I N S T I T U T   D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0130

DATA-DRIVEN DISCONTINUITY DETECTION IN DERIVATIVES OF A REGRESSION FUNCTION

I. GIJBELS AND A.-C. GODERNIAUX

http://www.stat.ucl.ac.be

Data-driven discontinuity detection in derivatives of a regression function

I. Gijbels and A-C. Goderniaux[1]

July 2001

Abstract

This paper provides a fully data-driven procedure for estimating the locations of jump discontinuities occuring in the k^{th} derivative of an unknown regression function. The basic ingredients for the procedure are a two-steps method for estimating the locations of the jump discontinuities, a bootstrap procedure for selecting the smoothing parameters involved in this estimation, and a cross-validation method for estimating the number of discontinuities in a derivative function. The paper extends ideas developed for change point detection in the regression function itself by Gijbels and Goderniaux (2000). A simulation study illustrates the performance of the procedure, and applications to some real data demonstrate its use in practice.

Key words: abrupt change, bandwidth, bootstrap, cross-validation, derivative function, least-squares fitting, local polynomial regression.

1 Introduction

Jump discontinuities represent only one type of irregularities that might occur in an otherwise smooth regression function. Other types of irregularities include changes in a derivative of the regression function. An example is a jump discontinuity in the first derivative which would appear as an 'abrupt change in the direction' of the regression

[1] Institut de Statistique, Université catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium.

function itself. A jump discontinuity in the second derivative would appear as a 'shoulderpoint' in the regression function itself. Such type of irregularities may show up when estimating regression curves in applications. Consider for example the Motorcycle data, reported on by Schmidt, Mattern and Schüler (1981). These are 123 measurements on test objects that underwent a simulated motorcycle collision. Recorded were the time (in milliseconds) after the impact (the variable X) and the head acceleration (in g) of the test object (the variable Y). Looking at the raw data, presented in Figure 5.3 in Section 5, one can 'suspect' some changes in the direction of the acceleration somewhere around roughly 15, 23 and 32 milliseconds. As a second example we consider data on winning speeds given in Poirier (1973). The data collected are the winning speeds (in miles per hour) at the Indianapolis 500 Race (a car race) for the years 1911 (time of conception of the race) till 1971. This race took place each year with the exception of the years 1917, 1918, 1942, 1943, 1944 and 1945, the war years. The data are displayed in Figure 5.5. Clearly, as we all know, there is an overall increase in the winning speeds. The question that is of interest for these data is to find out whether the two World Wars had an impact on the racing performance, through war-related technical advance (i.e. speeding up of developments in technology) on one hand and through the interruption of the activity (training and such) on the other hand. This question is related to find out whether there are accelerations or decelarations in the increase of the winning speeds. This translates into a detection of possible jump discontinuities in the third derivative. In Section 5 we will analyze both data sets using our data-driven procedure for detecting discontinuities in derivatives of the regression function.

Furthermore, if the interest is in estimating the k^{th} derivative of a regression function, then it is better to obtain an estimator that is adapted to possible jump discontinuities in the derivative function. See Figure 5.1 in Section 5, in which we represent three different regression functions (upper panels) for which a jump discontinuity occurs at the point 0.5 in the first derivative, for Figures 5.1 (a) and (b), and in the second derivative for Figure 5.1 (c). The lower panels of Figure 5.1 depict the corresponding true derivative functions as solid lines (first derivatives in Figures 5.1 (d) and (e), and the second derivative function in Figure 5.1 (f)). Presented in these lower panels are also the estimated derivative curves, adapted to the estimated jump discontinuities in the derivative functions (the dotted curves) together with smooth estimates of the derivative curves (the long-dashed curves). The estimates are based on simulated samples of size $n = 100$. Figure 5.1 clearly reveals that the estimates of the derivative functions assuming smooth derivatives are quite different from the estimates of the derivative functions allowing for possible jump points. For more details of these and other simulated examples see Section 5.

There are various approaches in the literature dealing with nonparametric regression with abrupt changes in a derivative. Hall and Titterington (1992) proposed a kernel-based estimation method to estimating curves with peaks and edges and Jose and Ismail (1997) rely on the analysis of residuals. Müller (1992) and Wu and Chu (1993), among others, have suggested methods based on differences of nonparametric kernel estimates. Eubank and Speckman (1994), Speckman (1994, 1995) and Cline, Eubank and Speckman (1995) considered semiparametric spline-based methods. Local polynomial procedures have been used by McDonald and Owen (1986), Horváth and Kokoszka (1997), Qui and Yandell (1998) and Spokoiny (1998), among others. For wavelet-based method see for example Wang (1995) and Raimondo (1998). All methods have in common that they involve the choice of some kind of smoothing parameters, and the performance of the methods often depends heavely on these choices.

In this paper we provide a fully data-driven procedure for estimating jump discontinuities in a derivative curve. The method also includes a data-driven way of determining the number of discontinuities in a derivative curve. The data-driven procedure developed here is a generalization of the procedure for estimating regression curves with jump discontinuities proposed by Gijbels, Hall and Kneip (1996, 1999) and studied further by Gijbels and Goderniaux (2000). Such a generalization requires an appropriate choice of a diagnostic function and a parametric family for least-squares fitting when dealing with estimation of a (known) number of jump discontinuities. In order to estimate the number of discontinuities we rely on a cross-validation method which basically combines cross-validation ideas discussed in Müller, Stadtmüller and Schmitt (1987) in the context of bandwidth selection for (smooth) derivative curves and in Müller and Stadtmüller (1999) in the context of estimation of unsmooth regression functions. So, this paper brings together several ideas that have been used in different contexts, and exploits them in the current context of detecting abrupt changes in a regression function and/or its derivatives.

The paper is organized as follows. In Section 2 we focus on the case of jump discontinuities in the first derivative. We introduce the estimation method and the algorithm to choose the bandwidth parameters. The generalization to the case of jump discontinuities in a higher order derivative is provided in Section 3. In Section 4, we discuss a cross-validation criterion to estimate the number of discontinuities appearing in the k^{th} derivative of the regression function. The performance of the data-driven detection procedure is illustrated via a simulation study in Section 5. In that section we also present the analysis of the two data sets described above. In a last section, we provide some discussion.

2 Estimation of jump discontinuities in the first derivative

2.1 Statistical model

We assume that a sample of n data pairs $\chi = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ is observed, generated from the model

$$Y_i = g(X_i) + \varepsilon_i, \qquad 1 \leq i \leq n.$$

We consider design points X_i which are either regularly spaced on I=[0,1] or are the order statistics of a random sample from a distribution having a density f supported on I. The errors ε_i are assumed to be independent and identically distributed with zero mean and finite variance σ^2. We suppose that $g(.)$, the unknown regression function is such that its first derivative is continuous except at an unknown finite number of jump discontinuities. Denote this unknown number of jump discontinuities by ν.

We consider first the case of a single jump discontinuity, appearing at the location $x_0 \in]0, 1[$ in the first derivative. The generalization to the case of more than one discontinuity in the derivative function is discussed in Section 2.4. In Section 3 we generalize the method to estimation of the locations of jump discontinuities in the k^{th} derivative of g.

The rate of convergence of the estimator $\hat{x}_0$ of x_0, the location of a jump discontinuity in the k^{th} derivative function is of order $\mathrm{O}(n^{-1/(2k+1)})$. This is a conjecture, based on the fact that we rely on the two-steps estimation method for estimating locations of jump discontinuities introduced by Gijbels, Hall and Kneip (1996, 1999) and for this method the rate $\mathrm{O}(n^{-1/(2k+1)})$ has been established under certain conditions on the bandwidth parameters. Our data-driven procedure is based on using a bootstrap procedure for the bandwidth selection. No theoretical behaviour of this bootstrap bandwidth selector has been established yet.

2.2 Estimation procedure

We adapt the two-steps estimation method as discussed by Gijbels, Hall and Kneip (1996, 1999) in the case of jump discontinuities in the regression function itself, to the case of jump discontinuities in the first derivative. This adaptation is straightforward, and consists of two steps: in a first step a preliminary estimator of x_0, the location of the jump discontinuity in the first derivative, is obtained via the evaluation of an appropriate diagnostic function; in a second step an improved estimator of x_0 is obtained by least-squares fitting of an appropriate parametric model in a small interval around the initial estimator of x_0.

2.2.1 Diagnostic step

A diagnostic function is used to obtain a first estimator $\tilde{x}_0$ of x_0. A way to detect a jump discontinuity in the first derivative is by looking at locations with high second order derivatives. So we suggest to consider the second derivative of a Nadaraya-Watson kernel estimator (see Nadaraya (1964) and Watson (1964)) and define the diagnostic function D by

$$D(x, h_1) = \frac{\partial^2}{\partial x^2} \left(\frac{\sum_{i=1}^n K\{(x - X_i)/h_1\} Y_i}{\sum_{i=1}^n K\{(x - X_i)/h_1\}} \right), \tag{2.1}$$

where K is a compactly supported twice differentiable kernel function and $h_1 > 0$ is a bandwidth. A first rough estimator of x_0 is then given by

$$\tilde{x}_0 = \text{argmax}_{x \in]vh_1, 1-vh_1[} |D(x, h_1)|,$$

where $[-v, v]$ denotes the support of K.

Note that in case of equally-spaced design we can also use the second derivative of the numerator of a Nadaraya-Watson kernel estimator as a diagnostic function:

$$D(x, h_1) = \frac{1}{nh_1^3} \sum_{i=1}^n K^{(2)}\{(x - X_i)/h_1\} Y_i ,$$

since this will be proportional to an estimator for the second derivative of the regression function.

2.2.2 Least-squares step

The aim of this second step is to improve the initial estimator $\tilde{x}_0$ of x_0. Therefore we construct an interval concentrated around $\tilde{x}_0$, denoted as $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$, with $h_2 > 0$, to which x_0 belongs with high probability. Denote by $\{i_1, i_1+1, \ldots, i_2\}$ the set of integers i such that $X_i \in [\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$. Suppose that the change point occurs between the two design points X_{i_0} and X_{i_0+1}. We discuss how to estimate i_0 and hence x_0. To improve the performance of $\tilde{x}_0$, we fit via the least-squares method a linear function (a first order polynomial) to the left and the right of the point X_{i_0} in the interval $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$. More precisely, we search for the value of i_0 that minimizes the sum of squares

$$\sum_{i=i_1}^{i_0} \{Y_i - (a_1 + b_1 X_i)\}^2 + \sum_{i=i_0+1}^{i_2} \{Y_i - (a_2 + b_2 X_i)\}^2 \tag{2.2}$$

where

$$b_j = \frac{\sum_{i=r_j}^{s_j} (X_i - \overline{X}_j)(Y_i - \overline{Y}_j)}{\sum_{i=r_j}^{s_j} (X_i - \overline{X}_j)^2} \qquad \text{and} \qquad a_j = \overline{Y}_j - b_j \overline{X}_j$$

with $r_1 = i_1$, $s_1 = i_0$ and $r_2 = i_0 + 1$, $s_2 = i_2$. The quantities a_1 and b_1, and a_2 and b_2 are nothing but the usual estimated coefficients of a linear regression model.

Denote by $\hat{\imath}_0$ the minimizer of the sum of squares (2.2). The final estimator for x_0 is then defined as the mid-point between $X_{\hat{\imath}_0}$ and $X_{\hat{\imath}_0+1}$:

$$\hat{x}_0 = \frac{1}{2}\left(X_{\hat{\imath}_0} + X_{\hat{\imath}_0+1}\right).$$

2.3 Choice of the bandwidth parameters

The two-steps method described above involves two smoothing parameters, the bandwidths h_1 and h_2. The choice of these parameters is rather crucial. We now discuss data-driven choices of these bandwidths.

The diagnostic function depends on the bandwidth h_1, but at a jump discontinuity it will consistently be large for many h_1 values. We then identify the jump discontinuity as the point x in the neighbourhood of which $|D(x, h_1)|$ is consistently large for a range of values of h_1, and take the smallest h_1 value for which this still holds (i.e. decreasing the h_1 value further would introduce artificial peaks at other locations). We consider a set of decreasing h_1 values, namely $h_{1,i} = h_0 r^i$, for $i = 0, 1, 2, \cdots$, with $h_0 > 0$ and $0 < r < 1$. The choice of the biggest h_1 value in this set (i.e. the value h_0) and the choice of the multiplication factor r are not important. Safe choices are h_0 big enough (say up to at most half of the length of the domain of the regression function) and r close to one.

See Gijbels and Goderniaux (2000) for more details when dealing with detection of jump discontinuities in the regression function itself.

The choice of the bandwidth h_2 is also crucial since the least-squares fit with a piecewise linear function will be quite bad if the interval $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$ is too large such that the unknown function g is far from a piecewise linear function in that interval. Suppose that we have obtained an estimator $\hat{\imath}_0$ (resp. $\hat{x}_0$) of i_0 (resp. x_0) by the two-steps method explained in Section 2.2, using the data-driven choice of h_1 as explained above and a certain fixed bandwidth h_2. The estimator $\hat{\imath}_0$ is integer-valued and may differ in absolute value from the theoretical (random) i_0 by 0, 1, 2, $\cdots$. Of course we would like the estimator $\hat{\imath}_0$ to be equal to i_0 with high probability. For choosing the bandwidth h_2 we propose a bootstrap procedure to estimate $P(\hat{\imath}_0 - i_0 = 0)$ for a large set of candidate h_2 values, denoted as $h_{2,j}$, $j = 0, \cdots, H$. We then select that bandwidth value h_2 for which the bootstrap estimate of the probability $P(\hat{\imath}_0 - i_0 = 0)$ is largest. The bootstrap algorithm for estimating this probability reads as follows (see also Gijbels, Hall and Kneip (1996)).

Step 1: Estimation of g and computation of residuals.
Let $\hat{x}_0 = \frac{1}{2}\left(X_{\hat{\imath}_0} + X_{\hat{\imath}_0+1}\right)$ denote the estimator introduced in Section 2.2. Using local linear regression (see for example Fan and Gijbels (1996)), with cross-validation bandwidth selector, we construct $\hat{g}$ on $[0, \hat{x}_0]$ and $[\hat{x}_0, 1]$. We define $\tilde{\varepsilon}_i = Y_i - \hat{g}(X_i)$ for $i = 1, \cdots, n$, and $\bar{\varepsilon}$ the mean of $\tilde{\varepsilon}_i$. Finally, we put $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \bar{\varepsilon}$, the centralized estimated residuals.

Step 2: Monte Carlo simulation.
Conditional on the observed sample $\chi = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, we consider $\varepsilon_1^*, \ldots, \varepsilon_n^*$ a resample drawn randomly with replacement from the set $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. We define

$$Y_i^* = \hat{g}(X_i) + \varepsilon_i^*, \qquad i = 1, \cdots, n. \tag{2.3}$$

Then $\chi^* = \{(X_1, Y_1^*), \cdots, (X_n, Y_n^*)\}$ is the bootstrap version of χ.

Step 3: Determination of the bootstrap probability.
Using the method described in Section 2.2, we compute the analogue $\hat{\imath}_0^*$ and $\hat{x}_0^* = \frac{1}{2}\left(X_{\hat{\imath}_0^*} + X_{\hat{\imath}_0^*+1}\right)$ of $\hat{\imath}_0$ and $\hat{x}_0$ for the resample χ^* rather than the sample χ. With B bootstrap replications, we have B values of $\hat{\imath}_0^*$, denoted by $\hat{\imath}_0^{*b}$, $b = 1, 2, \cdots, B$, and we

evaluate the discrete probability $P(\hat{\imath}_0^* - \hat{\imath}_0 = 0|\chi)$ via

$$\frac{1}{B}\sum_{b=1}^{B} \#\{b : \hat{\imath}_0^{*b} = \hat{\imath}_0\}.$$

With this bootstrap algorithm we have a data-driven procedure for estimating the jump discontinuity in the derivative curve: h_1 is automatically chosen as indicated above, and the bandwidth h_2 in the least-squares step is taken to be that bandwidth from the set of possible bandwidths for which the bootstrap estimate of the probability $P(\hat{\imath}_0 - i_0 = 0)$ is largest.

Note that with this data-driven procedure we opted for allowing for two possibly different bandwidths in the two steps of the estimation method. Alternatively, one could consider taking the same bandwidths in the diagnostic and the least-squares step, and select that single bandwidth via the bootstrap selection procedure described above. Both alternative data-driven procedures have been evaluated via extensive simulation by Gijbels and Goderniaux (2000) in the context of detecting jump discontinuities in the regression function itself. The conclusion was that the more general and more flexible two bandwidths option performs slightly better and hence we opted for this here too.

Note also that for estimating g in the bootstrap algorithm we estimate the jump point of the first derivative, and then estimate via local linear fitting the function g on each of the two intervals separated by the estimated jump point. This is surely not the most efficient way for estimating an unknown regression curve, knowing that the curve shows a jump point in its derivative. See also Section 6 for some further discussion.

2.4 The case of more than one jump discontinuity

The generalization to more than one jump discontinuity is quite straightforward, at least to some extent. Assume that there are ν jump discontinuities in the first derivative. One would then look for the ν local maxima in the diagnostic function $|D(x, h_1)|$ defined in (2.1), and would improve upon initial estimates of the locations by taking a small interval around each initial estimate and fitting via least-squares piecewise linear functions on each interval. Although this seems straightforward, we are not recommending to use this in practice. The reason is that the identification of local maxima corresponding to the jump discontinuities can be somewhat cumbersome. This identification problem can already be an issue when dealing with one single jump discontinuity, and is even more an issue when dealing with more than one jump discontinuity. In short, the identification problem might occur when handling functions g for which the derivative shows steep decreasing or increasing parts, which can blur the detection of jump points. It should be noted that any estimation method will show difficulties with such cases, not only the diagnostic function considered in this paper. For the specific case of our diagnostic function, this would mean that the diagnostic function could achieve its maximum at such points of steep increase or decrease. The solution to the problem is not so difficult though, and consists of a carefully-designed iterative algorithm which tracks back the maximum (or local maxima). For details of this identification problem and the iteration algorithm as a remedy for it, we refer the readers to Gijbels, Hall and Kneip (1996, 1999) or Gijbels and Goderniaux (2000).

In case of more than one jump discontinuity, the above mentioned identification problem can even be more severe, and hence in this case we recommend to use as a default the iterative algorithm which locates the local maxima of the diagnostic function associated to the jump discontinuities.

3 Generalization for detecting jump discontinuities in higher order derivatives

We now discuss how to generalize the data-driven procedure for detecting discontinuities in the k^{th} derivative of the regression function. Suppose that the function g is such that its k^{th} derivative is continuous except at a finite number of unknown points. For simplicity we restrict to the case of equally-spaced design points and as a diagnostic function we use the $(k+1)^{th}$ derivative of the numerator of a Nadaraya-Watson kernel estimator:

$$D(x, h_1) = \frac{1}{nh_1^{k+2}} \sum_{i=1}^{n} K^{(k+1)}\{(x - X_i)/h_1\} Y_i,$$

where K is a compactly supported ($[-v, v]$) kernel function that is $(k+1)$ times differentiable. For all other design cases it is preferable to work with the $(k+1)^{th}$ derivative of the Nadaraya-Watson kernel estimator or any other consistent nonparametric estimator of the unknown regression function. For simplicity we explain the procedure in case of only one discontinuity. Generalizations to more than one jump point are dealt with as indicated in Section 2.4 using the appropriate diagnostic function.

An initial estimator of the location x_0 of the jump discontinuity in the k^{th} derivative of the regression function is given by

$$\tilde{x}_0 = \text{argmax}_{x \in]vh_1, 1-vh_1[} |D(x, h_1)|.$$

This rough estimator is then refined by considering a small interval around this initial estimator and fitting via the least-squares method a piecewise k^{th} order polynomial in this interval. More precisely, assuming that the jump point x_0 falls between X_{i_0} and X_{i_0+1}, and denoting by $\{i_1, i_1+1, \cdots, i_2\}$ the set of all indices i for which $X_i \in [\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$, then we estimate i_0 by $\hat{\imath}_0$ the minimizer of the following sum of squares:

$$\sum_{i=i_1}^{i_0} \left\{ Y_i - \sum_{j=0}^{k} a_j X_i^j \right\}^2 + \sum_{i=i_0+1}^{i_2} \left\{ Y_i - \sum_{j=0}^{k} b_j X_i^j \right\}^2 .$$

The coefficients $a_j, b_j, j = 0, \cdots, k$ are the usual coefficients from a global least-squares fit with a polynomial of order k, and are given by

$$a = (a_0, \cdots, a_k)^T = (X^T X)^{-1} X^T Y \qquad \text{and} \qquad b = (b_0, \cdots, b_k)^T = (X'^T X')^{-1} X'^T Y' ,$$

where the superscript T denotes the transposed on a vector or matrix, and with

$$X = \begin{pmatrix} 1 & X_{i_1} & X_{i_1}^2 & \ldots & X_{i_1}^k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i_0} & X_{i_0}^2 & \ldots & X_{i_0}^k \end{pmatrix} \qquad Y = \begin{pmatrix} Y_{i_1} \\ \vdots \\ Y_{i_0} \end{pmatrix}$$

and

$$X' = \begin{pmatrix} 1 & X_{i_0+1} & X_{i_0+1}^2 & \ldots & X_{i_0+1}^k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i_2} & X_{i_2}^2 & \ldots & X_{i_2}^k \end{pmatrix} \qquad Y' = \begin{pmatrix} Y_{i_0+1} \\ \vdots \\ Y_{i_2} \end{pmatrix}.$$

The final estimator of x_0 is then defined as $\hat{x}_0 = \frac{1}{2}\left(X_{\hat{\imath}_0} + X_{\hat{\imath}_0+1}\right)$.

To choose the two different bandwidths involved in this two-steps estimation method we use the algorithm described in Section 2.3.

4 Estimating the number of discontinuities in a derivative

An important issue in discontinuity detection is to determine the number of discontinuities, which is often unknown in practice. Müller and Stadtmüller (1999) propose a cross-validation criterion to estimate the number of discontinuities that appear in the regression function itself. In this section we generalize this method to estimate the number of discontinuities appearing in the k^{th} derivatives of the regression function.

For ease of comprehension we first explain the method when the jump points occur in the first derivative (i.e. $k = 1$). In this case it seems natural to define the cross-validation quantity as follows:

$$\mathrm{CV}^{(1)}(\nu) = \frac{1}{n-1}\sum_{i=1}^{n-1}\left\{\frac{Y_{i+1}-Y_i}{X_{i+1}-X_i} - \hat{g}^{(1)}_{-(i,i+1),\nu}(X_i^{(1)})\right\}^2$$

where ν represents the number of discontinuities, $X_i^{(1)} = (X_{i+1}+X_i)/2$ and $\hat{g}^{(1)}_{-(i,i+1),\nu}(X_i^{(1)})$ is the leave-(2)-out kernel estimator of the first derivative of g, the regression function, based on the data $(X_1, Y_1), ..., (X_{i-1}, Y_{i-1}), (X_{i+2}, Y_{i+2}), ..., (X_n, Y_n)$ and adapted to the ν estimated jump points of the derivative function. This derivative estimator is obtained by carrying out local polynomial fitting of order 2 on all $\nu + 1$ intervals separated by the estimated jump points, using cross-validated bandwidths (adapted to estimation of the derivative function, see Müller, Stadtmüller and Schmitt (1987)). Note that $\frac{Y_{i+1}-Y_i}{X_{i+1}-X_i}$ represents the slope of the line between Y_{i+1} and Y_i. We compare this slope with an estimator of the first derivative of the regression function at the point $(X_{i+1} + X_i)/2$. It is clear that we want the cross-validation quantity to be as small as possible.

This method generalizes easily to the case of jump points occuring in the k^{th} derivative of the regression function. Define $X_i^{(0)} = X_i$ and $\delta_i^{(0)} = Y_i$ for $i = 1, ..., n$ and put

$$X_i^{(k)} = \frac{1}{2}(X_{i+1}^{(k-1)} + X_i^{(k-1)}), \qquad \delta_i^{(k)} = \frac{\delta_{i+1}^{(k-1)} - \delta_i^{(k-1)}}{X_{i+1}^{(k-1)} - X_i^{(k-1)}}$$

for $i = 1, ..., n - k$. The proposed generalization of the above cross-validation quantity is then

$$\mathrm{CV}^{(k)}(\nu) = \frac{1}{n-k}\sum_{i=1}^{n-k}(\delta_i^{(k)} - \hat{g}^{(k)}_{-(i,i+k),\nu}(X_i^{(k)}))^2 ,$$

where $\hat{g}^{(k)}_{-(i,i+k),\nu}(X_i^{(k)})$ is the leave-$(k+1)$-out kernel estimator of the k^{th} derivative function based on the data $(X_1, Y_1), ..., (X_{i-1}, Y_{i-1}), (X_{i+k+1}, Y_{i+k+1}), ..., (X_n, Y_n)$ and obtained via local polynomial approximation of order $(k + 1)$ on each of the $\nu + 1$ intervals defined by the ν estimated jump points of the k^{th} derivative function, and using cross-validation bandwidth selectors adapted to the estimation of derivative curves.

Such a type of cross-validation quantity has been proposed by Müller, Stadtmüller and Schmitt (1987) in the context of bandwidth selection for estimating the k^{th} derivative of a regression function.

So in practice we calculate the cross-validation quantity for each pre-specified number of discontinuities and then choose that number (of discontinuities) which corresponds with the smallest cross-validation value. To estimate the location of the pre-specified ν discontinuities, $\nu = 0, 1, 2, ...$ we propose to use the fully data-driven bootstrap procedure adapted to the derivative case. If two estimated locations are too close we consider that this value of ν is not a possible value of the number of discontinuities. Finally we estimate the number of discontinuities appearing in the k^{th} derivative by

$$\hat{\nu} = \text{argmin}_{\nu \in \{0,1,\ldots\}} \text{CV}^{(k)}(\nu).$$

Note that this data-driven procedure is developed for detecting jump points in a derivative function of pre-specified order. See Section 6 for a brief discussion on discontinuities appearing in several derivatives.

5 Simulation studies and applications

5.1 Simulation study

In this section we evaluate via a simulation study the data-driven estimation procedure. We consider the regression functions

$$g_1(x) = \begin{cases} 2x + 1 & \text{if } x \in [0, 0.5] \\ -2x + 3 & \text{if } x \in]0.5, 1] \end{cases}$$

$$g_2(x) = \begin{cases} 10x^2 & \text{if } x \in [0, 0.5] \\ -20/7\ x^3 + 20/7 & \text{if } x \in]0.5, 1] \end{cases}$$

$$g_3(x) = \begin{cases} 12x^2 - 2x + 2 & \text{if } x \in [0, 0.5] \\ -12x^2 + 22x - 4 & \text{if } x \in]0.5, 1]\ , \end{cases}$$

and work with fixed equidistant design, $x_i = i/n$ for $i = 1, \cdots, n$. The errors ε_i were taken to be Gaussian with variances $\sigma^2 = 0.1$ or 0.5. We present simulation results for sample sizes $n = 100$ or 200. In the upper panels of Figure 5.1 we depict the true regression functions g_1, g_2 and g_3 with typical simulated data sets for sample size $n = 100$

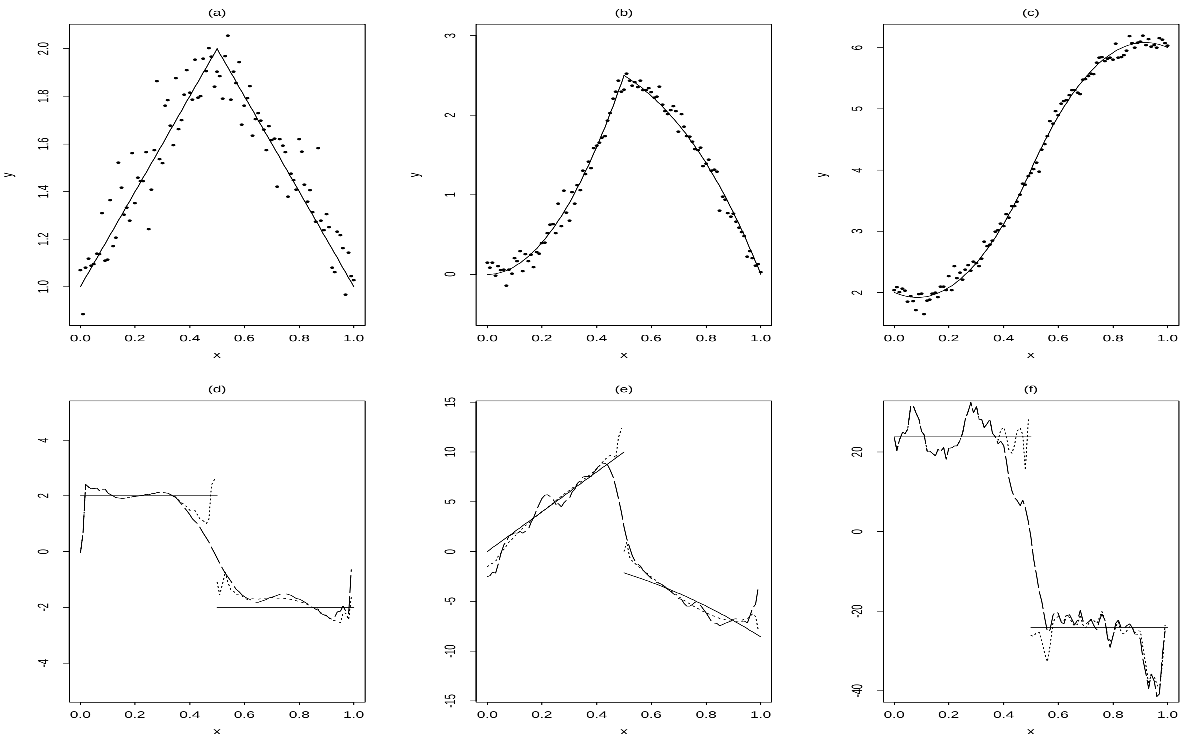

Figure 5.1: *Upper panels: The true regression functions (solid curves) with a typical simulated data set of size* $n = 100$ *and variance* $\sigma^2 = 0.01$*: Regression functions (a)* g_1*; (b)* g_2*; and (c)* g_3*. Lower panels: the true* k^{th} *derivative (solid curve) for the three functions along with a smooth local quadratic (for (d) and (e)) and local cubic (for (f)) estimator (long-dashed curve) and an unsmooth estimator (dotted curve). Regression function (d)* g_1*,* $k = 1$*; (e)* g_2*,* $k = 1$*; and (f)* g_3*,* $k = 2$*.*

and $\sigma^2 = 0.01$. Note that both functions g_1 and g_2 have a single jump discontinuity at $x_0 = 0.5$ appearing in the first derivative. The size of the jump is -4 for g_1 and $-85/7$ for g_2. The function g_3 presents a single jump discontinuity of size -48 at the point $x_0 = 0.5$ in the second derivative. In the lower panels of Figure 5.1 we present the true derivative functions $g_1^{(1)}$, $g_2^{(1)}$ and $g_3^{(2)}$ as solid curves along with smooth local quadratic (respectively cubic) estimators as long-dashed curves and estimators adapted to the estimated jump points as dotted curves. The adapted estimates were obtained by local quadratic fitting for the functions g_1 and g_2 and by local cubic fitting for the function g_3 on the two intervals separated by the estimated jump point. Figure 5.1 is based on simulations from an error with relatively small variance. This small error variance is only considered for producing

Figure 5.1 (especially focusing on (d)—(f)), in order to obtain estimated derivative curves that present nice visually. Recall that estimation of derivatives curves is more difficult than estimation of the regression function itself. For all other simulations, focusing on the estimation of the jump points, we consider larger error variances. Note, from Figure 5.1 (a)—(c), that even with such a small error variance, it is hard to tell from the data what is happening (smooth or non-smooth) with the first (second) derivative.

5.1.1 Estimation of the localisation of the jump

For the diagnostic function in (2.1), we used a standard Gaussian kernel. In all simulation studies we considered 1000 simulations and the number of bootstrap replicates was $B = 2000$. For determining h_1, the smoothing parameter of the diagnostic function, we search over the set of bandwidths $h_{1,i} = h_0 r^i$, for $i = 0, 1, 2, \cdots$, with $r = 0.9$ and $h_0 = 0.2$ (respectively $h_0 = 0.1$) when we study the functions g_1 and g_3 (respectively g_2). For the set of potential bandwidths $h_{2,j}$, $j = 0, \cdots, H$, in the least-squares step we took $h_{2,j} = 0.03 + 0.015j$ for $j = 0, \cdots, 7$.

Table 5.1: *Simulation results for the function g_1.*

	$\sigma^2 = 0.1$		$\sigma^2 = 0.5$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
% in $[0.0, 0.15[$	0	0	0	0
% in $[0.15, 0.25[$	0	0	0.5	0.2
% in $[0.25, 0.35[$	0.1	0	3.2	1.4
% in $[0.35, 0.45[$	9.7	8.1	21.1	18.2
% in $[0.45, 0.55[$	82.5	81.6	52.1	57.4
% in $[0.55, 0.65[$	7.7	10.3	18.9	20.7
% in $[0.65, 0.75[$	0	0	3.6	2.1
% in $[0.75, 0.85[$	0	0	0.6	0
% in $[0.85, 1.0[$	0	0	0	0
mean of $\hat{x}_0$	0.503200	0.501500	0.503630	0.502155
sd of $\hat{x}_0$	0.041326	0.038776	0.084613	0.068152

In Table 5.1 we summarize the simulation results for the function g_1. Presented are

Table 5.2: *Simulation results for the function g_2.*

	$\sigma^2 = 0.1$		$\sigma^2 = 0.5$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
% in $[0.0, 0.15[$	0	0	0	0
% in $[0.15, 0.25[$	0	0	0	0
% in $[0.25, 0.35[$	0	0	0.1	0
% in $[0.35, 0.45[$	3	2.7	13.2	15.9
% in $[0.45, 0.55[$	97	97.3	81.1	81.7
% in $[0.55, 0.65[$	0	0	4.7	2.3
% in $[0.65, 0.75[$	0	0	0.8	0.1
% in $[0.75, 0.85[$	0	0	0.1	0
% in $[0.85, 1.0[$	0	0	0	0
mean of $\hat{x}_0$	0.496810	0.499380	0.491730	0.488905
sd of $\hat{x}_0$	0.017943	0.014585	0.041763	0.035615

Table 5.3: *Simulation results for the function g_3.*

	$\sigma^2 = 0.1$		$\sigma^2 = 0.5$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
% in $[0.0, 0.15[$	0	0	0	0
% in $[0.15, 0.25[$	0	0	0.1	0
% in $[0.25, 0.35[$	0.1	0.1	0.6	1.1
% in $[0.35, 0.45[$	2.3	8.4	7.9	16.5
% in $[0.45, 0.55[$	80	85.3	55.7	58.9
% in $[0.55, 0.65[$	14.4	5.8	20.4	17.4
% in $[0.65, 0.75[$	2	0	7.9	1.6
% in $[0.75, 0.85[$	0	0.4	2.3	4.5
% in $[0.85, 1.0[$	1.2	0	5.1	0
mean of $\hat{x}_0$	0.533650	0.496960	0.560300	0.509960
sd of $\hat{x}_0$	0.073594	0.044082	0.121593	0.088881

the percentages that the estimated values $\hat{x}_0$ fall in the specified intervals. In the two bottom lines of the table we also list the means and standard deviations of $\hat{x}_0$ across the 1000 simulations. Table 5.2 is similar to Table 5.1, and reports on the simulation results for the function g_2. If we compare the results of the two functions that present a jump in the first derivative we can see that the results for the function g_1 are not as good as those for the function g_2. This is related to the fact that the size of the jump in the derivative function is smaller for the function g_1. Table 5.3 summarizes the results for the function g_3.

Table 5.4: *Simulation results.*

			0	1	2	3	4
g_1	$n = 100$	$\sigma^2 = 0.1$	0	78	12	10	0
one discontinuity in		$\sigma^2 = 0.5$	0	73	16	10	1
the first derivative	$n = 250$	$\sigma^2 = 0.1$	0	83	8	9	0
		$\sigma^2 = 0.5$	0	79	11	9	0
g_2	$n = 100$	$\sigma^2 = 0.1$	1	76	15	6	2
one discontinuity in		$\sigma^2 = 0.5$	1	73	17	8	1
the first derivative	$n = 250$	$\sigma^2 = 0.1$	0	86	4	9	1
		$\sigma^2 = 0.5$	0	83	5	12	0
g_3	$n = 100$	$\sigma^2 = 0.1$	2	64	22	8	4
one discontinuity in		$\sigma^2 = 0.5$	2	61	26	5	7
the second derivative	$n = 250$	$\sigma^2 = 0.1$	0	76	7	12	5
		$\sigma^2 = 0.5$	1	70	23	2	4

5.1.2 Estimation of the number of discontinuities

From each of the functions g_1, g_2 and g_3 we also applied the cross-validation method for determining the number of discontinuities. Here we used 100 simulations and $B = 1000$ bootstrap samples. Table 5.4 summarizes the simulation results for the three functions for sample sizes $n = 100$ and $n = 250$ and σ^2 equal to 0.1 and 0.5. Presented are the frequences (out of 100) that the estimated values $\hat{\nu}$ correspond to the specified values. As an illustration we show in Figure 5.2 the cross-validation quantity as function of ν for a simulated data set of size $n = 100$ generated from the function g_1 with $\sigma^2 = 0.1$. We conclude that for this sample the number of discontinuities should be taken to be one.

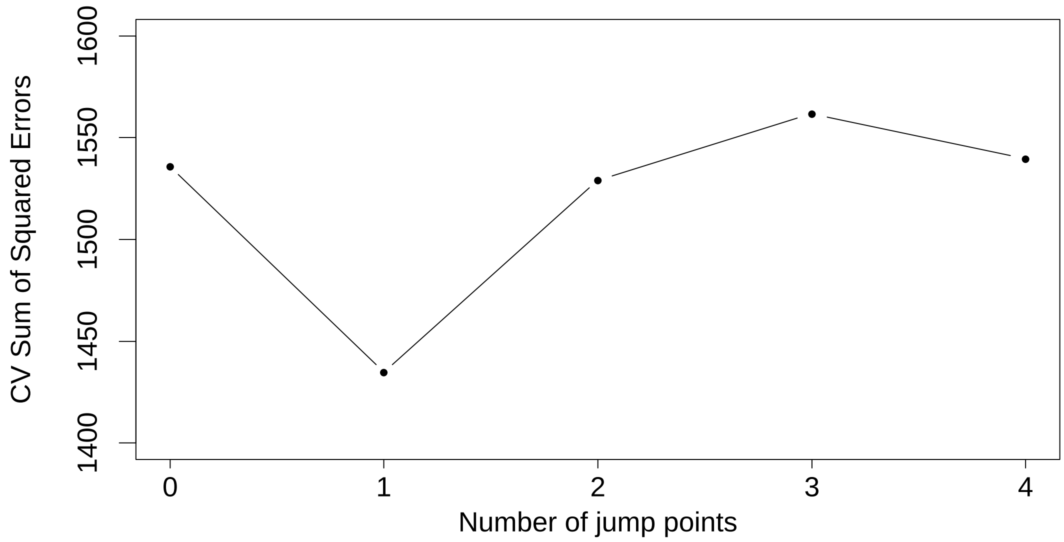

Figure 5.2: *Cross-validation sum of squares as function of the number of jump points for a simulated data set of size $n = 100$ from the function g_1 with $\sigma^2 = 0.1$.*

5.2 Applications

5.2.1 The Motorcycle data

For the Motorcycle data we applied the data-driven method to search for jump discontinuities in the first derivative of the regression function, since we aim at finding changes in direction in the acceleration. For the smoothing parameter h_1 in the diagnostic step we searched over the set $h_{1,i} = h_0 r^i$ with $h_0 = 7$ and $r = 0.9$. For the interval length in the least-squares step we used the set of potential bandwidths $h_{2,j} = 3.5 + 1.0j$ for $j = 0, 1, ..., 8$. The cross-validation quantity as function of ν is represented in Figure 5.4, and from this we fix the number of discontinuities to be three. The data-driven estimation method provided the estimations 14.2, 24.1 and 32.4 for the three change points. These results agree with those obtained by Speckman (1995) who uses a semiparametric change point method to identify the number and the locations of the change points. Figure 5.3 shows the data along with an adaptive local linear fit.

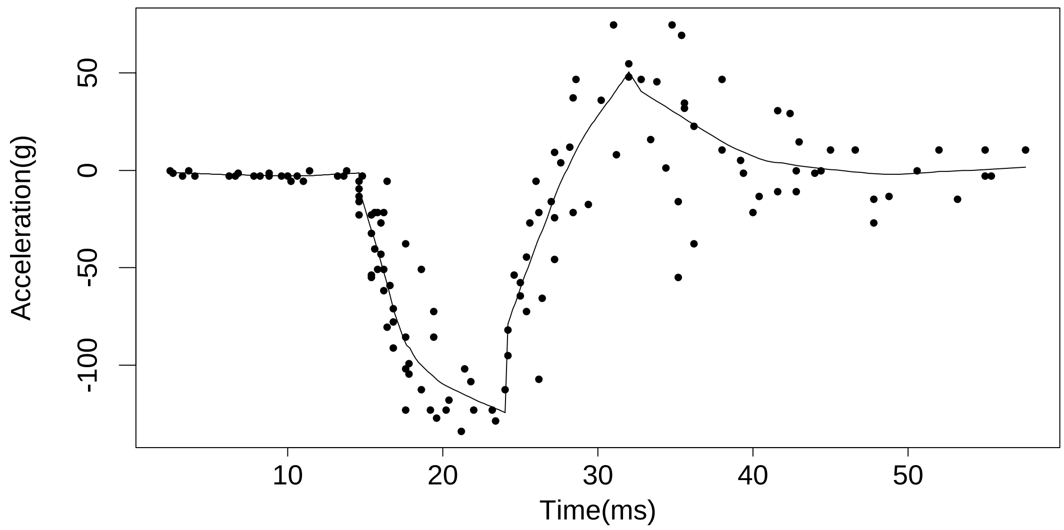

Figure 5.3: *The motorcycle data (points) with a local linear fit adapted to change-points.*

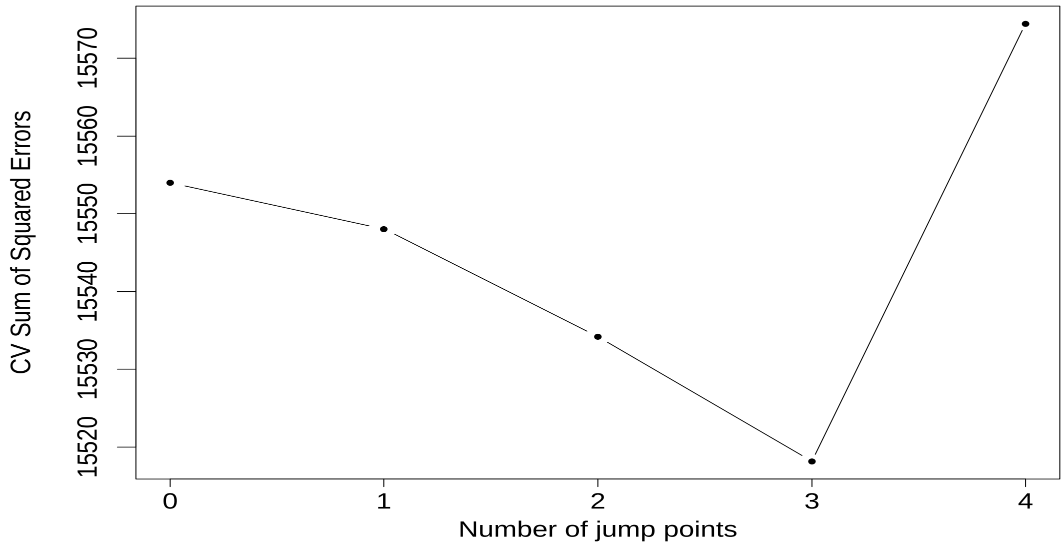

Figure 5.4: *Cross-validation sum of squares as function of the number of jump points for the motorcycle data.*

5.2.2 The winning speed data

We applied the data-driven method to the winning speed data set described in the introduction. For choosing h_1, the smoothing parameter for the diagnostic function, we search

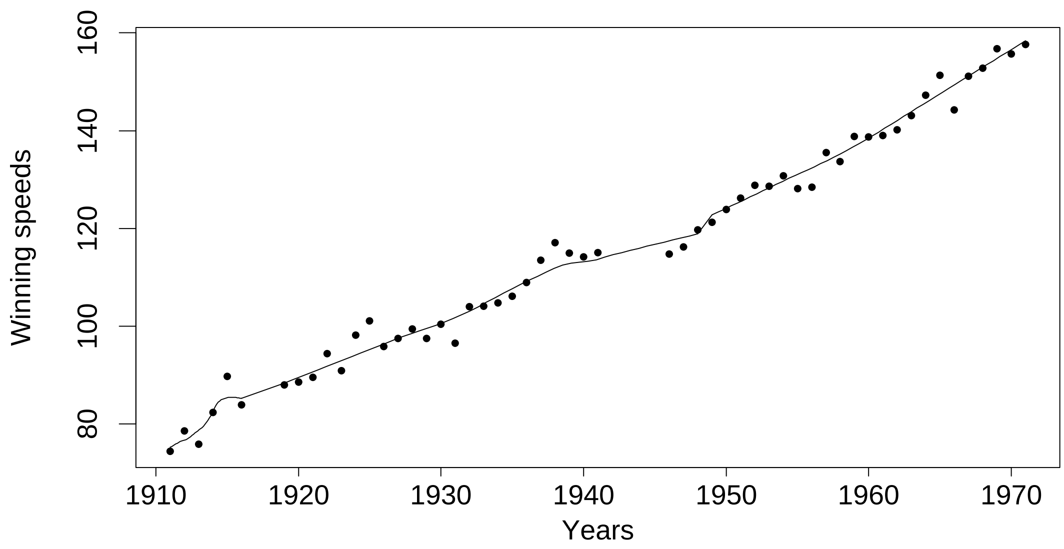

Figure 5.5: *The winnings speeds for the yearly Indianapolis 500 race (points) with a local linear fit adapted to the estimated change-points.*

over the set $h_{1,i} = h_0 r^i$ with $h_0 = 21$ and $r = 0.9$. For h_2, the length of the interval used for the least-squares step we used the set of potential bandwidths $h_{2,j} = 3.0 + 1.0j$ for $j = 0, 1, ..., 7$. The data-driven method tells us that there are two discontinuities located at 1917.5 and 1948.5 respectively with estimated sizes of jumps -0.005 and -0.001 respectively. Figure 5.5 shows the data along with a local linear estimate adapted to the two estimated jump discontinuities in the third derivative. For this data set we had problems applying the cross-validation criterion to estimate the number of discontinuities occuring in the third derivative due to the fact that there are very few data points before 1917.5.

6 Discussion

6.1 About estimation of the regression function

The estimation of the regression function itself was done as follows. The ν estimated jump discontinuities in the k^{th} derivative define $\nu + 1$ intervals and on each of these intervals we obtain the local linear fit. These fits are then joined together to get the global estimator for the regression function. This estimation method does not use the fact that the jump discontinuities occur in the k^{th} derivative of the function. One way to use more directly the information about jump discontinuities in the k^{th} derivative is to fit splines of order k with knots at the estimated ν jump discontinuities. In this way, the resulting estimator would have continuous first $k-1$ derivatives. We did not investigate further on this issue.

6.2 Curves with discontinuities of various kinds

A regression function of course can have change points of various kinds: it can have a jump discontinuity at a certain point and an abrupt change in direction at another point. In other words, it is possible that one has to deal with the case where there are jump discontinuities in the function itself but also in (for example) the first derivative. The data-driven procedure can also be used for such cases. One would then start by estimating the number and locations of jump discontinuities in the regression function itself, followed by estimating the number and locations of jump discontinuities in the derivative, and so on. This would then result into estimates of all the jump discontinuities occuring in any k^{th} order derivative ($k = 0, 1, 2, ...$). Say that this results into locating ℓ abrupt changes. Consider the $\ell+1$ intervals separated by these estimated abrupt changes. On each interval we would then obtain the local linear estimates using a cross-validated bandwidth.

Acknowledgements

Financial support of the 'Projet d'Actions de Recherche Concertées', No. 98/03–217, from the Belgian government is gratefully acknowledged.

References

Cline, D.B.H., Eubank, R. and Speckman, P. (1995). Nonparametric estimation of regression curves with discontinuous derivatives. *Journal of Statistical Research*, **29**, 17–30.

Eubank, R.L. and Speckman, P.L. (1994). Nonparametric estimation of functions with jump discontinuities. In *Change-point Problems*, IMS–Lecture Notes, Eds. E. Carlstein, H.-G. Müller and D. Siegmund, **23**, 130–143.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, New York.

Gijbels, I. and Goderniaux, A-C. (2000). Bandwidth selection for change point estimation in nonparametric regression. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0024. http://www.stat.ucl.ac.be

Gijbels, I., Hall, P. and Kneip, A. (1996). Interval and band estimation for curves with jumps. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 9609. http://www.stat.ucl.ac.be

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *The Annals of the Institute of Statistical Mathematics*, **51**, 231–251.

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

Horváth, L. and Kokoszka, P. (1997). Change-point detection with nonparametric regression. *Technical report*, University of Liverpool.

Jose, C.T. and Ismail, B. (1997). Estimation of jump points in nonparametric regression through residual analysis. *Communications in Statistics, Part A-Theory and Methods*, **26**, 2583–2607.

McDonald, J.A. and Owen, A.B. (1986). Smoothing with split linear fits. *Technometrics*, **28**, 195–208.

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

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *The Annals of Statistics*, **27**, 299–337.

Müller, H.-G., Stadtmüller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, **74**, 743–749.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and Applications*, **9**, 141–142.

Poirier, D.J. (1973). Piecewise regression using cubic splines. *Journal of American Statistical Association*, **68**, 515–524.

Qiu, P. and Yandell, B. (1998). A local polynomial jump-detection algorithm in nonparametric regression. *Technometrics*, **40**, 141–152.

Raimondo, M. (1998). Minimax estimation of sharp change points. *The Annals of Statistics*, **26**, 1379–1397.

Schmidt, G., Mattern, R. and Schüler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact. EEC Research Program on Biomechanics of Impacts. Final Report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Germany.

Speckman, P.L. (1994). Detection of change-points in nonparametric regression. *Unpublished manuscript.*

Speckman, P.L. (1995). Fitting curves with features: semiparametric change-point methods. *Computing Science and Statistics*, **26**, 257–264.

Spokoiny, V.G. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, **26**, 1356–1378.

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

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā* Series A, **26**, 359–372.

Wu, J.S. and Chu, C.K. (1993). Modification for boundary effects and jump points in nonparametric regression. *Journal of Nonparametric Statistics*, **2**, 341–354.