Electronic Journal of Statistics
Vol. 6 (2012) 478–500
ISSN: 1935-7524
DOI: 10.1214/12-EJS682

Local bandwidth selection via second derivative segmentation

Alexander Aue[*] **and Thomas C. M. Lee**[†]

Department of Statistics, University of California at Davis
4118 Mathematical Sciences Building, One Shields Avenue
Davis, CA 95616, USA
e-mail: alexaue@ucdavis.edu; tcmlee@ucdavis.edu

Haonan Wang[‡]

Department of Statistics, Colorado State University
Fort Collins, CO 80523, USA
e-mail: wanghn@stat.colostate.edu

Abstract: This paper studies the problem of local bandwidth selection for local linear regression. It is known that the optimal local bandwidth for estimating the unknown curve f at design point x depends on the curve's second derivative $f''(x)$ at x. Therefore one could select the local bandwidth $h(x)$ at x via estimating $f''(x)$. However, as typically estimating $f''(x)$ is a much harder task than estimating $f(x)$ itself, this approach for choosing $h(x)$ tends to produce less accurate results. This paper proposes a method for choosing $h(x)$ that bypasses the estimation of $f''(x)$, yet at the same time utilizes the useful fact that the optimal local bandwidth depends on $f''(x)$. The main idea is to first partition the domain of $f(x)$ into different segments for which the second derivative of each segment is approximately constant. The number and the length of the segments are assumed unknown and will be estimated. Then, after such a partition is obtained, any reliable, well-studied global bandwidth selection method can be applied to choose the bandwidth for each segment. The empirical performance of the proposed local bandwidth selection method is evaluated by numerical experiments.

AMS 2000 subject classifications: Primary 62G08.
Keywords and phrases: Bandwidth function, break point detection, local linear regression, optimal bandwidth.

Received January 2012.

1. Introduction

Local linear regression is a popular method for nonparametric curve estimation. An important aspect in its implementation is the choice for the amount of smoothing; i.e., the selection of the so-called bandwidth. If the target curve does not possess too much spatial variation in its structure, then it is well known that it could be well estimated by using one single (global) bandwidth throughout

[*] Supported in part by NSF under Grant DMS 0905400.
[†] Supported in part by NSF under Grant DMS 1007520.
[‡] Supported in part by NSF under Grants DMS 0706761, DMS 0854903 and DMS 1106975.

its whole domain. However, if the curve demonstrates a large amount of spatial inhomogeneities, then local bandwidth smoothing, sometimes also known as variable bandwidth smoothing, should be used. That is, different bandwidths are allowed to be used at different locations. This constitutes the so-called bandwidth function $h(x)$: the optimal local bandwidth $h(x)$ for estimating the regression function at location x is a function of x. The goal of this paper is to propose a method for choosing this bandwidth function $h(x)$.

In the literature different approaches have been proposed for choosing $h(x)$. The so-called plug-in approach relies on the asymptotic expression for the optimal bandwidth function. In this approach $h(x)$ is obtained by replacing the unknowns in this asymptotic expression with their estimates; e.g., see Fan and Gijbels (1992) and Gijbels and Mammen (1998). Another popular approach, sometimes known as the risk estimation approach, is to first construct an estimator of the mean squared error between the true and estimated function, and then choose $h(x)$ to minimize such an estimator. Examples include Fan and Gijbels (1995), Ruppert (1997) and Doksum, Peterson and Samarov (2000). Most recently Gluhovsky and Gluhovsky (2007) proposed a different approach, in which $h(x)$ is modeled as a smoothing spline and is defined as the minimizer of a novel penalty criterion.

The proposed method of this paper is motivated by the fact that the asymptotic expression for the optimal bandwidth at x depends on the second derivative of the unknown curve at x. We shall use Figure 1 to aid describing the main ideas of its major steps. A set of noisy observations together with the true but unknown curve are given in Figure 1(a). The noisy observations are then partitioned into different segments with the goal that the second derivative within each segment is approximately constant. The number of segments and the locations of the break points (i.e., the points at which adjacent segments meet) are automatically estimated by the minimum description length (MDL) principle (e.g., see Rissanen, 1989, 2007). Some asymptotic properties of this segmentation procedure will be provided below. See Figure 1(b) for the true second derivative and the corresponding segmentation. The next step is to calculate a single (global) bandwidth for each segment. These bandwidths are then joined together to form a piecewise constant function $h(x)$; see Figure 1(c). Notice that this bandwidth function is smaller near the middle of the curve, indicating that comparatively smaller bandwidths are required to recover the peak structure around $x = 0.5$. In order to preserve continuity, the partial local smoothing rule of Hall, Marron and Titterington (1995) is applied to this piecewise constant bandwidth function to obtain a final continuous bandwidth function, which is shown in Figure 1(d). Lastly this final bandwidth function is used to estimate the unknown curve. The resulting curve estimate is displayed in Figure 1(e). For comparative purposes, an estimate obtained by using a global bandwidth is shown in Figure 1(f). This global bandwidth was chosen by the AIC$_{\mathrm{C}}$ method of Hurvich, Simonoff and Tsai (1998). Observe that this "single bandwidth estimate", although recovering the peak structure at $x = 0.5$ reasonably well, undersmoothes the linear structures at both ends.

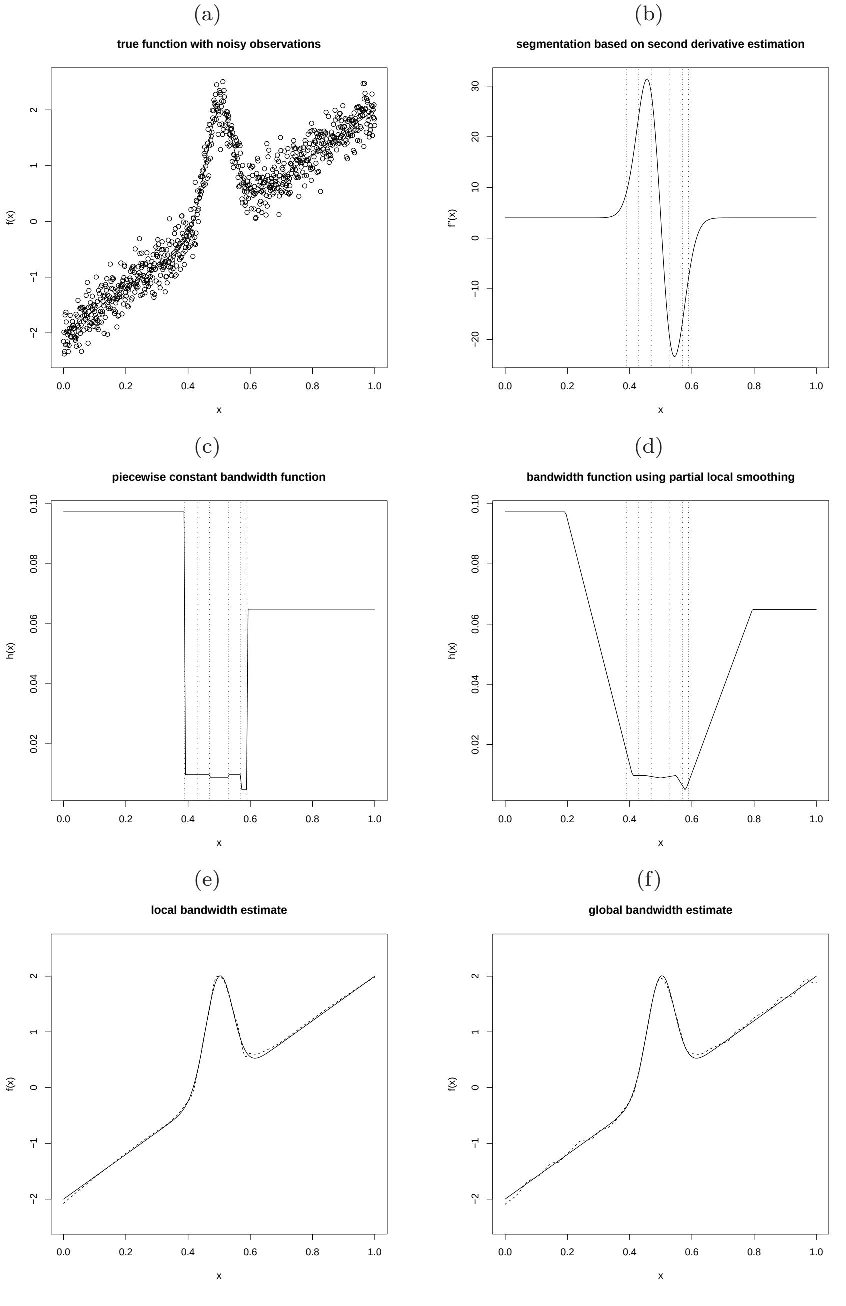

Fig 1. *An illustration of the steps involved in the proposed method. Panel (a): Noisy observations (circles) and true function (solid line). Panel (b): True second derivative (solid line) and seven segments obtained by partitioning the second derivative (vertical dotted lines indicate break point locations). Panel (c): Piecewise constant bandwidth function (solid line). Panel (d): Continuous bandwidth function obtained by partial local smoothing (solid line). Panel (e): Local linear smoothing estimate (dotted line) obtained from the bandwidth function in (d). Panel (f): Local linear smoothing estimate (dotted line) with a global bandwidth.*

The rest of this article is organized as follows. The proposed method is described in detail in Section 2. Some of its theoretical properties are provided in Section 3. Section 4 reports numerical simulation results while concluding remarks are offered in Section 6. Lastly technical details are delayed to the Appendix.

2. The proposed method

2.1. Background

Suppose observed are n pairs of observations $\{(x_i, y_i)\}_{i=1}^n$ satisfying

$$y_i = f(x_i) + \epsilon_i, \quad x_1 < \cdots < x_n, \quad \epsilon_i \sim \text{ iid } (0, \sigma^2),$$

where $f(x)$ is the unknown regression function of interest. For the moment we assume that the design points x_i's are uniformly distributed in $[a, b]$; non-uniform design densities will be discussed later. At any point x the local linear regression estimate of $f(x)$ is given by $\hat{f}_{h(x)}(x) = \hat{\alpha}_x$, where $\hat{\alpha}_x$, together with $\hat{\beta}_x$, are defined as the joint minimizer of

$$\sum_{i=1}^{n} [y_i - \{\alpha_x + \beta_x(x_i - x)\}]^2 \, K_{h(x)}(x - x_i)$$

(e.g., see Fan and Gijbels, 1996, Ch. 2). In the above $h(x)$ is the local bandwidth that controls the amount of smoothing at x, $K(\cdot)$ is the kernel function, and $K_{h(x)}(x) = K\{x/h(x)\}/h(x)$. Note that we view a kernel as a symmetric probability density function, not necessarily of bounded support.

If the goal is to minimize the expected local squared error $E\{f(x) - \hat{f}_{h(x)}(x)\}^2$, then it is well-known that the optimal choice of $h(x)$ admits the following asymptotic expression (e.g., see Fan and Gijbels, 1996, Ch. 3):

$$h_{\text{opt}}(x) = \left[\frac{\sigma^2 (b-a) \int K^2(u) du}{n \{f''(x) \int u^2 K(u) du\}^2} \right]^{\frac{1}{5}}. \tag{1}$$

Observe that in this expression for $h_{\text{opt}}(x)$, the only quantity that depends on x is the second derivative $f''(x)$. Therefore one way to select $h(x)$ is to first estimate $f''(x)$ and then plug-in this estimate into (1). However, as the estimation of $f''(x)$ is a much harder task than the estimation of $f(x)$, this approach for choosing local bandwidth tends to produce less satisfactory results.

Our proposed method for choosing $h(x)$ will bypass the estimation of $f''(x)$, but at the same time utilize the fact that $h_{\text{opt}}(x)$ depends on x only through $f''(x)$. The main idea is to first partition the domain of $f(x)$ into different segments for which the second derivative of each segment is approximately constant. Then one could use any reliable, well-studied global bandwidth selection method to choose the bandwidth for each segment. In other words, the key is to estimate $f''(x)$ with a best fitting piecewise constant function.

Now we return to the case when the density function for the design points x_i's are not uniform. In this case the term $(b-a)$ in the optimal bandwidth expression (1) will need to be replaced by the reciprocal of the density function at x, and an ideal segmentation of the regression function domain should take that into account. However, our numerical experience suggests that, unless the density function is highly skewed, the resulting segmentation using the uniform density assumption often leads to satisfactory empirical results. Results from simulation experiments to be reported below support this claim.

2.2. Second differencing

Fitting a piecewise constant function to $f''(x)$ would have been a standard problem if we had direct noisy observations of $f''(x)$. That is, if we could observe measurements like

$$y_i^* = f''(x_i) + e_i, \tag{2}$$

where the e_i's are iid zero mean errors. However, we do not observe such y_i^* and we suggest applying second differencing to y_i to obtain "pseudo data" that play a similar role as y_i^*. In the sequel we write $f_i = f(x_i)$ for all i.

We first apply a differencing operator to y_i and calculate (x_i', y_i') for $i = 1, \ldots, n-1$ as:

$$x_i' = \frac{x_{i+1} + x_i}{2}, \quad y_i' = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} + \frac{\epsilon_{i+1} - \epsilon_i}{x_{i+1} - x_i}.$$

Now apply another differencing operation to y_i' and obtain (x_i'', y_i'') for $i = 1, \ldots, n-2$ as:

$$\begin{aligned} x_i'' &= \frac{x_{i+1}' + x_i'}{2} = \frac{1}{4}(x_{i+2} + 2x_{i+1} + x_i), \\ y_i'' &= \frac{y_{i+1}' - y_i'}{x_{i+1}' - x_i'} \\ &= \frac{2}{x_{i+2} - x_i}\left\{\frac{f_{i+2} - f_{i+1}}{x_{i+2} - x_{i+1}} - \frac{f_{i+1} - f_i}{x_{i+1} - x_i} + \frac{\epsilon_{i+2} - \epsilon_{i+1}}{x_{i+2} - x_{i+1}} - \frac{\epsilon_{i+1} - \epsilon_i}{x_{i+1} - x_i}\right\}. \end{aligned} \tag{3}$$

Notice that y_i'' may be viewed as a discrete but noisy approximation of $f''(x_i'')$.

To simplify notation, write

$$\begin{aligned} z_i &= y_i'' \\ g_i &= \frac{2}{x_{i+2} - x_i}\left\{\frac{f_{i+2} - f_{i+1}}{x_{i+2} - x_{i+1}} - \frac{f_{i+1} - f_i}{x_{i+1} - x_i}\right\} \\ \eta_i &= \frac{2}{x_{i+2} - x_i}\left\{\frac{\epsilon_{i+2} - \epsilon_{i+1}}{x_{i+2} - x_{i+1}} - \frac{\epsilon_{i+1} - \epsilon_i}{x_{i+1} - x_i}\right\}. \end{aligned}$$

By noting that g_i is in fact a discrete version of $f''(x_i'')$, one could write (3) in the form of (2) as

$$z_i = g_i + \eta_i, \quad i = 1, \ldots, m \equiv n - 2.$$

We shall treat (x_i'', z_i) as our "pseudo data" and fit a piecewise constant function to them. However, the noise term η_i, although mean-zeroed, is now no longer independent. To derive the correlation structure of η_i, first write $d_i = x_{i+1} - x_i$. Then straightforward algebra shows that

$$\begin{aligned}
\operatorname{var}(\eta_i) &= \frac{2\sigma^2}{d_{i+1}^2 d_i^2}\left(\frac{2}{x_{i+2}-x_i}\right)^2 (d_{i+1}^2 + d_i^2 + d_{i+1}d_i) \quad \text{for } i = 1, \ldots, m,\\
\operatorname{cov}(\eta_i, \eta_{i-1}) &= -\frac{4\sigma^2}{d_{i+1}d_i^2 d_{i-1}} \frac{(2d_{i+1}d_{i-1} + d_i d_{i-1} + d_{i+1}d_i)}{(x_{i+2}-x_i)(x_{i+1}-x_{i-1})} \quad \text{for } i = 2, \ldots, m,\\
\operatorname{cov}(\eta_i, \eta_{i-2}) &= \frac{\sigma^2}{d_i d_{i-1}} \frac{4}{(x_{i+2}-x_i)(x_i - x_{i-2})} \quad \text{for } i = 3, \ldots, m,\\
\operatorname{cov}(\eta_i, \eta_j) &= 0 \quad \text{if } |i-j| > 2.
\end{aligned}$$

We will denote the covariance matrix, of size $m \times m$, specified by these equations as $\sigma^2 \boldsymbol{V}$. We note that the above expressions were derived by conditioning on the x_i's; i.e., they are conditional variances and covariances.

2.3. Second derivative segmentation using minimum description length

The next task is to fit a piecewise constant function to $\{(x_i'', z_i)\}_{i=1}^m$. To do so, we need to decide on how many pieces are required, and on the locations of the break points at which these pieces join. This is a model selection problem, in the sense that different candidate models (i.e., piecewise constant functions) may have a different number of parameters. We will use the minimum description length (MDL) principle (e.g., see Rissanen, 1989, 2007) to solve this problem.

The basic idea of the MDL principle can be explained as follows. Suppose a set of observed data $\boldsymbol{w}$ and a set of candidate models $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ for $\boldsymbol{w}$ are given. The goal is to select a "best" model for $\boldsymbol{w}$ from Θ. It is allowed that different $\boldsymbol{\theta}_i$'s may have a different number of parameters. One typical example is subset selection in the multiple linear regression context. The MDL principle defines the "best" model as the one that permits the most economical representation (or compression) of the data $\boldsymbol{w}$. That is, the best fitted model is the one that produces the shortest codelength for storing $\boldsymbol{w}$.

One general method for calculating the codelength for $\boldsymbol{w}$ is to decompose $\boldsymbol{w}$ into two components: a fitted model $\hat{\boldsymbol{\theta}}$ plus the corresponding residuals $\hat{\boldsymbol{r}}$. We shall use the notation $CL(a)$ to denote the codelength for an arbitrary object a. With this we have

$$CL(\boldsymbol{w}) = CL(\hat{\boldsymbol{\theta}}) + CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{\theta}}).$$

The MDL principle defines the best $\hat{\boldsymbol{\theta}}$ as the one that gives the smallest $CL(\boldsymbol{w})$. In the above expression we have stressed that $\hat{\boldsymbol{r}}$ is "conditional" on $\hat{\boldsymbol{\theta}}$.

For the piecewise constant function fitting problem that we consider here, $\boldsymbol{w}$ corresponds to $\boldsymbol{z} = (z_1, \ldots, z_m)^T$, $\hat{\boldsymbol{\theta}}$ corresponds to any fitted candidate piecewise constant function $\hat{\boldsymbol{g}}$, and $\hat{\boldsymbol{r}} = \boldsymbol{z} - \hat{\boldsymbol{g}}$. In other words, the MDL principle suggests that $\hat{\boldsymbol{\theta}}$ should be chosen as the one that minimizes

$$CL(\boldsymbol{z}) = CL(\hat{\boldsymbol{g}}) + CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{g}}). \tag{4}$$

Thus to apply MDL to solve the current segmentation problem, we need to derive computable expressions for $CL(\hat{\boldsymbol{g}})$ and $CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{g}})$, which in turn requires the calculation of $\hat{\boldsymbol{g}}$.

Suppose that there are $B+1$ segments in the candidate piecewise constant function (i.e., there are B break points), and that the number of x_i'''s in the j-th segment is m_j (such that $m_1 + \cdots + m_{B+1} = m$). Let $\lambda_1 < \cdots < \lambda_B$ be the locations of the B break points relative to the sample size (basically $\lambda_j = n_j/m$, where $n_j = m_1 + \cdots + m_j$; see Section 3 for the formal definition), and write $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_B)$. Also, define the ij-th element X_{ij} of the "model matrix" $\boldsymbol{X}$ as

$$X_{ij} = \begin{cases} 1 & \text{if } x_i'' \text{ is in the } j\text{-th segment,} \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, B+1$. Deleting repeated values, we next convert $\hat{\boldsymbol{g}}$ into $\hat{\boldsymbol{h}} = (\hat{g}_{n_1}, \ldots, \hat{g}_{n_{B+1}})^T$. To determine the candidate piecewise constant function maximum likelihood or generalized least squares can be applied to obtain

$$\hat{\boldsymbol{h}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{z}, \tag{5}$$

from which $\hat{\boldsymbol{g}}$ can be easily computed by reintroducing the corresponding number of repetitions m_j for each coordinate $\hat{h}_j$. Using this, it is shown in Appendix A that $CL(\boldsymbol{z})$ can be approximated by

$$\begin{aligned} \mathrm{MDL}(B, \boldsymbol{\lambda}) = \; & \log(B+1) + B \log(m-1) + \frac{1}{2} \sum_{j=1}^{B+1} \log m_j \\ & + \frac{m}{2} \log \frac{1}{m} (\boldsymbol{z} - \hat{\boldsymbol{g}})^T \boldsymbol{V}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{g}}). \end{aligned} \tag{6}$$

Notice that, for any given $\boldsymbol{z}$, any candidate piecewise constant function can be completely specified by $(B, \boldsymbol{\lambda})$ if $\hat{\boldsymbol{g}}$ is computed with (5). This fact is reflected in the notation of $\mathrm{MDL}(B, \boldsymbol{\lambda})$. We propose selecting the best fitting piecewise constant function as the minimizer of (6). Some theoretical properties of $\mathrm{MDL}(B, \boldsymbol{\lambda})$ is established in Section 3 below.

We also note that the criterion $\mathrm{MDL}(B, \boldsymbol{\lambda})$ can be straightforwardly modified to handle the situation when the noise variance varies with the segments. In this case the second last term will be replaced by $0.5 \sum \log(m_j + 1)$ while the last term will be replaced with a sum of such terms. The theoretical results in Section 3 can be slightly modified to accommodate this new criterion.

2.4. Practical minimization of MDL$(B, \boldsymbol{\lambda})$

This subsection describes a practical algorithm for minimizing (6). The idea is similar to performing forward selection followed by backward elimination in the multiple linear regression setting.

At the beginning of the algorithm, we fit only one segment to (x_i'', z_i); i.e., no break points. Then we add one break point to this initial fit. The location of this break point is chosen in a way that it provides the largest reduction of MDL$(B, \boldsymbol{\lambda})$ amongst all possible break point locations. Then a second break point is added to this two-piece constant function. As before, the location of this break point is chosen to maximize the reduction of MDL$(B, \boldsymbol{\lambda})$. This forward selection process continues until the adding of any new break points actually increases the value of MDL$(B, \boldsymbol{\lambda})$.

The second and last stage of this algorithm is backward elimination. The idea is to successively remove one break point at a time from those that were introduced in the previous forward selection process. At each time step the break point to be removed is chosen such that it permits the largest reduction of MDL$(B, \boldsymbol{\lambda})$ after its removal. This elimination process continues until no more removal of break points will cause a reduction in MDL$(B, \boldsymbol{\lambda})$.

The algorithm is akin to the knot addition and deletion idea of the highly successful smoothing method MARS (Friedman, 1991). In the context of regression spline fitting, MARS is known to perform empirically better than other knot addition/deletion strategies (Lee, 2002). It also worked exceptionally well in all our numerical work.

If the number of observations in any segment is too small, it may lead to unreliable estimates. Therefore we have imposed the constraint that each segment contains at least 5 observations.

We close this section with the following remark which outlines how the candidate segmentation given by $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_B < \lambda_{B+1} = 1$ can greatly facilitate numerical computations. To do so, utilize first the candidate segmentation to decompose the $m \times m$ matrix $\boldsymbol{V}$ into B block square submatrices $\boldsymbol{V}_j$ with dimension $m_j \times m_j$, where $m_j = \lfloor \lambda_j m \rfloor$ and $m_1 + \cdots + m_B = m$. This has the effect that the dependence between the different pieces in the segmentation is suppressed and we can work with independent blocks for the asymptotics. Since the MA(2) errors in the pseudo-data model $y_i = g_i + \eta_i$ are independent if they are more than two lags apart, the block creation does not affect the large sample properties. On the other hand, as a consequence of the above, one can simplify calculations involving the limit of the generalized least squares estimator $\hat{\boldsymbol{h}} = (\hat{h}_1, \ldots, \hat{h}_B)^T$. Each of its components is now of the form

$$\hat{h}_j = (\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{e}_j)^{-1} \boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{z}(\lambda_{j-1}, \lambda_j), \qquad j = 1, \ldots, B+1,$$

where $\boldsymbol{e}_j = (1, \ldots, 1)^T$ is the m_j-dimensional vector whose entries are all equal to one and $\boldsymbol{z}(\lambda_{j-1}, \lambda_j) = (z_{\lfloor \lambda_{j-1} m \rfloor + 1}, \ldots, z_{\lfloor \lambda_j m \rfloor})^T$. In Lemmas B.1 and B.2 below we show that both $\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{e}_j$ and $\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{z}(\lambda_{j-1}, \lambda_j)$ can be represented as certain fifth-order polynomials and the (ill-conditioned) inverse matrix $\boldsymbol{V}^{-1}$ does not need to be calculated explicitly.

2.5. Partial local smoothing

After a segmentation is obtained, the next task is to choose a (global) bandwidth for each segment. This can be achieved by applying any reliable global bandwidth selection method. In our numerical work to be reported in Section 4 below, we use the AIC$_\mathrm{C}$ method of Hurvich, Simonoff and Tsai (1998). Once a (global) bandwidth is obtained for each segment, all these bandwidths are then joined together to form a piecewise constant bandwidth function $h_0(x)$.

When the bandwidth function $h_0(x)$ is piecewise constant, it is customary to smooth those "corners" at which adjacent pieces are merged (e.g., see Fan and Gijbels, 1995), so that the resulting $h(x)$ is continuous. We also follow this custom and apply the partial local smoothing rule of Hall, Marron and Titterington (1995) to make $h_0(x)$ continuous. This partial local smoothing rule employs the following interpolation formula. Let τ_j and τ_{j+1} be the midpoints of the j-th and $(j+1)$-th pieces of the piecewise constant function $h_0(x)$ respectively. Therefore $h_0(\tau_j)$ is the (global) bandwidth obtained for the j-th segment; similarily for $h_0(\tau_{j+1})$. For any $x \in [\tau_j, \tau_{j+1})$, the partial local interpolation rule defines the final bandwidth function $h_1(x)$ as

$$h_1(x) = \begin{cases} h_0(\tau_1) & a \le x < \tau_1, \\ \frac{1}{\tau_{j+1}-\tau_j}\big\{h_0(\tau_j)(\tau_{j+1}-x) & \\ \qquad + h_0(\tau_{j+1})(x-\tau_j)\big\} & \tau_j \le x < \tau_{j+1}, \quad j = 1, \ldots, B, \\ h_0(\tau_{B+1}) & \tau_{B+1} \le x \le b. \end{cases} \tag{7}$$

Supportive theoretical and empirical results of this partial local smoothing rule can be found in Hall, Marron and Titterington (1995).

2.6. Summary

The main steps of the proposed method can be summarized as follows.

1. Apply the second differencing operation (3) and obtain (x_i'', z_i).
2. Find the "best" fitting piecewise constant function for (x_i'', z_i). This "best" fitting function is defined as the minimizer of (6), and it can be practically minimized using the algorithm described in Section 2.4.
3. From the "best" fitting piecewise constant function obtained in the previous step, a segmentation for (x_i, y_i) can be obtained. For each segment in this segmentation, apply a global bandwidth selector to choose a bandwidth. Merge the resulting global bandwidths together to form a piecewise constant bandwidth function $h_0(x)$. In our implementation the AIC$_\mathrm{C}$ method of Hurvich, Simonoff and Tsai (1998) is adopted as the global bandwidth selector.

4. Apply the partial local smoothing rule (7) to $h_0(x)$ to form a continuous bandwidth function $h_1(x)$.
5. Compute the estimate $\hat{f}_h(x)$ for $f(x)$ with local linear regression with bandwidth $h = h_1(x)$.

3. Theoretical properties

In this section, we study the asymptotic behavior of the proposed second differencing segmentation procedure. To do so, we have to further specify the form of the regression function f. For our purposes, we henceforth restrict the discussion on theoretical properties to regression functions f_0 that are once continuously differentiable with a piecewise constant second derivative f_0''. This is enabled in the following way. Without loss of generality, let $[a,b] = [0,1]$. Set $\lambda_0^0 = 0$ and $\lambda_{B^0+1}^0 = 1$. Then, we assume that $f_0''(x) = f_{0,j}''$ is constant for $x \in (\lambda_{j-1}^0, \lambda_j^0)$, $j = 1, \dots, B^0+1$, where $0 < \lambda_1^0 < \cdots < \lambda_{B^0}^0 < 1$ denote the B^0 break points. The second differencing procedure aims to partition f_0'' via noisy versions of the discrete approximations g_i^0 for which we then obtain

$$g_i^0 = h_j^0, \qquad n_{j-1}^0 < i < n_j^0 - 1, \quad j = 1, \dots, B^0 + 1. \tag{8}$$

The connection between λ_j^0 and n_j^0 is given by

$$n_j^0 = \lfloor \lambda_j^0 m \rfloor, \qquad \ell = 1, \dots, B^0 + 1,$$

with $\lfloor \cdot \rfloor$ denoting integer part and $m = n - 2$ as before. The number of g_i^0 in segment j is therefore equal to $m_j^0 = n_j^0 - n_{j-1}^0$. Certain edge effects in (8) have been left out. These occur when one transitions with the second differencing procedure from one segment into the next. As the number of these occurrences is clearly not larger than B^0, they do not affect the asymptotic.

Since the true partition is unknown, the MDL procedure is utilized as described in Section 2 and we select the best piecewise constant approximation of f_0'', which is determined by the parameters $(B, \boldsymbol{\lambda})$ according to the MDL criterion (6), adjusted for known B^0,

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \frac{2}{m} \mathrm{MDL}(B^0, \boldsymbol{\lambda}).$$

The following consistency result can be proved.

Theorem 3.1. *Assume the true second derivative f_0'' is piecewise constant. If B^0, the true number of breaks in the partition, is known, then the estimated break points $\hat{\boldsymbol{\lambda}} = (\lambda_1, \dots, \lambda_{B^0})$ converge with probability one to the true break points $\boldsymbol{\lambda}^0 = (\lambda_1^0, \dots, \lambda_{B^0}^0)$. That is,*

$$\hat{\boldsymbol{\lambda}} \xrightarrow{a.s.} \boldsymbol{\lambda}^0 \qquad (n \to \infty),$$

provided that $\sigma \sim d^2$, where $d = \max d_i$.

The proof of Theorem 3.1 is provided in Section B of the Appendix.

Note that the application of the differencing operator introduces dependence. For equally spaced design points with $d = d_i$, $\{\eta_i\}$ is a second order moving average process given by the difference equations

$$d^2\eta_i = \epsilon_i - 2\epsilon_{i+1} + \epsilon_{i+2}, \qquad i = 1, \ldots, m.$$

The moving average polynomial $\theta(z) = 1 - 2z + z^2 = (1-z)^2$ has two unit roots and imposes a special structure on the matrix $\boldsymbol{V}$ defined in Section 2.2 (see Appendix B below). Matrices of similar kind have been used in the detection of trend in time series and are discussed in depth in Anderson (1971). It should also be noted that it is critical here that the unit roots are known in advance and do not have to be estimated from the data. In the latter case, which has been dealt with for example in Anderson and Takemura (1986), certain pile-up effects cause the maximum likelihood estimator of the moving average unit roots to select an invertible set of parameters with positive probability, even asymptotically.

While the unit roots complicate matters for theoretical derivations, they also induce a superconsistent procedure under the piecewise constant second derivative assumption. That is, the rate of convergence is faster than the typical parametric rate of "root n"; see Lemma B.4 for the exact rate. The reason for this lies roughly in the fact that partial sums of the $\{\eta_i\}$ are telescoping, namely

$$d^2\sum_{i=1}^{m}\eta_i = \sum_{i=1}^{m}(\epsilon_{i+2} - \epsilon_{i+1}) - \sum_{i=1}^{m}(\epsilon_{i+1} - \epsilon_i) = (\epsilon_n - \epsilon_{n-1}) - (\epsilon_2 - \epsilon_1)$$

consists of exactly four terms for any m. Since the second differencing procedure utilizes the generalized least squares estimator $\hat{\boldsymbol{h}}$ in (5), the exact proof will deal with weighted versions of the above partial sums. We discuss details in Appendix B. These findings imply, and add theoretical justification for, the excellent finite sample performance of our procedure to be reported in Section 4 below.

In proving Theorem 3.1, we have assumed the number of break points, B^0, to be known. There are, as of now, only a few estimation procedures known in the literature whose theoretical foundation covers the case of unknown B^0. Two deal with independent random variables with common variance confounded by changes in the mean. Yao (1988) addresses the normal case and Horváth and Serbinowska (1995) the multinomial case. Recently Aue and Lee (2011) generalized the results of Yao (1988) to more complex image segmentation problems. While the theory behind the MDL-based second differencing procedure is difficult to establish, we conjecture that under a Gaussianity assumption one can retain Theorem 3.1 also for B^0 unknown. Since a formal proof of this conjecture would add unnecessary length to the paper with only marginal gains from a more practical point of view, we do not pursue this further. The simulations in this paper, however, indicate that the performance is very satisfactory also when B^0 is unknown and even when the true model is different from the one assumed in this section.

4. Simulation results

Two sets of numerical experiments were conducted. The first set of experiments was to evaluate the performance of the proposed method when the design points are regularly spaced, and to compare its estimation results with those obtained by the recent method proposed by Gluhovsky and Gluhovsky (2007). In the second set of experiments the proposed method is compared with other common bandwidth selection methods when the design density is non-uniform. For easy referencing, we shall call the proposed local bandwidth selection method SDS, short for Second Derivative Segmentation.

4.1. Regularly spaced data

Since we were unable to obtain the codes for the method proposed in Gluhovsky and Gluhovsky (2007), we repeated their simulation experiments with identical settings, and compare our numerical findings with those reported in their paper.

First, 100 sets of noisy observations were generated from the regression function

$$f(x) = x + 2\exp(-16x^2), \tag{9}$$

with $n = 81$ design points equally spaced in $[-2, 2]$ and $\sigma^2 = 0.5^2$. This test function is the same as the one in Figure 1, except the domain now is linearly "stretched" to $[-2, 2]$ from $[0, 1]$. For each of these noisy data sets, we applied the proposed method and the EBBS local bandwidth method of Ruppert (1997) to obtain estimates of $f(x)$. Denote, for the I-th noisy data set, the corresponding estimates obtained by the proposed method and the EBBS method as $\hat{f}_I(x)$ and $\tilde{f}_I(x)$ respectively. We calculated mean squared errors (MSEs) for $\hat{f}_I(x)$ as

$$\mathrm{MSE}(\hat{f}_I) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}_I(x_i) - f(x_i)\}^2,$$

and similarly for $\tilde{f}_I(x)$. Following Gluhovsky and Gluhovsky (2007), we then calculated the MSE ratio

$$\frac{\sum_{I=1}^{100}\mathrm{MSE}(\hat{f}_I)}{\sum_{I=1}^{100}\mathrm{MSE}(\tilde{f}_I)}$$

and the standard deviation of the MSE values for $\hat{f}_I(x)$ divided by the average of the MSE values for $\tilde{f}_I(x)$. These two values are $(0.73, 0.37)$, while the corresponding "best possible" pair from Table 1 of Gluhovsky and Gluhovsky (2007) is $(0.74, 0.26)$. The reason for using the words "best possible" in the previous sentence is as follows. The practical calculation of the proposed local bandwidth estimate of Gluhovsky and Gluhovsky (2007) involves the choices of (i) a tuning parameter λ and (ii) a fitting method $\hat{\beta}^{(i)}$. However, no automatic selection procedures were provided by these authors for choosing λ and $\hat{\beta}^{(i)}$. Instead, they reported results obtained from using different combinations of λ's and $\hat{\beta}^{(i)}$. The

Table 1

MSE ratios for the 15 normal mixture functions, denoted by F1 to F15. Numbers in parentheses are standard deviations adjusted by the averaged MSE of EBBS

	EBBS	$\hat{\beta}^{(4)}$	$\hat{\beta}^{(5)}$	SDS
F1	1.00 (0.32)	0.42 (0.18)	0.52 (0.20)	0.23 (0.15)
F2	1.00 (0.32)	0.45 (0.20)	0.55 (0.22)	0.30 (0.18)
F3	1.00 (0.28)	0.82 (0.24)	0.71 (0.22)	0.52 (0.22)
F4	1.00 (0.28)	0.88 (0.27)	0.80 (0.23)	0.73 (0.29)
F5	1.00 (0.26)	0.92 (0.28)	0.80 (0.24)	0.55 (0.30)
F6	1.00 (0.33)	0.42 (0.18)	0.53 (0.20)	0.22 (0.16)
F7	1.00 (0.33)	0.47 (0.18)	0.55 (0.21)	0.34 (0.19)
F8	1.00 (0.33)	0.45 (0.20)	0.53 (0.21)	0.26 (0.16)
F9	1.00 (0.33)	0.43 (0.18)	0.53 (0.20)	0.23 (0.13)
F10	1.00 (0.24)	0.63 (0.14)	0.69 (0.16)	0.45 (0.14)
F11	1.00 (0.32)	0.44 (0.19)	0.53 (0.20)	0.22 (0.14)
F12	1.00 (0.28)	0.52 (0.16)	0.59 (0.18)	0.30 (0.12)
F13	1.00 (0.31)	0.46 (0.18)	0.55 (0.20)	0.25 (0.11)
F14	1.00 (0.26)	0.64 (0.16)	0.68 (0.17)	0.49 (0.18)
F15	1.00 (0.27)	0.67 (0.19)	0.68 (0.20)	0.50 (0.17)

above pair $(0.74, 0.26)$ is the one that corresponds to the smallest MSE ratio. For reference, the worst pair is $(1.22, 0.61)$.

As in Gluhovsky and Gluhovsky (2007), we repeated the above experiment with 15 other regression functions. They are the 15 normal mixture functions listed in Marron and Wand (1992). The number of design points is $n = 181$, while σ^2 remains the same. The resulting MSE ratios and their scaled standard errors are calculated as before, and are listed in Table 1. Also listed in Table 1 are the corresponding values of the proposal of Gluhovsky and Gluhovsky (2007), using fitting methods $\hat{\beta}^{(4)}$ and $\hat{\beta}^{(5)}$ with their best possible λ's. Judging from these numerical values, one could conclude that, for regularly spaced data, the proposed method SDS is to be preferred over the method of Gluhovsky and Gluhovsky (2007) or the EBBS method of Ruppert (1997).

4.2. Non-uniform design densities

Recall that the proposed second derivative segmentation procedure assumes that the design density is uniform. In this second set of experiments we tested its performance when the design density was actually non-uniform. Altogether six beta densities with different parameters were used as the design density: $\text{Beta}[\frac{s+4}{5}, \frac{11-s}{5}]$ with $s = 1, \ldots, 6$. They are plotted in Figure 2. Two testing regression functions were used. The first regression function is essentially the same as (9), but with its domain mapped from $[-2, 2]$ to $[0, 1]$. The second regression function is

$$f(x) = \sin[2(4x-2)] + 2\exp[-16(4x-2)^2], \quad x \in [0,1],$$

which is displayed in Figure 3.

For each combination of design density and test function, 200 data sets were generated with $n = 200$ and a signal-to-noise ratio (snr) of 3, where snr is defined

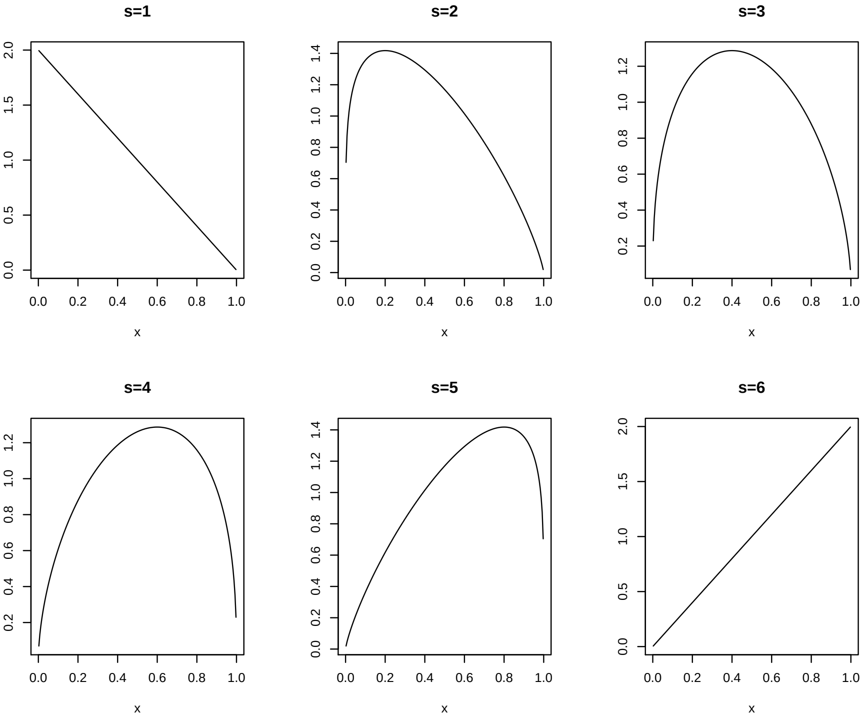

FIG 2. *The six beta densities used in the non-uniform design density experiments: Beta*$[\frac{s+4}{5}, \frac{11-s}{5}]$, $s = 1, \ldots, 6$.

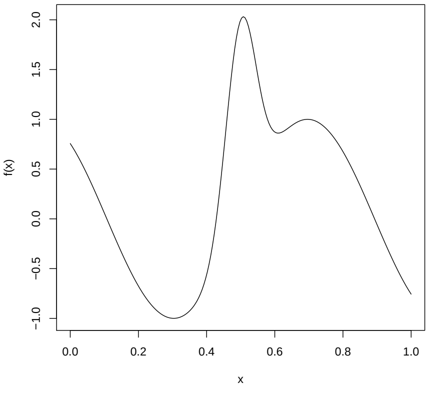

FIG 3. *The second regression function tested in the non-uniform design density experiments.*

Table 2

Averaged MSE values for different regression estimates for the non-uniform design density experiments. The first 6 rows are for the first test function (displayed in Figure 1), while the last 6 rows are for the second test function (displayed in Figure 3). Numbers in parentheses are standard errors

design density Beta$[\frac{s+4}{5}, \frac{11-s}{5}]$	*global*	*plug-in*	*EBBS*	*proposed*
$s = 1$	0.0309 (0.0011)	0.0263 (0.00104)	0.0329 (0.000927)	0.0225 (0.00101)
$s = 2$	0.0202 (0.000539)	0.0200 (0.000628)	0.0263 (0.000617)	0.0191 (0.00152)
$s = 3$	0.0193 (0.000467)	0.0193 (0.00055)	0.0256 (0.000503)	0.0134 (0.000503)
$s = 4$	0.0195 (0.000593)	0.0201 (0.00067)	0.0262 (0.000577)	0.0135 (0.000502)
$s = 5$	0.0229 (0.000677)	0.0206 (0.000693)	0.0275 (0.000671)	0.0155 (0.000616)
$s = 6$	0.0261 (0.000781)	0.0243 (0.000817)	0.0314 (0.000782)	0.0193 (0.000733)
$s = 1$	0.0155 (0.000472)	0.0135 (0.000433)	0.0143 (0.000381)	0.0121 (0.000433)
$s = 2$	0.0117 (0.00028)	0.0125 (0.000321)	0.0150 (0.000307)	0.0105 (0.000319)
$s = 3$	0.0106 (0.00025)	0.0109 (0.000284)	0.0137 (0.000272)	0.00926 (0.000239)
$s = 4$	0.0141 (0.000362)	0.0146 (0.000406)	0.0171 (0.000369)	0.0124 (0.00037)
$s = 5$	0.0122 (0.000300)	0.0123 (0.000325)	0.0149 (0.000314)	0.0102 (0.000328)
$s = 6$	0.0171 (0.000396)	0.0158 (0.000462)	0.0167 (0.000441)	0.0151 (0.000506)

as snr $= \|f\|/\sigma$ with $\|\cdot\|$ as the Euclidean norm. Then, for each generated data set, four regression estimates were obtained:

1. *global*: local linear regression using global bandwidth selected by the AIC$_\mathrm{C}$ method of Hurvich, Simonoff and Tsai (1998),
2. *plug-in*: kernel regression with the local plug-in bandwidth strategy of Herrmann (1997),
3. *EBBS*: the local bandwidth EBBS method of Ruppert (1997), and
4. *SDS*: the proposed method.

Finally, MSE values for all regression estimates are calculated. The averages of these MSE values, together with their standard errors, are reported in Table 2. From Table 2, one could see that, even for non-uniform design densities, SDS still performed favorably when comparing to other commmon methods.

We have also repeated the above experiments with $n = 400$ and snr $= 5$. Since these additional experiments provide similar empirical conclusions, their numerical results are omitted.

5. Real data

In this section the proposed procedure is applied to two real data sets. The first one is the motorcycle data set that has been analyzed by various previous authors (e.g., Fan and Gijbels, 1996). Here the design points x_i are the time at which the responses y_i were recorded after a simulated motorcycle impact experiment. These responses are the head acceleration of the test object. The data are displayed in the left panel of Figure 4. Since there are sharp changes near $x = 15$ and $x = 30$, a global constant bandwidth will not work well here. The proposed procedure correctly identified such changes and used smaller bandwidths to estimate their values. The estimated regression function, together

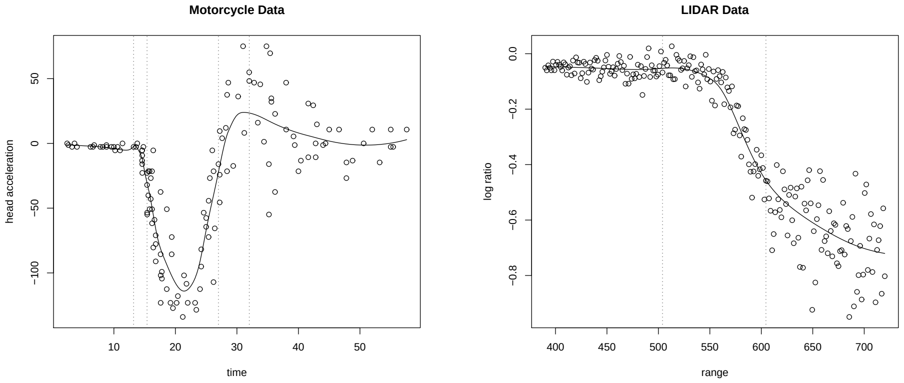

Fig 4. *Left panel: the motorcycle data set; right panel: the LIDAR data set. In each panel circles represent the data points, the solid line is the estimated regression curve obtained by the proposed method, and the vertical lines denote the locations of the break points.*

with the locations of the break points, are also displayed in the left panel of Figure 4.

Displayed in the right panel of Figure 4 is the so-called LIDAR data set (e.g., Ruppert, Wand and Carroll, 2003). LIDAR is a laser based technique for detecting chemical compounds in the atmosphere. The x-variable is the distance traveled by the laser light before it is bounced back to its origin. The y-variable is the log of the ratio of laser light received from two different frequency sources. As similar to the above motorcycle data set, a global constant bandwidth will not work well for this LIDAR data set. The proposed method was capable of first dividing it into three regions of approximately constant curvature and then selecting a tailored local bandwidth for each region; see the right panel of Figure 4.

Lastly we point out that for both data sets the noise levels are heteroscedastic. This violates the constant noise variance assumption made by the proposed procedure, but still the proposed procedure performed well.

6. Concluding remarks

In this article a method is proposed for choosing the bandwidth function for local linear smoothing. A major component of the method is the second derivative segmentation procedure. This procedure aims to partition the curve domain into homogeneous regions, so that a tailored bandwidth can be obtained for each region. Although this segmentation procedure is computationally expensive, it has been shown to be superconsistent if the underlying second derivative is piecewise constant. In addition, via theoretical results and numerical experiments, we have also demonstrated the superior empirical properties of the resulting local bandwidth selection method. We have further outlined how the procedure can handle hetereoscedastic data. Lastly, the second derivative segmentation idea can be combined with other smoothing methods. For example, the local linear regression used in this article can be straightforwardly replaced by smoothing splines.

Acknowledgment

The authors are grateful to the associate editor for many valuable comments which help clarify some issues present in the original version of this paper.

Appendix A: Derivation of MDL$(\boldsymbol{B}, \boldsymbol{\lambda})$

This part of the appendix derives the criterion MDL$(B, \boldsymbol{\lambda})$ given in (6), which approximates the codelength $CL(\boldsymbol{z})$. First we recall that B is the number of break points (i.e., there are $B+1$ segments) and that m_j is the number of x_i'' in the j-th segment.

The codelength $CL(\boldsymbol{z})$ is decomposed into two parts, $CL(\boldsymbol{z}) = CL(\hat{\boldsymbol{g}}) + CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{g}})$, and we begin with the first part. To completely specify a fitted piecewise constant function $\hat{\boldsymbol{g}}$, we need to specify (i) the number of segments, (ii) the locations of the break points, and (iii) the function values of each segments. Since there are $(B+1)$ segments, the codelength for (i) is $\log(B+1)$. For (ii), we restrict the location of any break point to be one of those midpoints that are equi-distanced to any two adjacent x_i'''s. As there are m x_i'''s and hence $(m-1)$ such midpoints, each of the B break point locations can be specified by an integer from $[1, \ldots, m-1]$. So the total codelength for (ii) is $B\log(m-1)$. We apply the following result from Rissanen (1989) to derive codelength for (iii): a parameter estimated from N data points can be effectively encoded with codelength $0.5 \log N$. Since the function value of the j-th segment is (approximately) estimated using m_j data points, so the total codelength for (iii) is $0.5 \sum_j \log m_j$. Combining these results we have

$$CL(\hat{\boldsymbol{g}}) = \log(B+1) + B\log(m-1) + \frac{1}{2}\sum_{j=1}^{B+1} \log m_j.$$

The next part is to calculate the codelength $CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{g}})$ for the residuals given $\hat{\boldsymbol{g}}$. Rissanen (1989) demonstrates that this is given by the negative of the conditional likelihood of $\hat{\boldsymbol{r}}$ given $\hat{\boldsymbol{g}}$, which for the current problem is

$$CL(\hat{\boldsymbol{r}}|\hat{\boldsymbol{g}}) = \frac{m}{2} \log \frac{1}{m} (\boldsymbol{z} - \hat{\boldsymbol{g}})^T \boldsymbol{V}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{g}}).$$

Combining the above two codelength expressions we arrive (6).

Appendix B: Proof of Theorem 3.1

The proof of Theorem 3.1 is given in several steps. In Section B.1, we discuss properties of the variance-covariance matrix $\boldsymbol{V}$ and establish certain key auxiliary results. These will then be invoked to derive the statement of Theorem 3.1 in

Section B.2. We assume throughout the proof that the regression $y_i = f(x_i) + \epsilon_i$ is canonical in order to contain the complexity of the proofs. It is expected that similar arguments apply also to the case non-canonical regression case.

B.1. The banded Toeplitz matrix $\boldsymbol{V}$

Let $\mathbb{T}$ be the complex unit circle and let $b : \mathbb{T} \to \mathbb{C}$ be the Laurent polynomial $b(t) = 6 - 4(t + t^{-1} + (t^2 + t^{-2})$. The symbol b induces the banded Toeplitz operator $T(b)$ that takes the values 6, -4 and 1 on the main diagonal, the first off-diagonals and the second-off diagonals, respectively. The symbol is unbounded as b has a zero of order four at $t = 1$, so that the smallest eigenvalues of the corresponding finite $(m \times m)$ Toeplitz matrices $T_m(b)$ are of exact order m^{-4}. This, in turn, implies that the largest elements of $T_m^{-1}(b)$ grow with rate m^4; e.g., see Böttcher and Grudsky (2005) for details on Toeplitz matrices. It is now easy to see that the $m \times m$ variance-covariance matrix $\boldsymbol{V}$ of Section 2.2 can be rewritten in terms of $T_m(b)$ simply as $\boldsymbol{V} = d^4 T_m(b)$. Most of the theory of banded Toeplitz matrices is based on boundedness of the symbol and is therefore not applicable in the current setting.

We need the following two important auxiliary results. Note that we do not need to compute the ill-conditioned inverse matrix $\boldsymbol{V}^{-1} = T_m^{-1}(b)$ directly.

Lemma B.1. *Let $\boldsymbol{e} = (1, \ldots, 1)^T$ be the m-dimensional vector whose elements are all equal to 1. Then,*

$$d^{-4}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e} = \frac{1}{24}\left(\frac{1}{30}m^5 + \frac{1}{3}m^4 + \frac{7}{6}m^3 + \frac{5}{3}m^2 + \frac{4}{5}m\right).$$

In particular, $d^{-4}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e} \sim \frac{1}{720}m^5$.

Proof. Let $\boldsymbol{a} = d^{-4}\boldsymbol{V}^{-1}\boldsymbol{e}$. Direct computations yield that the components a_i of $\boldsymbol{a}$ are given by

$$\begin{aligned} a_i = \frac{1}{24}\Bigg[&\left(\frac{m+1}{2}\right)^2\left(\frac{m+3}{2}\right)^2 \\ &- \left\{\left(\frac{m+1}{2}\right)^2 + \left(\frac{m+3}{2}\right)^2\right\}\left(i - \frac{m+1}{2}\right)^2 + \left(i - \frac{m+1}{2}\right)^4\Bigg]. \end{aligned}$$

To see that this is correct, it is most convenient to verify that $d^4\boldsymbol{V}\boldsymbol{a} = \boldsymbol{e}$. Now, $d^{-4}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e} = \boldsymbol{a}^T\boldsymbol{e} = \sum_{j=1}^m a_j$ and the statement of the lemma can be verified directly by elementary but lengthy calculations. □

Lemma B.2. *Choose $\kappa \in [0, 1]$, and let $\boldsymbol{a}(\lfloor \kappa m \rfloor)$ be defined as the vector in the proof of Lemma B.1 with dimension $\lfloor \kappa m \rfloor$. Let $0 \le \kappa_1 < \kappa_2 \le 1$. Then,*

$$S_m(\kappa_1, \kappa_2, \kappa) = \sum_{i=\lfloor \kappa_1 m \rfloor + 1}^{\lfloor \kappa_2 m \rfloor} a_i(\lfloor \kappa m \rfloor) = \frac{1}{24}\sum_{\ell=1}^{5}\left[p_\ell(\kappa_2, \kappa) - p_\ell(\kappa_1, \kappa)\right] m^\ell,$$

where

$$\begin{aligned}
p_1(x,y) =& \frac{4}{5}x,\\
p_2(x,y) =& \frac{5}{3}xy,\\
p_3(x,y) =& -x^3 + \frac{3}{2}x^2y + \frac{2}{3}xy^2,\\
p_4(x,y) =& -\frac{2}{3}x^3y + x^2y^2,\\
p_5(x,y) =& \frac{1}{5}x^5 - \frac{1}{2}x^4y + \frac{1}{3}x^3y^2.
\end{aligned}$$

Proof. Similar to the proof of Lemma B.1. □

B.2. Establishing Theorem 3.1

Recall that, since the value of B^0 is assumed known, the candidate segmentation is specified by the values $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_{B^0} < \lambda_{B^0+1} = 1$. Given such a candidate segmentation, we need to derive its large-sample behavior, in particular the bias that is induced when compared to the true segmentation $0 = \lambda_0^0 < \lambda_1^0 < \cdots < \lambda_{B^0}^0 < \lambda_{B^0+1}^0 = 1$.

To do so, we utilize the candidate segmentation and decompose the $m \times m$ matrix $\boldsymbol{V}$ into B^0 block square submatrices $\boldsymbol{V}_j$ with dimension $m_j \times m_j$, where $m_j = \lfloor \lambda_j m \rfloor$ and $m_1 + \cdots + m_{B^0} = m$. This has the effect that the dependence between the different pieces in the segmentation is suppressed and we can work with independent blocks for the asymptotics. Since the MA(2) errors in the pseudo-data model $y_i = g_i + \eta_i$ are independent if they are more than two lags apart, the block creation does not affect the large sample properties.

As a consequence of the above, we can simplify calculations involving the limit of the generalized least squares estimator $\hat{\boldsymbol{h}} = (\hat{h}_1, \ldots, \hat{h}_{B^0})^T$. Each of its components is now of the form

$$\hat{h}_j = (\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{e}_j)^{-1} \boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{z}(\lambda_{j-1}, \lambda_j), \qquad j = 1, \ldots, B^0 + 1, \tag{10}$$

where $\boldsymbol{e}_j = (1, \ldots, 1)^T$ is the m_j-dimensional vector whose entries are all equal to one and $\boldsymbol{z}(\lambda_{j-1}, \lambda_j) = (z_{\lfloor \lambda_{j-1} m \rfloor + 1}, \ldots, z_{\lfloor \lambda_j m \rfloor})^T$. To these, Lemmas B.1 and B.2 can be applied.

Lemma B.3. *If* $\lambda_{j-1} < \lambda_k^0 < \cdots < \lambda_{k+L-1}^0 < \lambda_j < \lambda_{k+L}^0$, *then*

$$\hat{h}_j \xrightarrow{\text{a.s.}} \sum_{\ell=0}^{L+1} w_{j,k+\ell} h_{k+\ell}^0,$$

where $w_{j,k} = w(0, \lambda_k^0 - \lambda_{j-1}, \nu_j)$, $w_{j,k+\ell} = w(\lambda_{k+\ell}^0, \lambda_{k+\ell+1}^0, \nu_j)$ *for* $\ell = 1, \ldots, L$, *and* $w_{j,k+L} = w(\lambda_{k+L}^0, \lambda_j, \nu_j)$ *with*

$$w(\kappa_1, \kappa_2, \kappa_3) = \frac{6(\kappa_2^5 - \kappa_1^5)}{\kappa_3^5} - \frac{15(\kappa_2^4 - \kappa_1^4)}{\kappa_3^4} + \frac{10(\kappa_2^3 - \kappa_1^3)}{\kappa_3^3}$$

and $\nu_j = \lambda_j - \lambda_{j-1}$.

Proof. Observe first that the factors d^4 involving the design spacing d cancel out, since they appear both in the numerator and the denominator of the right-hand side in (10). For the remaining denominator Lemma B.1 implies a leading term of exact order $\frac{1}{720}\lfloor \nu_j^5 m^5 \rfloor$. For the numerator we first decompose $\boldsymbol{z}(\lambda_{j-1}, \lambda_j) = \boldsymbol{g}^0(\lambda_{j-1}, \lambda_j) + \boldsymbol{\eta}(\lambda_{j-1}, \lambda_j)$, where the quantities of the right-hand side are the accordingly defined deterministic and random components. Adopting the notations $n_j = \lfloor \lambda_j m \rfloor$, $n_j^0 = \lfloor \lambda_j m \rfloor$, $m_j = n_j - n_{j-1}$ and $m_j^0 = n_j^0 - n_{j-1}^0$, we obtain

$$
\begin{aligned}
&d^{-4}\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{g}^0(\lambda_{j-1}, \lambda_j)\\
&\quad = \sum_{i=1}^{m_j} a_i(m_j) g^0_{\lambda_{j-1}+i}\\
&\quad = h_k^0 \sum_{i=1}^{n_k^0 - n_{j-1}} a_i(m_j) + \sum_{\ell=1}^{L-1} h^0_{k+\ell} \sum_{i=n^0_{k+\ell-1}+1}^{n^0_{k+\ell}} a_i(m_j) + h^0_{k+L} \sum_{i=n_{k+L-1}+1}^{n_j} a_i(m_j)\\
&\quad = h_k^0 S_m(0, \lambda_k^0 - \lambda_{j-1}, \nu_j) + \sum_{\ell=1}^{L-1} h^0_{k+\ell} S_m(\lambda^0_{k+\ell-1}, \lambda^0_{k+\ell}, \nu_j)\\
&\qquad + h^0_{k+L} S_m(\lambda^0_{k+L-1}, \lambda_j, \nu_j),
\end{aligned}
$$

where each S_m term refers to the fifth order polynomial defined in Lemma B.2. Now, applying Lemma B.2, one obtains for $\ell = 1, \dots, L-1$ that

$$
\begin{aligned}
&h^0_{k+\ell} S_m(\lambda^0_{k+\ell-1}, \lambda^0_{k+\ell}, \nu_j)\\
&\quad = \frac{h^0_{k+\ell}}{24} \sum_{u=1}^{5} \left[p_u(\lambda^0_{k+\ell}, \nu_j) - p_u(\lambda^0_{k+\ell-1}, \nu_j) \right] m^j\\
&\quad \sim \frac{h^0_{k+\ell}}{24} \bigg[\frac{1}{5}\big(\{\lambda^0_{k+\ell}\}^5 - \{\lambda^0_{k+\ell-1}\}^5\big) - \frac{1}{2}\big(\{\lambda^0_{k+\ell}\}^4 - \{\lambda^0_{k+\ell-1}\}^4\big)\nu_j\\
&\qquad + \frac{1}{3}\big(\{\lambda^0_{k+\ell}\}^3 - \{\lambda^0_{k+\ell-1}\}^3\big)\nu_j^2 \bigg] m^5\\
&\quad = v_{j,k+\ell} m^5.
\end{aligned}
$$

Similar expressions can be computed for the first term $h_k^0 S_m(0, \lambda_k^0 - \lambda_{j-1}, \nu_j)$ and the last term $h^0_{k+L} S_m(\lambda^0_{k+L-1}, \lambda_j, \nu_j)$. From the preceding it follows that $\hat{h}_j$ is of the form specified in the lemma. It remains to determine the form of the asymptotic weights $w_{j,k+\ell}$. These are given as the limit as $m \to \infty$ of the ratios

$$
\left(\frac{\nu_j^5 m^5}{720} \right)^{-1} v_{j,k+\ell} m^5
$$

and are easily shown to coincide with the expressions given in the statement of the lemma. Since the random components $(\boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{e}_j)^{-1} \boldsymbol{e}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{\eta}(\lambda_{j-1}, \lambda_j)$ satisfy a strong law of large numbers the assertion of the lemma follows. □

Under the piecewise second derivative assumption, the next lemma establishes that the term $(\boldsymbol{e}_j^T\boldsymbol{V}_j^{-1}\boldsymbol{e}_j)^{-1}\boldsymbol{e}_j^T\boldsymbol{V}_j^{-1}\boldsymbol{\eta}$ does not only satisfy a strong law of large numbers but that typically the rate of convergence to the zero limit is fast. The speed is controlled by the existence of higher order moments of the innovations ϵ_i and even the assumption of a finite variance only yields superconsistency.

Lemma B.4. *Let $\boldsymbol{e}$ be as in Lemma B.1. The weighted random sums $(\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta}$ converges almost surely to zero and satisfies*

$$(\sigma\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta} = \mathcal{O}\left(\frac{1}{m^{2-\delta}}\right)$$

for $\delta > 0$. The rate is better than $\mathcal{O}(m^{-2})$ if in addition $E[|\epsilon_1|^{2+\Delta}|] < \infty$ for some $\Delta > 0$.

Proof. Fix $\epsilon > 0$ and note that $\mathrm{Var}((\sigma\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta}) = (\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1} \sim 720m^{-5}$. An application of Tchebyshev's inequality yields

$$P\left(\frac{m^{2-\delta}}{\sigma}\left|(\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta}\right| \geq \epsilon\right) \leq \frac{C}{\epsilon^2 m^{1+2\delta}}$$

for some constant $C > 0$. The latter implies

$$\sum_{m=1}^{\infty} P\left(\frac{m^{2-\delta}}{\sigma}\left|(\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta}\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\sum_{m=1}^{\infty}\frac{1}{n^{1+2\delta}} < \infty$$

and, on account of the Borel-Cantelli lemma, $(\sigma\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{\eta} = \mathcal{O}(m^{\delta-2})$ with probability one. This proves the first part of the lemma. The second follows by similar arguments from a higher-order Markov inequality. □

Proof of Theorem 3.1. Recall the MDL criterion in (6) and observe that

$$\frac{2}{m}\mathrm{MDL}(B^0, \boldsymbol{\lambda}) \sim \log\frac{1}{m}(\boldsymbol{z} - \hat{\boldsymbol{g}})^T\boldsymbol{V}^{-1}(\boldsymbol{z} - \hat{\boldsymbol{g}})$$

as $m \to \infty$. Using $\hat{\boldsymbol{\eta}} = \boldsymbol{z} - \hat{\boldsymbol{g}} = \boldsymbol{\eta} + \boldsymbol{g} - \hat{\boldsymbol{g}}$, it follows first that

$$\hat{\boldsymbol{\eta}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}^T\boldsymbol{V}^{-1}\boldsymbol{\eta} + (\boldsymbol{g} - \hat{\boldsymbol{g}})^T\boldsymbol{V}^{-1}(\boldsymbol{g} - \hat{\boldsymbol{g}}) + 2\boldsymbol{\eta}^T\boldsymbol{V}^{-1}(\boldsymbol{g} - \hat{\boldsymbol{g}}).$$

It is clear from Lemmas B.3 and B.4 that $\frac{1}{m}\hat{\boldsymbol{\eta}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\eta}} \sim \sigma^2 + R_m$, where the remainder term R_m is positive almost surely if $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^0$. The assertion of Theorem 3.1 is therefore implied by the strict concavity of the logarithm. □

References

Anderson, T. W. (1971). *The statistical analysis of time series.* Wiley: New York. MR0283939

Anderson, T. W. and Takemura, A. (1986). Why do noninvertible estimated moving averages occur? *Journal of Time Series Analysis* **7** 235-254. MR0883008

Aue, A. and Lee, T. C. M. (2011). On image segmentation using information theoretic criteria. *The Annals of Statistics* **39** 2912-2935.

Böttcher, A. and Grudsky, S. M. (2005). *Spectral properties of banded Toeplitz matrices*. Society for Industrial Mathematics: Philadelphia.

Doksum, K., Peterson, D. and Samarov, A. (2000). On variable bandwidth selection in local polynomial regression. *Journal of the Royal Statistical Society Series B* **62** 431-448. MR1772407

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20** 2008-2036. MR1193323

Fan, J. and Gijbels, I. (1995). Data–driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society Series B* **57** 371-394. MR1323345

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines (with discussion). *The Annals of Statistics* **19** 1-141. MR1091842

Gijbels, I. and Mammen, E. (1998). Local adaptivity of kernel estimates with plug-in local bandwidth selectors. *Scandinavian Journal of Statistics* **25** 503-520. MR1650027

Gluhovsky, I. and Gluhovsky, A. (2007). Smooth location-dependent bandwidth selection for local polynomial regression. *Journal of the American Statistical Association* **102** 718-725. MR2370862

Hall, P., Marron, J. S. and Titterington, D. M. (1995). On partial local smoothing rules for curve estimation. *Biometrika* **82** 575-587. MR1366283

Herrmann, E. (1997). Local Bandwidth Choice in Kernel Regression Estimation. *Journal of Computational and Graphical Statistics* **6** 35-54. MR1451989

Horváth, L. and Serbinowska, M. (1995). Testing for changes in multinomial observations: the Lindisfarne problem. *Scandinavian Journal of Statistics* **22** 371-384. MR1363219

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B* **60** 271-293. MR1616041

Lee, T. C. M. (2002). On Algorithms for Ordinary Least Squares Regression Spline Fitting: A Comparative Study. *Journal of Statistical Computation and Simulation* **72** 647-663. MR1930486

Marron, J. S. and Wand, M. P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics* **20** 712-736. MR1165589

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore. MR1082556

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Springer. MR2287233

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92** 1049-1062. MR1482136

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters* **6** 181-189. MR0919373