residuals

$$\mathrm{SSR}(m) = \sum_{i=1}^{n} \left\{Y_i - \widehat{f}_m(x_i)\right\}^2.$$

When $m < m_0$, it can be checked that the sum of squares of residuals around a jump point missed by the jump detection procedure, say the $j$th jump, is

$$\sum_{x_i \in D_{h_n/2,j}} \left\{ Y_i - \widehat{f}_m(x_i) \right\}^2 = \mathcal{O}\left(nh_n\right).$$

When $m > m_0$, the sum of squares of residuals around a spurious jump is about $\mathcal{O}\{(nh_n \log n)^{1/2}\}$ which is much smaller than $\mathcal{O}\left(nh_n\right)$. Although both quantities seem negligible compared to $\text{SSR}(m) = \mathcal{O}(n)$, their difference needs to be taken into account when determining the number of jumps. Another important factor is related to the estimators of jump sizes. When $m \leq m_0$, all $m$ estimators of jump sizes converge to non-zero constants. In the case when $m > m_0$, there are $m - m_0$ spurious jumps and the estimators of their jump sizes will converge to zero in the rate of $\{\log n/(nh_n)\}^{1/2}$. Based on these two considerations, our jump information criterion is

$$\text{JIC}(m) = n \log\{\text{SSR}(m)/n\} + P(n) \sum_{j=1}^{m} \frac{1}{\left|\widehat{d}_j(m)\right|^{\gamma}}, \tag{9}$$

where $\gamma \geq 0$ is a tuning parameter, and $P(n)$ is an adjustment factor.

The criterion $\text{JIC}(m)$ in (9) has two terms. The first term measures the distance from the estimated regression curve to the observed data. It can be checked that this term is a decreasing function of $m$ with the rate $\mathcal{O}(nh_n)$ when $m \leq m_0$, and the rate becomes $\mathcal{O}\{(nh_n \log n)^{1/2}\}$ when $m > m_0$. The second term is a penalty consisting of two quantities $P(n)$ and $\sum_{j=1}^{m} |\widehat{d}_j(m)|^{-\gamma}$. The quantity $\sum_{j=1}^{m} |\widehat{d}_j(m)|^{-\gamma}$ is used as a panelty to take into account both the number of jumps and the jump sizes of the jump part of model (1). It can be checked that this is an increasing function of $m$, and the rate of increase is faster when $m > m_0$, compared to when $m \leq m_0$. The adjustment factor $P(n)$ is used to guarantee that $\text{JIC}(m)$ is decreasing in $m$ when $m \leq m_0$ and increasing in $m$ when $m > m_0$. It can be checked that such properties are guaranteed when

$$\left(\frac{nh_n}{\log n}\right)^{-\gamma/2} nh_n^2 < P(n) < nh_n. \tag{10}$$

By minimizing $\text{JIC}(m)$, we can obtain an estimator $\widehat{m}$ of $m$, and then we can obtain estimators of the jump locations and jump sizes. More specifically, we define

$$\widehat{m} = \arg\min_{m \geq 0} \text{JIC}(m), \tag{11}$$

$$\widehat{s}_j = \widehat{s}_j(\widehat{m}), \tag{12}$$

$$\widehat{d}_j = \widehat{d}_j(\widehat{m}), \tag{13}$$

$$\widehat{f}(x) = \widehat{f}_{\widehat{m}}(x). \tag{14}$$

From Theorem 3, in the case when $P(n) = (nh_n/\log n)^{-\gamma/2}\, nh_n$, which is in the range specified by (10), the estimator $\widehat{m}$ will converge to $m_0$ at the optimal rate of $h_n + \{\log n/(nh_n)\}^{(\gamma+1)/2}$. When $P(n) > (nh_n/\log n)^{-\gamma/2}\, nh_n$, the distribution of $\widehat{m}$ is skewed to the right, and it is skewed to the left when $P(n) < (nh_n/\log n)^{-\gamma/2}\, nh_n$. In the latter two cases, the optimal convergence rate of $\widehat{m}$ would not be reached any more.

As mentioned in Section 1, in the literature on change-point detection, several papers proposed versions of the Bayesian information criterion for determining the number of change points using the Bayes factor (Schwarz, 1978; Zhang & Siegmund, 2007; Hannart & Naveau, 2012). Under

several assumptions, including that the continuous part $f_C(x)$ of $f(x)$ is known, we derive the following criterion in the Supplementary Material,

$$\text{BIC}(m) = n\log\{\text{SSR}(m)/n\} + m\log(nh_n). \tag{15}$$

This is a special case of (9) when $\gamma = 0$ and $P(n) = \log(nh_n)$, and this selection of $P(n)$ is beyond the range in (10). Also, the penalty term in (15) depends on $m$, but not on the jump sizes, making it less sensitive to the jump structure of the observed data, compared to JIC$(m)$. This intuition is confirmed by the numerical results in Section 4.

## 3. Statistical Properties

In this section, we study the statistical properties of the estimators defined in (11)–(14) when the number of jumps is estimated by the proposed jump information criterion.

THEOREM 3. *If all assumptions of Theorem* 1 *hold, then almost surely*

$$\frac{K_n}{R_n}\,|\widehat{m} - m_0| = \mathcal{O}(1), \tag{16}$$

*where* $K_n = \min\left\{nh_n, P(n)(nh_n/\log n)^{\gamma/2}\right\}$, *and* $R_n = nh_n^2 + P(n)\{\log n/(nh_n)\}^{1/2}$.

Theorem 3 shows that it is necessary to have $K_n/R_n \to \infty$ in order for $\widehat{m}$ to be consistent. The condition (10) is mainly derived from the condition that $K_n/R_n \to \infty$. In such cases, we can conclude immediately from (16) that

$$\text{pr}(\lim_{n\to\infty} \widehat{m} = m_0) = 1, \tag{17}$$

and the convergence reaches the optimal rate of $O(h_n)$ when $P(n) = (nh_n/\log n)^{-\gamma/2}\, nh_n$. From (17), $\widehat{m} = m_0$ almost surely when the sample size $n$ is large enough. So, in the next theorem about the almost sure consistency of the estimators $\{\widehat{s}_j : j = 1, \ldots, \widehat{m}\}$ and $\{\widehat{d}_j : j = 1, \ldots, \widehat{m}\}$, the index $j$ will be specified to be one of $\{1, \ldots, m_0\}$, without loss of generality.

THEOREM 4. *Under the conditions of Theorem* 3,

$$\left(\frac{n}{h_n \log n}\right)^{1/2} |\widehat{s}_j - s_j| = \mathcal{O}(1) \quad (j = 1, \ldots, m_0),$$

$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left|\widehat{d}_j - d_j\right| = \mathcal{O}(1) \quad (j = 1, \ldots, m_0),$$

$$\left(\frac{nh_n}{\log n}\right)^{1/2} \left\|\widehat{f} - f\right\|_{L^\infty(\overline{D}_{\delta_n, 1\to m_0})} = \mathcal{O}(1),$$

*where all equations hold almost surely, and* $\delta_n$ *is defined in Theorem* 2.

Theorem 4 shows that estimators of the jump locations, jump sizes and the whole discontinuous curve converge to their true values, as if the number of jumps is known in advance.

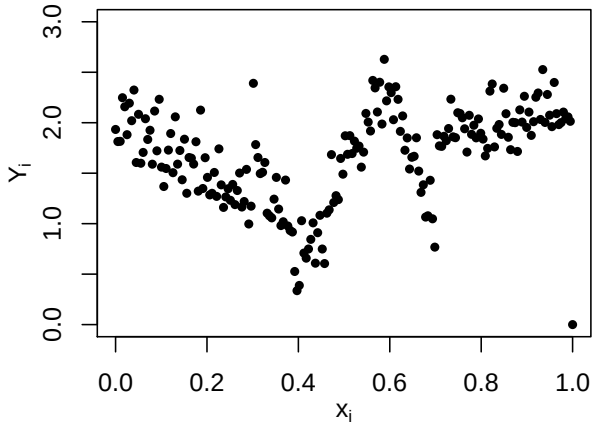


Fig. 1. Graph for simulation data based on Model1.

## 4. NUMERICAL STUDY

### 4·1. *Monte Carlo simulation examples*

Assume that $n$ observations $\{(x_i, Y_i) : i = 1, \ldots, n\}$ are generated from the model (1). The regression function is assumed to be

$$f(x) = \begin{cases} -3x + 2, & [0, 0.3), \\ -3x + 3 - \sin\{(x - 0.3)\pi/0.2\}, & [0.3, 0.7), \\ x/2 + 1.55, & [0.7, 1), \end{cases}$$

which has $m_0 = 2$ jump points. Because the central part of $f$ is steep around the two jump points, the jumps are difficult to detect. See Gijbels et al. (2007) for a related discussion. We then consider the following three cases.

Case 1: The design points $\{x_i = i/n : i = 1, \ldots, n\}$ are equally spaced in $[0, 1]$, and the random errors $\{\varepsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed with distribution $N(0, 0.2^2)$.

Case 2: The design points $\{x_i : i = 1, \ldots, n\}$ are unequally spaced in $[0, 1]$ with a smooth design density $q(x) = 0.6I_{\{x \le 0.1\}} + 1.1I_{\{0.1 < x < 0.9\}} + 0.6I_{\{x \ge 0.9\}}$, and the random errors $\{\varepsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed with distribution $0.2t_{10}$.

Case 3: The design points $\{x_i : i = 1, \ldots, n\}$ are equally spaced in $[0, 1]$, and the random errors $\{\varepsilon_i : i = 1, \ldots, n\}$ follow the autoregressive model $\varepsilon_i = 0.1\varepsilon_{i-1} + u_i$ $(i = 1, \ldots, n)$, where $\varepsilon_0 = 0$ and $\{u_i : i = 1, \ldots, n\}$ are independent and identically distributed with distribution $N(0, 0.2^2)$.

Case 2 considers a random design setup, and Case 3 considers a scenario with dependent observations. They do not satisfy the assumptions of model (1), and are considered here to study the robustness of the proposed method. One realization of $n = 200$ observations in Case 1 when $\sigma = 0.2$ is shown in Figure 1. Both jumps, especially the left one, are indeed quite difficult to detect visually.

Next, we use the criterion JIC$(m)$ to determine the number of jumps. In the simulation, we use the Epanechnikov kernel functions $K_r(x) = 0.75(1 - x^2)$, for $x \in [0, 1/2]$, and $K_l(x) = K_r(-x)$, which satisfy Asssumption 2. The bandwidth is chosen to be $h_n = 0.3n^{-1/5}$. The factor $n^{-1/5}$ in $h_n$ is mainly due to Assumption 1. The constant $0.3$ in $h_n$ is chosen using the following consideration: this constant should be chosen relatively small so that we can include multiple jump points in the simulation and they will not have overlapping neighbourhoods, but it should not be too small, or there will be too few observations in each neighborhood for data

Table 1. *Percentages of 1000 replicated simulations for which the $\widehat{m}$ values obtained by each method are equal to, smaller than, and larger than the true number of jumps $m_0$.*

| Method | $\widehat{m}$ | Case 1 | | | Case 2 | | | Case 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n$ =200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| JIC($m$) | $> m_0$ | 2.8 | 0.0 | 0.0 | 7.0 | 0.4 | 0.0 | 6.0 | 0.6 | 0.2 |
| (small penalty) | $= m_0$ | 96.4 | 100.0 | 100.0 | 93.0 | 99.6 | 100.0 | 93.0 | 99.4 | 99.8 |
| | $< m_0$ | 0.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| JIC($m$) | $> m_0$ | 0.0 | 0.0 | 0.0 | 1.2 | 0.0 | 0.0 | 0.8 | 0.0 | 0.0 |
| (moderate penalty) | $= m_0$ | 96.6 | 100.0 | 100.0 | 95.6 | 100.0 | 100.0 | 93.4 | 100.0 | 100.0 |
| | $< m_0$ | 3.4 | 0.0 | 0.0 | 3.2 | 0.0 | 0.0 | 5.8 | 0.0 | 0.0 |
| JIC($m$) | $> m_0$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| (large penalty) | $= m_0$ | 54.0 | 94.4 | 100.0 | 58.0 | 90.6 | 99.2 | 51.0 | 90.6 | 99.8 |
| | $< m_0$ | 46.0 | 5.6 | 0.0 | 42.0 | 9.4 | 0.8 | 49.0 | 9.4 | 0.2 |
| BIC($m$) | $> m_0$ | 31.8 | 10.8 | 4.0 | 42.8 | 15.8 | 9.4 | 38.8 | 25.8 | 13.8 |
| | $= m_0$ | 68.2 | 89.2 | 96.0 | 57.2 | 84.2 | 90.6 | 61.2 | 74.2 | 86.2 |
| | $< m_0$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| wavelets | $> m_0$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | $= m_0$ | 16.0 | 88.0 | 100.0 | 37.0 | 86.0 | 100.0 | 21.0 | 84.0 | 100.0 |
| | $< m_0$ | 84.0 | 12.0 | 0.0 | 63.0 | 14.0 | 0.0 | 79.0 | 16.0 | 0.0 |

smoothing. In JIC($m$), we fix $\gamma = 1$, and choose $P(n)$ to be $\left\{nh_n^2(\log n)^2\right\}^{1/2}$ as a small penalty, $(nh_n \log n)^{1/2}$ as a moderate penalty, or $(nh_n)^{1/2} \log n$ as a large penalty. The selected moderate penalty is also the one that results in the optimal convergence rate of $\widehat{m}$, discussed immediately after Theorem 3. Apart from these three choices of $P(n)$, we also consider the case when $\gamma = 0$ and $P(n) = \log(nh_n)$, which corresponds to BIC($m$) in (15). The wavelet method proposed by Wang (1995) using the Haar wavelet is also considered, with the true value of $\sigma$ used in computing the threshold value for jump detection. The sample size $n$ is chosen to be $200, 500$, or $1000$, and the simulation is repeated 1000 times.

The results are summarized in Table 1. For the proposed method, the results show that the frequency of $\widehat{m} = m_0$ tends to be larger when $n$ increases in all cases considered, which confirms Theorem 3. The distribution of $\widehat{m}$ is skewed to the right when $P(n)$ is chosen to be the small penalty, skewed to the left if $P(n)$ is chosen to be the large penalty, and more symmetric when $P(n)$ is chosen to be the moderate penalty. Furthermore, the frequency of $\widehat{m} = m_0$ is the largest when $P(n)$ is chosen to be the moderate penalty, which confirms the statement made immediately after Theorem 3 about the optimal convergence rate of $\widehat{m}$. Based on these results, we suggest using the moderate penalty for $P(n)$ in practice. By comparing the results in Cases 1–3, it seems that the proposed method is robust to the assumptions of independent observations and fixed design points. Theoretical study of these properties is left to future research. As a comparison, both the BIC and wavelet methods are not as effective as the proposed method with the moderate penalty in all cases considered.

Table 1 concerns the estimated number of jumps. Next, we describe how close the estimated jumps are to the true jumps. Let $\widehat{S} = \{\widehat{s}_1, \ldots, \widehat{s}_{\widehat{m}}\}$ and $S = \{s_1, \ldots, s_{m_0}\}$ denote the sets of

Table 2. *Averaged Hausdorff distances between $\widehat{S}$ and $S$ and the corresponding standard errors (in parentheses) based on 1000 replicated simulations. All numbers are in the unit of $1 \times 10^{-2}$.*

| | $n$ | JIC($m$) (small penalty) | JIC($m$) (moderate penalty) | JIC($m$) (large penalty) | BIC($m$) | wavelets |
|---|---|---|---|---|---|---|
| Case 1 | 200 | 1.13(4.31) | 1.73(7.22) | 14.17(19.01) | 4.96(7.20) | 38.60(14.74) |
| | 500 | 0.14(0.03) | 0.14(0.03) | 2.22(8.88) | 1.77(4.96) | 9.80(13.06) |
| | 1000 | 0.07 (0.01) | 0.07(0.01) | 0.07(0.01) | 0.57(2.70) | 5.00(0.01) |
| Case 2 | 200 | 1.53(3.85) | 1.95(7.18) | 14.04(18.92) | 7.20(8.24) | 30.20(19.41) |
| | 500 | 0.24(0.79) | 0.19(0.04) | 3.62(11.20) | 2.52(5.79) | 10.60(13.95) |
| | 1000 | 0.09 (0.02) | 0.09(0.02) | 0.41(3.56) | 1.62(5.05) | 5.00(0.01) |
| Case 3 | 200 | 1.66(5.16) | 2.77(9.26) | 15.44(19.36) | 6.08(7.58) | 36.60(16.37) |
| | 500 | 0.22(0.95) | 0.15(0.04) | 3.65(11.31) | 3.98(7.05) | 11.40(14.74) |
| | 1000 | 0.09 (0.47) | 0.07(0.01) | 0.15(1.79) | 1.93(0.05) | 5.00(0.01) |

detected jump points and of true jump points. The distance between $\widehat{S}$ and $S$ can be used as a quantitative performance measure. A natural choice of such a distance is the Hausdorff distance defined below.

$$d_H(A, B) = \max\left(\sup_{x\in A}\inf_{y\in B}\|x-y\|, \sup_{y\in B}\inf_{x\in A}\|x-y\|\right),$$

where $A$ and $B$ are two point sets and $\|\cdot\|$ denotes the Euclidean distance. The smaller the value of $d_H\left(\widehat{S}, S\right)$, the better the jump detection procedure performs. The averaged Hausdorff distances for various methods are presented in Table 2, which are computed based on 1000 replicates. For the proposed method, the table shows that when $n$ gets larger, the Hausdorff distance between $\widehat{S}$ and $S$ gets smaller, and when $n$ is reasonably large, e.g., $n = 1000$, this distance does not depend on $P(n)$ much. As in Table 1, both the BIC and wavelet methods are not as effective as the proposed method in terms of the Hausdorff distance.

### 4·2. *Analysis of the sea-level pressure data*

Our example concerns December sea-level pressures during 1921–1992 observed by the Bombay weather station in India. The data are shown in Figure 2. Qiu & Yandell (1998) confirmed a jump around the year 1960 in this dataset, using their local polynomial jump detection algorithm. But this method relies on a threshold which will directly affect the estimated number of jumps. To determine the threshold, we need to choose a significance level in advance. Therefore, the difficulty in determining the number of jumps is transferred by this method to the difficulty in choosing the significance level, as discussed in Section 1. As a comparison, we do not need to choose such parameters in the proposed method, so, it should be able to provide a more objective analysis.

Next, we apply our proposed method to this dataset. The kernel function and the bandwidth $h_n$ are chosen to be the same as those used in the simulation examples. The adjustment factor is chosen to be $P(n) = (nh_n \log n)^{1/2}$. Our jump detection procedure identifies two jumps in 1938 and 1960, with corresponding estimated jump sizes $2.27(0.55)$ and $-2.18(0.55)$, respectively,

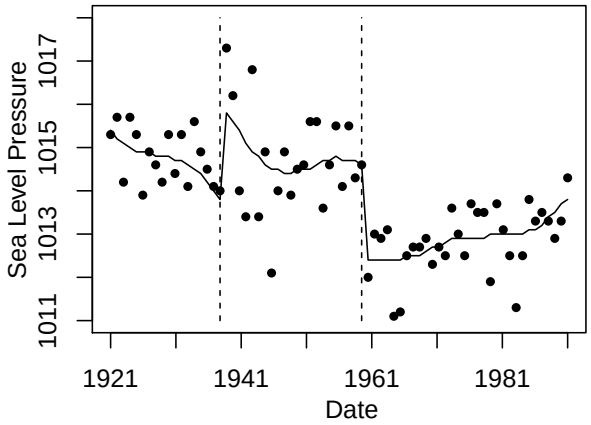


Fig. 2. Sea-level pressure data, the estimated regression curve (solid), and the detected jump dates (vertical dotted).

where the numbers in parentheses are the standard errors. The two jumps detected, shown in Figure 2 by vertical lines, catch the discontinuities in the observed data well. For this data, Qiu & Yandell (1998) only detect the jump around the year 1960. As a comparison, the proposed method in this paper detects that jump and another jump around the year 1938. The jump-preserving curve estimator is shown by the solid curve.

## ACKNOWLEDGEMENT

The authors are grateful for constructive comments from the Editor, the Associate Editor and two referees. This research is supported by a U.S. National Science Foundation grant, a Natural Sciences Foundation of China grant, and a Chinese Ministry of Education Foundation grant.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes technical proofs of Theorem 1–4 and Proposition 1. It also contains the derivation of the Bayesian information criterion (15).

## REFERENCES

ANTOCH, J., GRÉGOIRE G. & HUŠKOVÁ, M. (2007). Tests for continuity of regression functions. *Journal of Statistical Planning and Inference* **137**, 753–777.

ANTONIADIS, A. & GIJBELS, I. (2002). Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics* **14**, 7–29.

EFRON, B. (2004). The estimation of prediction error: covariate penalties and cross-validation (with discussion). *Journal of the American Statistical Association* **99**, 619–642.

FRICK, K., MUNK, A. & SIELING H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society (Series B)* **76**, 495–580.

GIJBELS, I. & GODERNIAUX, A. C. (2004). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics* **46**, 76–86.

GIJBELS, I., LAMBERT, A. & QIU, P. H. (2007). Jump-preserving regression and smoothing using local linear fitting: a compromise. *Annals of the Institute of Statistical Mathematics* **59**, 235–272.

GRÉGOIRE, G. & HAMROUNI, Z. (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis* **83**, 56–83.

HALL, P. & TITTERINGTON, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429–440.

HANNART, A. & NAVEAU, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics* **54**, 256–268.

HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd edition)*. Springer, New York.

JOO, J. & QIU, P. H. (2009). Jump detection in a regression curve and its derivative. *Technometrics* **51**, 289–305.

KANG, K. H., KOO, J. Y. & PARK, C. W. (2000). Kernel estimation of discontinuous regression functions. *Statistics and Probability Letters* **47**, 277–285.

KOO, J. Y. (1997). Spline estimation of discontinuous regression functions. *Journal of Computational and Graphical Statistics* **6**, 266–284.

LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory* **21**, 21–59.

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *The Annals of Statistics* **24**, 1667–1678.

LOADER, C. R. (1999). *Local Regression and Likelihood*. Springer, New York.

MCDONALD, J. A. & OWEN, A. B. (1986). Smoothing with split linear fits. *Technomerics* **28**, 375–402.

MA, S. & YANG, L. (2011). A jump-detecting procedure based on polynomial spline estimation. *Journal of Nonparametric Statistics* **23**, 67–81.

MÜLLER, C. H. (2002). Robust estimators for estimating discontinuous functions. *Metrika* **55**, 99–109.

MÜLLER, H. -G. (1992). Change-points in nonparametric regression analysis. *Annals of Statistics* **20**, 737–761.

QIU, P. H. (1994). Estimation of the number of jumps of the jump regression functions. *Communications in Statistics—Theory and Methods* **23**, 2141–2155.

QIU, P. H., ASANO, C. & LI, X. (1991). Estimation of jump regression functions. *Bulletin of Informatics and Cybernetics* **24**, 197–212.

QIU, P. H. & YANDELL, B. (1998). A local polynomial jump detection algorithm in nonparametric regression. *Technometrics* **40** 141–152.

QIU, P. H. (2005). *Image Processing and Jump Regression Analysis*. Wiley, New York.

SCHWARZ, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135–1151.

WANG, Y. (1995). Jump and sharp cusp detection via wavelets. *Biometrika* **82**, 385–397.

WU, J. S. & CHU, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *The Annals of Statistics* **21**, 1545–1566.

YAO, Y. -C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters* **6**, 181–189.

YIN, Y. Q. (1988). Detecting of the number, location and magnitudes of jumps. *Communications in Statistics—Stochastic Models* **4**, 445–455.

ZHANG, N. R. & SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22–32.