Bandwidth selection for change point estimation in nonparametric regression

Irène Gijbels and Anne-Cécile Goderniaux[1]

October 2000

Abstract

Nonparametric estimation of abrupt changes in a regression function involves the choice of smoothing (bandwidth) parameters. The performance of estimation procedures heavily depends on this choice. So far little attention has been paid to the crucial issue of choosing appropriate bandwidth parameters in practice. In this paper we propose a bootstrap procedure for selecting the bandwidth parameters in a nonparametric two-steps estimation method. This results in a fully data-driven procedure for estimating a finite number of change points in a regression function. The performance of the data-driven procedure is evaluated via a simulation study, which reveals that the procedure performs quite well. As an illustration we apply the procedure to some real data.

Key words: bandwidth, bootstrap, discontinuity points, cross-validation, least-squares fitting.

1 Introduction

There is a vast literature on estimation of smooth regression functions. In many applications though the regression function is smooth except at a finite number of locations where

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

it shows, for example, a jump discontinuity. Examples where this occurs can be found in physical sciences, engineering, medicine and economics, among others. See the recent paper by Müller and Stadtmüller (1999) for more background information and many examples of data applications. Estimation of an unknown regression function with a finite number of discontinuities is usually done in a few stages. In a first stage the locations of the possible jump discontinuities are estimated. Once these locations are estimated one turns to the second stage which is the actual estimation of the regression curve. A common procedure here is to fit nonparametrically smooth curves to the left and to the right of the estimated locations of jump discontinuities, relying on techniques available for estimating smooth curves. So the first stage is the most interesting one from this point of view. Another approach has been followed by Kang, Koo and Park (2000) who estimate the locations of the jump discontinuities as well as the sizes of the jumps. Using these estimations they then adjust the data suitably and apply ordinary smoothing techniques to these adjusted data. So there are various aspects in estimating regression functions with jump discontinuities: estimation of the locations of the jump discontinuities, estimation of the jump sizes and finally estimation of the regression function itself.

The literature on nonparametric estimation of regression functions which are smooth except at a number of points is by now quite large. Kernel-based estimation methods have been studied by Canny (1986), Korostelev (1987), Hall and Titterington (1992), Müller (1992), Wu and Chu (1993a,b,c), Chu (1994), Speckman (1994), Eubank and Speckman (1994), Bunt, Koch and Pope (1995) and Kang, Koo and Park (2000), among others. Local-polynomial methods were used by McDonald and Owen (1986), Leclerc and Zucker (1987), Loader (1996), Horváth and Kokoszka (1997), Qui and Yandell (1998), Spokoiny (1998) and Hamrouni (1999), among others. References for spline-based methods include Laurent and Utreras (1986), Girard (1990) and Koo (1997). For wavelet-based methods see for example Mallat and Hwang (1992), Potier and Vercken (1994), Wang (1995), Raimondo (1998), Oudshoorn (1998) and Antoniadis and Gijbels (2000). Müller and Song (1997) and Gijbels, Hall and Kneip (1999) proposed two-steps procedures.

Two important issues arise when dealing with nonparametric estimation of a regression curve with jump discontinuities. A first issue is knowledge of the finite number of jump discontinuities. Often it is assumed that this number is known, whereas this is not always the case in applications. A second issue is that any nonparametric estimation method involves the choice of parameters, call them smoothing parameters, and that the performance of the estimation procedures heavily depends on the choice of these parameters. Hence it is very important to address the issue of how to choose these parameters in practice. Some attention has been paid to these issues in the literature. References dealing with the first issue include Yin (1988), Wu and Chu (1993a), Bunt, Koch and Pope (1998), Beunen (1998), Oudshoorn (1998), Müller and Stadtmüller (1999) and Antoniadis and Gijbels (2000). As far as we know little work has been done on the second issue. See for example Chu and Wu (1993c) and Spokoiny (1998) (for a theoretical contribution).

In this paper we will focus on estimation of the locations of the jump discontinuities, and in particular on the practical choice of the smoothing (bandwidth) parameters involved. The objective is to come up with an estimation procedure with a data-driven choice of the bandwidth parameters, which performs well in practice. We assume throughout the paper that the number of finite jump discontinuities is known. We use the two-steps estimation method proposed by Gijbels, Hall and Kneip (1999) for which it has been shown that the estimator for the location of a jump discontinuity achieves the optimal rate n^{-1}, where n is the sample size. Moreover the performance of this estimation method has been evaluated intensively via simulations, which showed a very good performance for a variety of regression functions. This simulation study also revealed once again that the choice of the bandwidth parameters are really crucial in practice. This in fact motivated the study presented here. We propose a bootstrap procedure for selecting the bandwidth parameters appearing in the estimation procedure. Some of the basic ingredients for this bootstrap procedure are similar to these used in Gijbels, Hall and Kneip (1996), in the context of interval and band estimation for curves with jumps. A theoretical justification of the bootstrap bandwidth selection procedure introduced here relies on theoretical results established in Gijbels, Hall and Kneip (1996). We tested our data-driven estimation procedure on various simulation models, and got for all a very good performance.

The paper is organized as follows. In Section 2 we briefly describe the estimation method and discuss the bandwidth parameters it involves. In Section 3 we introduce the bootstrap bandwidth selection procedure. A numerical study of the data-driven estimation method is presented in Section 4, along with an illustration with some real data. Some concluding remarks and a brief discussion on related research issues are provided in Section 5.

2 Estimation of jump discontinuities

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

$$Y_i = g(X_i) + \varepsilon_i, \qquad 1 \leq i \leq n. \tag{2.1}$$

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. The unknown function $g(\cdot)$ is smooth except at a (known) finite number of jump discontinuities. For ease of presentation we will explain the estimation and bootstrap selection procedure for the case of a single jump discontinuity in the regression function at the point $x_0 \in]0, 1[$. The case of more than one jump discontinuity will be addressed in Section 3.3.

2.2 Estimation procedure

We now briefly describe the two-steps estimation procedure for estimating the location x_0 of the jump discontinuity introduced by Gijbels, Hall and Kneip (1999). The first step of the procedure aims at obtaining a rough estimate of x_0, say $\tilde{x}_0$. The second step then consists of restricting to an interval around $\tilde{x}_0$, say $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$ with $h_2 > 0$ a bandwidth parameter, and using least-squares fitting in this interval to get an improved estimator $\widehat{x}_0$ of x_0. The first step, called the diagnostic step, is based on the derivative of the Nadaraya-Watson estimator of the regression function and involves a bandwidth parameter h_1.

2.2.1 Diagnostic step

A way to detect a jump discontinuity is by looking at locations with high derivatives. So a possible diagnostic function is the derivative of the Nadaraya-Watson estimator (see

Nadaraya (1964) and Watson (1964)), defined as

$$D(x,h_1) = \frac{\partial}{\partial x}\left(\frac{\sum_{i=1}^{n} K\{(x-X_i)/h_1\}Y_i}{\sum_{i=1}^{n} K\{(x-X_i)/h_1\}}\right) \tag{2.2}$$

where K is a compactly supported (with support $[-v,v]$), differentiable kernel function and $h_1>0$ is a bandwidth. A first rough estimator of x_0 is then given by

$$\tilde{x}_0 = \text{argmax}_{x\in]vh_1,1-vh_1[}|D(x,h_1)|. \tag{2.3}$$

The rate of convergence of this preliminary estimator $\tilde{x}_0$ is $n^{-1}(\log n)^{1/2}$. The second, least-squares, step will lead to a n^{-1} estimator. See Gijbels, Hall and Kneip (1999).

2.2.2 Least-squares step

We construct an interval concentrated around $\tilde{x}_0$, namely $[\tilde{x}_0-h_2,\tilde{x}_0+h_2]$, to which x_0 belongs with high probability. Denote by $\{i_1,i_1+1,\cdots,i_2\}$ the set of integers i for which $X_i\in[\tilde{x}_0-h_2,\tilde{x}_0+h_2]$. We fit a step function on the interval $[\tilde{x}_0-h_2,\tilde{x}_0+h_2]$ using least-squares. We estimate the jump discontinuity to occur between design points X_{i_0} and X_{i_0+1}, where i_0 is chosen to minimize the sum of squares

$$\sum_{i=i_1}^{i_0}\left\{Y_i-(i_0-i_1+1)^{-1}\sum_{j=i_1}^{i_0}Y_j\right\}^2+\sum_{i=i_0+1}^{i_2}\left\{Y_i-(i_2-i_0)^{-1}\sum_{j=i_0+1}^{i_2}Y_j\right\}^2. \tag{2.4}$$

Denote by $\hat{\imath}_0$ the minimizer of this sum of squares.

The final estimator of the jump discontinuity x_0 is then defined as the mid-point between $X_{\hat{\imath}_0}$ and $X_{\hat{\imath}_0+1}$:

$$\hat{x}_0 = \frac{1}{2}\left(X_{\hat{\imath}_0}+X_{\hat{\imath}_0+1}\right). \tag{2.5}$$

2.3 Choice of the bandwidth parameters

Each of the two steps in the estimation procedure involves a bandwidth parameter: in the diagnostic step we need to choose h_1 and the least-squares step requires the choice of h_2. Both bandwidth choices are in fact quite crucial for the performance of the procedure, as has been explored and discussed in detail in Gijbels, Hall and Kneip (1999).

We now provide some discussion on these bandwidth parameters, necessary for a good understanding of the sequel.

The preliminary estimator $\tilde{x}_0$ clearly depends on the bandwidth h_1 since the diagnostic function changes with h_1. At a jump discontinuity though the diagnostic function will always be large, and hence we search for locations that maximize $|D(x,h_1)|$ for a range of values of h_1. 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.

More precisely, to select the bandwidth h_1 we propose to use an iterative method. Let $h_{1,i} = h_0 r^i$, for $i = 0, 1, 2, \cdots$, be a range of decreasing values of h_1, with $h_0 > 0$ and $0 < r < 1$. For $i = 0, 1, \ldots$ we compute $\tilde{x}(h_{1,i})$, the value of x that maximizes $|D(x,h_{1,i})|$. Further, for $m = 0, 1, ...$ we compute

$$\tilde{x}_-(m) = \sup_{h_1 \in \{h_{1,0},\cdots,h_{1,m}\}} [\tilde{x}(h_1) - 2vh_1] \tag{2.6}$$

and

$$\tilde{x}_+(m) = \inf_{h_1 \in \{h_{1,0},\cdots,h_{1,m}\}} [\tilde{x}(h_1) + 2vh_1] \tag{2.7}$$

and verify that $\tilde{x}_-(m) \le \tilde{x}_+(m)$. We stop the iteration when this relation is no longer correct for the first time, say when $m = m_1$. We then choose the bandwidth $\tilde{h}_1 = h_{1,m_1-1}$ and define the preliminary estimator for x_0 as

$$\tilde{x}_0 = \tilde{x}(\tilde{h}_1). \tag{2.8}$$

In other words we choose the smallest bandwidth, denoted by $\tilde{h}_1$, such that the relation $\tilde{x}_-(m) \le \tilde{x}_+(m)$ is still correct. The preliminary estimator is the value of x at which $|D(x,\tilde{h}_1)|$ is maximal. This procedure allows to control that the different estimations $\tilde{x}(h_{1,i})$ for $i = 0, 1, 2, \cdots$ stay always in the same neighbourhood. We stop the iteration when we obtain an estimation $\tilde{x}(h_{1,m})$ that falls outside this neighbourhood. This mean that $\tilde{x}(h_{1,m})$ is too different from the other estimations $\tilde{x}(h_{1,i})$ and that the bandwidth $h_{1,m}$ is too small and results in a diagnostic curve with a lot of artificial peaks.

The importance of the bandwidth h_2 used in the second step is clear. Since we approximate the unknown function g by a step function using least-squares, the choice of the interval where we carry out this approximation, namely $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$ is crucial.

If this interval is taken large and there g differs a lot from a step function, then a local constant approximation may be very poor. As a result, the least-squares estimation on the large interval $[\tilde{x}_0 - h_2, \tilde{x}_0 + h_2]$ leads to a bad estimation of i_0 and hence of x_0. There is thus interest to choose the interval as small as possible, but not too small so that there are still enough data points in the interval for a reasonable least-squares fit. In the next section we propose a bootstrap procedure for choosing the bandwidth h_2.

3 Data-driven estimation procedures

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. 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. Now the idea is to use a bootstrap procedure to estimate the probability $P(\hat{\imath}_0 - i_0 = 0)$ and to select the bandwidth for which this estimated probability is maximal. We first describe in detail the bootstrap algorithm for estimating the probability $P(\hat{\imath}_0 - i_0 = 0)$, and then specify the actual bandwidth selection. Examples of implementations are provided in Section 4.

3.1 Bootstrap algorithm

Step 1: Estimation of g and computation of residuals.
Let $\hat{x}_0 = \frac{1}{2}(X_{\hat{\imath}_0} + X_{\hat{\imath}_0+1})$ denote the estimator introduced in Section 2.2. Using local linear regression (see for example Fan and Gijbels (1996)), 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

$$Y_i^* = \hat{g}(X_i) + \varepsilon_i^*, \qquad i = 1, \cdots, n. \tag{3.1}$$

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

$$P(\hat{\imath}_0^* - \hat{\imath}_0 = 0|\chi) \tag{3.2}$$

via

$$\frac{1}{B}\sum_{b=1}^{B} \#\{b : \hat{\imath}_0^{*b} = \hat{\imath}_0\}. \tag{3.3}$$

Using the above procedure one can of course evaluate all discrete probabilities $P(\hat{\imath}_0^* - \hat{\imath}_0 = k|\chi)$, $k = 0, -1, 1, -2, 2, \cdots$. These have been studied in Gijbels, Hall and Kneip (1996), where it was shown that under some regularity conditions and for fixed design $\sup_{k=0,-1,1,\ldots} \mid P(\hat{\imath}_0^* - \hat{\imath}_0 = k|\chi) - P(\hat{\imath}_0 - i_0 = k) \mid \longrightarrow 0$ in probability. Moreover, the same holds for random design. The only difference is that probabilities have to be considered conditionally on the realized design $X_1, \cdots, X_n$.

3.2 Bootstrap bandwidth selection methods

We now discuss how the above bootstrap algorithm for estimating $P(\hat{\imath}_0 - i_0 = 0) \equiv p_0$ can be applied for the bandwidth selection purpose. A first direct way consists of simply choosing the bandwidth h_1 as explained in Section 2.3 and choosing the bandwidth h_2 for which the bootstrap estimate of $P(\hat{\imath}_0 - i_0 = 0)$ is maximum. More precisely, choose a set of potential bandwidth values for h_2, denoted by $h_{2,j}$, $j = 0, \cdots, H$ and for each of these bandwidths we obtain the estimated probability defined in (3.3). Finally, we choose that bandwidth h_2 from the set $h_{2,j}$, $j = 0, \cdots, H$ which is associated with the largest bootstrap estimated probability. We denote this bandwidth by $\hat{h}_{2,\text{boot}}$. To stay on the safe side one should consider a large enough set of potential bandwidths. We refer to this approach as the *first bootstrap selection method*. Note that two parameters h_0 and r appearing in the first step need to be chosen. These choices however are of little importance. They represent respectively the starting value of the range of h_1 values that are considered and the multiplicative decreasing step in the sequence of h_1 values. The only precaution is to choose h_0 and r not too small. In other words, the grid of h_1 values should be fine enough and should be of a reasonable range. Safe choices are h_0 large (for example half of the length of the domain of the regression function) and r closed to 1.

It is worthwhile to consider an alternative approach for selecting the bandwidths. Note that with the above approach we in fact propose to allow for different bandwidths in the two estimation steps. What happens if we restrict ourselves to take the same bandwidth in the two estimation steps? So, instead of using the method explained in Section 2.3 for selecting h_1, we use the same potential set of smoothing parameters as in the second step. More precisely, consider a set of potential bandwidths h_j, $j = 0, \cdots, H$. For each of them, we then proceed as follows. Obtain a preliminary estimator $\tilde{x}(h_j)$, i.e. the value of x that maximizes the diagnostic function $|D(x, h_j)|$. Next construct an interval of length $2h_j$ around $\tilde{x}(h_j)$ and use least-squares fitting to obtain the final estimator corresponding to this bandwidth value h_j. Using the bootstrap algorithm described above, we then estimate the probability p_0, associated to that fixed bandwidth h_j. Finally, we select the bandwidth which is associated with the largest bootstrap estimated probability. Use this bandwidth to calculate the final estimator of the jump discontinuity. This alternative approach has a few advantages. There is no longer a need to choose the parameters h_0 and r, and this selection method requires less computing time. We refer to it as the *second bootstrap selection method*. The first bootstrap selection method is of course more flexible since it allows for two different bandwidths h_1 and h_2. In Section 4 we will evaluate via simulations the performance of both bootstrap selection methods.

Before the above bootstrap selection methods can be claimed to be fully data-driven, it remains to discuss the choice of the bandwidth parameters in the local linear fit $\hat{g}$ of g on $[0, \hat{x}_0]$ and on $[\hat{x}_0, 1]$ in the first step of the bootstrap algorithm. We propose to use cross-validation to choose the bandwidth(s) in this step. Assuming that the same bandwidth is used to estimate g to the left of $\hat{x}_0$ and to the right of $\hat{x}_0$, and denoting the local linear fit on $[0, \hat{x}_0]$ by $\hat{g}_1$ and the local linear fit on $[\hat{x}_0, 1]$ by $\hat{g}_2$ we choose h which minimizes the cross-validation quantity

$$\mathrm{CV}(h) = \sum_{i=1}^{\hat{\imath}_0} \left\{ \hat{g}_1^{-i}(X_i; h) - Y_i \right\}^2 + \sum_{i=\hat{\imath}_0+1}^{n} \left\{ \hat{g}_2^{-i}(X_i; h) - Y_i \right\}^2 , \tag{3.4}$$

where $\hat{g}_1^{-i}(\cdot)$ and $\hat{g}_2^{-i}(\cdot)$ denote the estimators $\hat{g}_1$ and $\hat{g}_2$ obtained by disgarding the i^{th} data point, on the interval $[0, \hat{x}_0]$ and $]\hat{x}_0, 1]$ respectively. The cross-validated bandwidth selector is then defined as

$$\hat{h}_{\mathrm{CV}} = \mathrm{argmin}_h \mathrm{CV}(h). \tag{3.5}$$

If needed one can built in some more flexibility by allowing two different bandwidths for estimating g to the left and to the right of $\hat{x}_0$. A cross-validation procedure for selecting two separate smoothing parameters h_L and h_R would be:

$$\begin{aligned} \hat{h}_{\mathrm{L,CV}} &= \mathrm{argmin}_{h_{\mathrm{L}}} \sum_{i=1}^{\hat{i}_0} \left\{ \hat{g}_1^{-i}(X_i; h_{\mathrm{L}}) - Y_i \right\}^2 \\ \hat{h}_{\mathrm{R,CV}} &= \mathrm{argmin}_{h_{\mathrm{R}}} \sum_{i=\hat{i}_0+1}^{n} \left\{ \hat{g}_2^{-i}(X_i; h_{\mathrm{R}}) - Y_i \right\}^2 . \end{aligned} \tag{3.6}$$

In our simulation study presented in Section 4 we allowed for two different bandwidths, applying (3.6).

3.3 Generalization for more than one discontinuity

We now explain how to generalize the data-driven estimation procedure for more than one jump discontinuity. When there are $k \geq 1$ jump points denoted by $x_1, x_2, \cdots, x_k$, with k known, we can still proceed in the same way: we identify their positions using a kernel method and improve the rate of convergence using a least-squares step. We use a four steps method for determining the location of k discontinuities. We recall the definition of $h_{1,i}$ given in Section 2.3: $h_{1,i} = h_0 r^i$ for $i = 0, 1, 2, \cdots$ with $h_0 > 0$ and $0 < r < 1$.

Step 1: Initialization.
Let M denote the number of local maxima of $|D(., h_{1,0})|$ on $]vh_{1,0}, 1-vh_{1,0}[$. Let $\{\xi_{0,j}, 1 \leq j \leq M\}$ be the set of points at which the maxima are achieved.

Step 2: Iteration.
Given a set $\{\xi_{i,j}, 1 \leq j \leq M\}$ of local maxima of $|D(., h_{1,i})|$ on $]vh_{1,0}, 1 - vh_{1,0}[$, let $\xi_{i+1,j}$ denote the local maximum of $|D(., h_{1,i+1})|$ that is nearest to $\xi_{i,j}$.

Step 3: Termination.
We stop the algorithm at iteration $i = \tilde{\imath}$, when the number of data values in some interval $]x - h_{1,i}, x + h_{1,i}[\subseteq]vh_{1,0}, 1 - vh_{1,0}[$ first falls below a predetermined value.
We take the preliminary estimates $\tilde{x}_1, \ldots, \tilde{x}_k$ of the k jump discontinuities $x_1, x_2, \cdots, x_k$ to be those values of $\xi_{\tilde{\imath},j}$ for which $|D(\xi_{\tilde{\imath},j}, h_{1,\tilde{\imath}}) - D(\xi_{0,j}, h_{1,0})|$ is one of the k largest.

Step 4: Least-squares.
Fit local least-squares within the interval $[\tilde{x}_j - h_2, \tilde{x}_j + h_2]$ for $j = 1, \ldots, k$.
Finally, we obtain the final estimators $\hat{x}_1, \ldots, \hat{x}_k$ of $x_1, \ldots, x_k$.

The three first steps of the algorithm allow to obtain preliminary estimators for $x_1, \ldots, x_k$, involving an automatic choice of the bandwidth comparable to the proposal in Section 2.3. For the least-squares fitting, we again need to select the bandwidth h_2. And again here we proceed as before.We select a (large) set of potential bandwidths $h_{2,j}$ for $j = 0, \cdots, H$ and then estimate, for each jump point x_ℓ, $\ell = 1, \cdots, k$, the discrete probability $P(\hat{\imath}_{0,\ell} - i_{0,\ell} = 0) \equiv p_{0,\ell}$, where obviously the extra subscript ℓ refers to the jump discontinuity x_ℓ. For each jump discontinuity we choose that bandwidth h_2 from the candidate set which maximizes the bootstrap estimate of the probability $p_{0,\ell}$.

An extra precaution which has to be taken here is that we need to make sure that the interval around $\tilde{x}_\ell$ in the least-squares step contains only one jump discontinuity. Hence, the set of potential bandwidths should contain no too large bandwidth values.

4 Numerical study

In this section we evaluate, via a simulation study, the data-driven estimation procedures developed in Section 3 and provide an illustration with some real data.

4.1 Simulation study

We consider the following regression functions:

$$g_1(x) = 4x^2 + I(x > 0.5);$$

$$g_2(x) = \cos\{8\pi(0.5 - x)\} - 2\cos\{(8\pi(0.5 - x)\}I(x > 0.5);$$

$$g_3(x) = \begin{cases} 4x^2 & \text{if } x \in [0, 0.2[\\ 4x^2 + 1.2 & \text{if } x \in [0.2, 0.5[\\ 4x^2 + 2 & \text{if } x \in [0.5, 1], \end{cases}$$

where $I(A)$ indicates the indicator function on a set A.

We considered 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

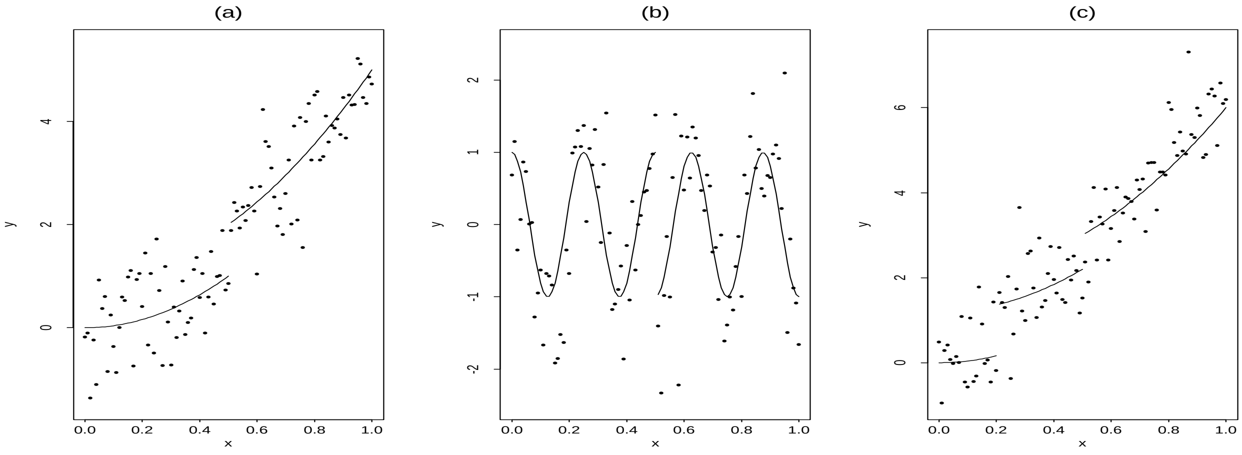

Figure 4.1: *The true regression functions (solid curves) with a typical simulated data set of size $n = 100$ and variance $\sigma^2 = 0.5$. (a). Regression function g_1; (b). Regression function g_2; (c). Regression function g_3.*

sample sizes $n = 50, 100$ or 200. Figure 4.1 presents the true regression functions g_1, g_2 and g_3 with typical simulated data sets for sample size $n = 100$ and $\sigma^2 = 0.5$. Note that both functions g_1 and g_2 have a single jump discontinuity at $x_0 = 0.5$ of size 1 and 2 respectively, whereas the function g_3 has two jump discontinuities of size 1.2 and 0.8 occuring at x_0 =0.2 and x_0 =0.5 respectively.

For the diagnostic function in (2.2), we use a standard Gaussian kernel. In all studies we considered 1000 simulations and the number of bootstrap replicates was $B = 2000$. For each example we use the cross-validation bandwidths defined in (3.6) to choose the smoothing parameters in the first step of the bootstrap algorithm described in Section 3.1.

We investigate the performance of the *first bootstrap selection method* and the *second bootstrap selection method* by simulating from model (2.1) with unknown regression function g_1. For determining $\tilde{h}_1$, defined in Section 2.3, we search over the set of bandwidths $h_{1,i} = h_0 r^i$, for $i = 0, 1, 2, \cdots$, where $h_0 = 0.2$ and $r = 0.9$. This means that the biggest bandwidth that we consider is $h_0 = 0.2$, which is one quarter of the domain of the regression function. 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 $j = 0, \cdots, 18$. Figure 4.2 (a) depicts a simulated data set for $n = 50$ and $\sigma^2 = 0.1$ together with the true function

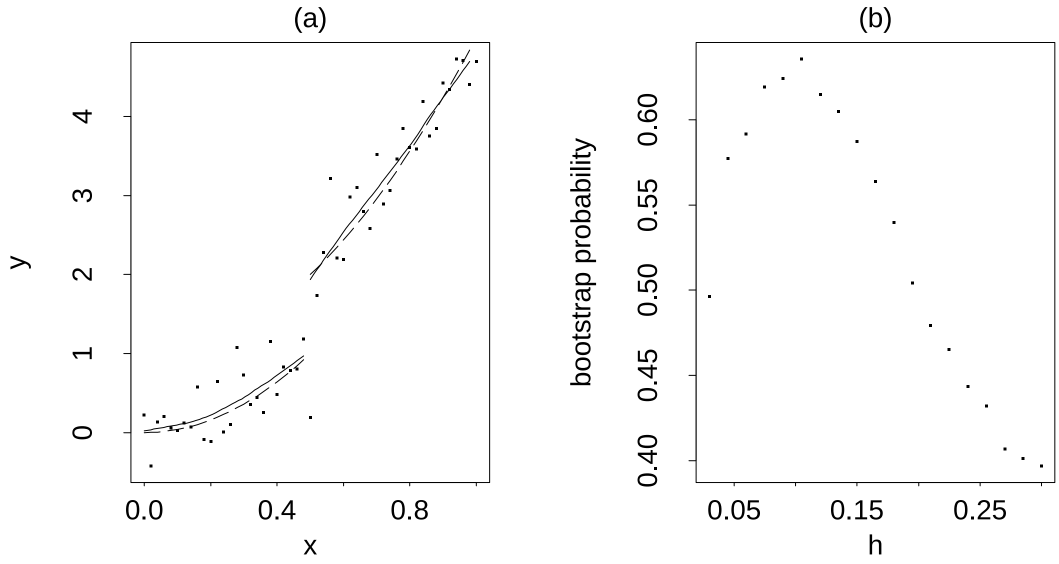

Figure 4.2: *(a). A simulated data set with $n = 50$ and $\sigma^2 = 0.1$, together with the true regression function $g_1(x)$(solid curve) and a local linear estimator, adapted to the estimated change-point (dashed curve) ; (b). The bootstrap estimate of the probability p_0 for a range of values of the bandwidth h_2.*

$g_1(\cdot)$. Presented is also a local linear estimator of the regression function, obtained by applying local linear fitting to the left and right of the estimated jump point. Figure 4.2 (b) shows the bootstrap estimates of the probability p_0, associated with the potential bandwidths $h_{2,j}$ for $j = 0, \cdots, 18$. For this particular simulation the bandwidth selected by the bootstrap algorithm is $\hat{h}_{2,\mathrm{boot}} = 0.105$, which is the bandwidth with the largest bootstrap estimated probability. With this bandwidth the final estimator of the discontinuity point is $\hat{x}_0$=0.51. To give an idea about the variability of the bootstrap algorithm for selecting h_2 we present in Figure 4.3 a kernel density estimate of the 1000 bandwidths $\hat{h}_{2,\mathrm{boot}}$ selected via the bootstrap algorithm for the 1000 simulations. This graph shows that for these 1000 simulations, the bootstrap selected bandwidth $\hat{h}_{2,\mathrm{boot}}$ has a density function concentrated around the value 0.0511.

In Table 4.1 we summarize the simulation results for the function g_1. Presented are 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

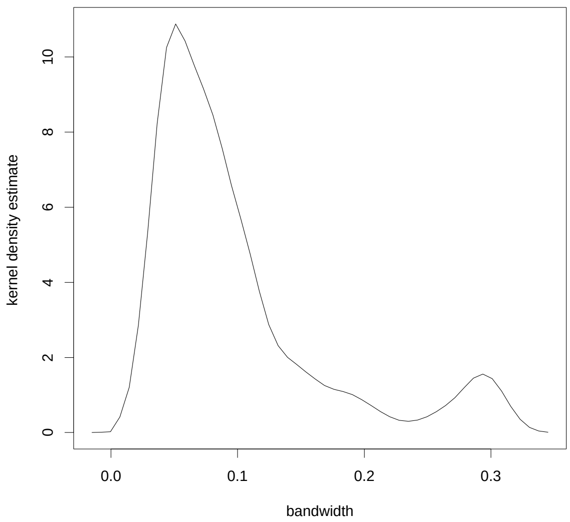

Figure 4.3: *A kernel density estimate of the bandwidth* $\hat{h}_{2,\mathrm{boot}}$ *selected by the bootstrap procedure, based on 1000 simulations from regression function* g_1.

simulations. Results are given for the two bootstrap selection methods described in Section 3.2. Recall that the *first bootstrap selection method* is the more general one allowing for two different bandwidths in the two steps of the estimation method, whereas the *second bootstrap selection method* only uses one bandwidth. Note that for the second selection method the percentages of values of $\hat{x}_0$ in the interval $[0.45, 0.55[$ are slightly smaller (see columns one versus columns two) showing a slightly better performance of the first method. The means and standard deviations for the two methods are quite close. Overall, the two methods perform quite comparably, with a slightly better performance for the more flexible first method. Note that the data-driven methods perform very well even for these rather small sample sizes.

The function g_2 presents some specific difficulties because it shows many fluctuations: regions with steep increase followed by regions with a steep decrease. As a consequence, the diagnostic function $|D(x,h)|$ typically will have a lot of local maxima. Since the function $g_2(\cdot)$ has steep gradients at several locations, the largest maxima in the diagnostic function may as well be due to a steep gradient rather than to a jump point. To overcome

Table 4.1: *Simulation results for the function g_1, for the first selection method (two h's) and the second method (only one h).*

	$\sigma^2 = 0.1$				$\sigma^2 = 0.5$	
	$n = 50$		$n = 100$		$n = 100$	
	one h	two h's	one h	two h's	one h	two h's
% in $[0.0, 0.15[$	0	0	0	0	0.2	0
% in $[0.15, 0.25[$	0	0	0	0	0.4	0
% in $[0.25, 0.35[$	0	0	0	0	0.7	0.6
% in $[0.35, 0.45[$	0.1	0.1	0.1	0	2.2	2.1
% in $[0.45, 0.55[$	93.9	95.3	99.5	99.8	78.4	80.1
% in $[0.55, 0.65[$	4.9	3	0.4	0.2	9.8	6.3
% in $[0.65, 0.75[$	0.1	0.2	0	0	3	3.1
% in $[0.75, 0.85[$	0.6	0.6	0	0	3.2	4.5
% in $[0.85, 1.0[$	0.4	0.8	0	0	2.1	3.4
mean of $\hat{x}_0$	0.517640	0.518100	0.50685	0.506610	0.533130	0.540420
sd of $\hat{x}_0$	0.035288	0.044805	0.007330	0.006675	0.091399	0.100969

Table 4.2: *Simulation results for the function g_2.*

	$\sigma^2 = 0.1$		$\sigma^2 = 0.5$
	$n = 100$	$n = 200$	$n = 200$
% in $[0.0, 0.2[$	1.9	1	8.9
% in $[0.2, 0.4[$	5	2.3	13.4
% in $[0.4, 0.6[$	85.7	93.8	57.1
% in $[0.6, 0.8[$	4.1	1.7	11.6
% in $[0.8, 1.0[$	3.3	1.2	9
mean of $\hat{x}_0$	0.497210	0.484165	0.49356
sd of $\hat{x}_0$	0.097851	0.069340	0.174536

this problem, we suggest to use the algorithm developed in Section 3.3 for the case of multiple jump points. We simulated 1000 samples of size $n = 100$ or $n = 200$ from model

Table 4.3: *Simulation results for the function g_3.*

	$\sigma^2 = 0.1$				$\sigma^2 = 0.5$	
	$n = 50$		$n = 100$		$n = 100$	
	0.2	0.5	0.2	0.5	0.2	0.5
% in $[0.0, 0.15[$	0.1	0	0	0	1.2	0
% in $[0.15, 0.25[$	95.5	0	99.8	0	83	0.4
% in $[0.25, 0.35[$	1	0	0	0	4.6	0.2
% in $[0.35, 0.45[$	0	0.7	0	0.2	0.4	3
% in $[0.45, 0.55[$	0.5	84	0	96.2	1.6	65.4
% in $[0.55, 0.65[$	0	6.6	0	0.6	1	11.8
% in $[0.65, 0.75[$	0.4	2.8	0.2	1.6	2.6	9.2
% in $[0.75, 0.85[$	1.7	4	0	0.8	3.8	7.6
% in $[0.85, 1.0[$	0.8	1.9	0	0.6	1.8	2.4
mean of $\hat{x}_0$	0.228880	0.536200	0.206200	0.514000	0.266820	0.560800
sd of $\hat{x}_0$	0.102347	0.080836	0.023935	0.048503	0.168740	0.111661

(2.1) with $g = g_2$. As before we took $h_{1,i} = 0.2\ 0.9^i$ for $i = 0, 1, \cdots$, and the set of potential bandwidths h_2 for the least-squares step was taken to be $h_{2,j} = 0.03 + 0.015j$ for $j = 0, 1, \cdots, 5$. Table 4.2 shows the percentages that the estimated values $\hat{x}_0$ fall in the specified intervals. Gijbels, Hall and Kneip (1999) also provide simulation results for the function g_2 (for different sample sizes and values of σ) without data-driven choices of the parameters involved. Looking at the results presented there we can conclude that the results in Table 4.2 are far better, confirming again that the bandwidth parameter choices are crucial.

Function g_3 shows two jump discontinuities of different sizes. Here we searched over the discrete set of bandwidths $h_{1,i} = h_0 r^i$, for $i = 0, 1, 2, \cdots$, where $h_0 = 0.1$ and $r = 0.9$. The set of potential bandwidths for the least-squares step was taken to be $h_{2,j} = 0.03 + 0.015j$, for $j = 0, \cdots, 5$. The results are summarized in Table 4.3. Note that the discontinuity point 0.2 is better detected than the second jump point 0.5. This can be explained by the fact that the jump-size at 0.2 is bigger than the jump-size at 0.5. Also here the results are quite good for such small sample sizes.

4.2 Applications

As an illustration we now apply the data-driven estimation method to some real data sets. The first data set consists of measurements in mils of the thickness of 90 US Lincoln pennies, two each year from 1945 through 1989. Penny thickness was reduced in World War II and restored to its original thickness sometime around 1960 and reduced again in the 70's. To analyse the data one sometimes replaces the original data by the mean of each pair per year. These data are given in Scott (1992) and are displayed in Figure 4.4. A semiparametric change-point type analysis of the data was done by Speckman (1994) who found that there were changes in thickness around the years 1958 and 1974.

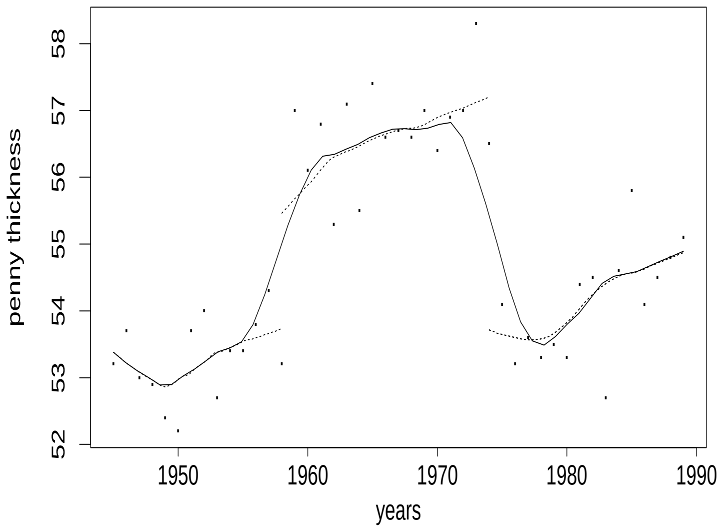

Figure 4.4: *Thickness of pennies from 1945 to 1989: the data, a nonparametric estimate adapted to the estimated change point of the regression function (solid curve) together with a non-adapted estimate of the regression function (dotted curve.)*

The bootstrap procedure tells us that the two abrupt changes occur in 1958.5 and 1974.5 with a value of the bootstrap selected bandwidth equal to 5 for the estimation of the two jump points. We have noted that in this application the choice of the bandwidth seems to be less crucial: a number of alternative bandwidths lead to the same estimated locations of the jump points. A reason for this is probably that this is a rather 'clear-cut' case: even by simply looking at the data one could 'guess' the locations.

A second application concerns 215 average annual temperatures measured in Prague from 1775 to 1989, discussed in Horváth and Kokoszka (1997). This data set was analyzed by Horváth and Kokoszka (1997) in order to detect climatic changes that took place over a span of several years or a decade. Antoniadis and Gijbels (2000) also studied these data from 1775 to 1902 and found, using a wavelet method, two jump points occuring at 1787 and 1837.

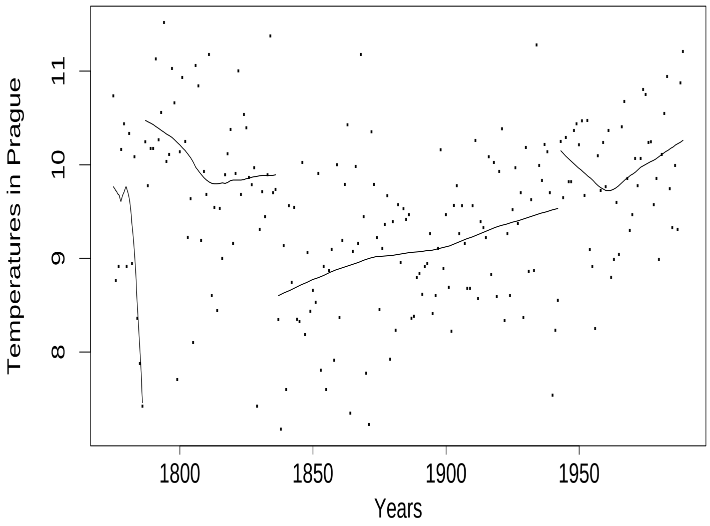

Figure 4.5: *The average annual temperatures in Prague (points) displayed with a local linear estimator adapted to three estimated change-points.*

We apply our data-driven method and search for two change-points in the data between 1775 and 1902. The obtained estimators for the locations of the jump points are 1786.5 and 1836.5, which agrees with what was found by Antoniadis and Gijbels (2000). If we fix the number of jump points to be three (four) between 1775 and 1989, we obtained the estimated locations (in years) 1786.5, 1836.5 and 1942.5 (or 1786.5, 1836.5, 1858.5 and 1942.5 in case of four change-points). These results agree with those obtained by Horváth and Kokoszka (1997). Finally, the entire function was fitted, using a local linear estimator (to the left, centre and right of the respective estimated points 1787.5, 1836.5 and 1942.5). The resulting estimate is displayed in Figure 4.5.

5 Concluding remarks and discussion

In this paper we addressed the important issue of selecting smoothing parameters when estimating jump discontinuities in an unknown regression function. The proposed data-driven estimation procedure seems to perform very well, even for the more difficult cases. The selection method is based on a bootstrap procedure and therefore performs also well for small sample sizes.

Jump discontinuities represent only one type of irregularities that might occur in an otherwise smooth regression function. Other types of irregularities include changes in the derivative functions. The data-driven estimation procedure developed in this paper can be adapted for detecting jumps in derivative functions. This essentially requires the use of an appropriate diagnostic function and an appropriate family of parametric functions in the least-squares step. Results on this subject will be reported elsewhere.

One also might be interested in testing whether an unknown regression function has jump discontinuities or not. Such testing problems are rather difficult and mostly are dealt with via asymptotic theory. Bootstrap testing procedures based on the two-steps estimation method have been proposed by the authors, and seem to perform very well, in comparison with other testing procedures available in the literature. They do not rely on asymptotics and are data-driven, so that the user is not left with a difficult and crucial choice of some smoothing parameter.

Acknowledgements

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

References

Antoniadis, A. and Gijbels, I. (2000). Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics*, to appear.

Beunen, J.A. (1998). Implementation of a method for detection and location of discontinuities. *Thesis* for Master of Scientific Studies (Statistics). University of Newcastle, Australia.

Bunt, M., Koch, I. and Pope, A. (1995). Kernel-based nonparametric regression for discontinuous functions. *Statistics Research Report SRR 012-95*, The Australian National University.

Bunt, M., Koch, I. and Pope, A. (1998). Counting discontinuities in nonparametric regression. *Research Report 98-2*, University of Newcastle, Ausytralia.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **8**, 679–698.

Chu, C.K. (1994). Estimation of change-points in a nonparametric regression function through kernel density estimation. *Communications in Statistics, Part A–Theory and Methods*, **23**, 3037–3062.

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., 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.

Girard, D. (1990). Détection de discontinuités dans un signal par inf-convolution spline et validation croisée. *Technical Report RR 702-I-M*, University of Grenoble, France.

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

Hamrouni, Z. (1999). Inférence statistique par lissage linéaire local pour une fonction de régression présentant des discontinuités. *Doctoral Thesis*, Université de Joseph Fourier, Grenoble, France.

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

Kang, K.-H., Koo, J.-Y. and Park, C.-W. (2000). Kernel estimation of discontinuous regression functions. *Statistics & Probability Letters*, **47**, 277–285.

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

Korostelev, A.P. (1987). On minimax estimation of a discontinuous signal. *Theory of Probability and its Applications*, **32**, 727–730.

Laurent, P.J. and Utreras, F. (1986). Optimal smoothing of noisy broken data with spline functions. *Journal of Approximation Theory and its Applications*, **2**, 71–94.

Leclerc, Y.C. and Zucker, S.W. (1987). The local structure of image discontinuities in one dimension. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **9**, 341–355.

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

Mallat, S. and Hwang, W.L. (1992). Singularity detection and processing with wavelets. *IEEE Trans. Information Theory*, **2**, 617–643.

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 Song, K.-S. (1997). Two-stage change-point estimators in smooth regression models. *Statistics & Probability Letters*, **34**, 323–335.

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

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

Oudshoorn, C.G.M. (1998). Asymptotically minimax estimation of a function with jumps. *Bernoulli*, **4**, 15–33.

Potier, C. and Vercken, C.(1994). Spline fitting numerous noisy data with discontinuities. In *Curves and Surfaces*, Eds Laurent, P.J. et al., 477–480, Academic Press.

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.

Scott, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* Wiley, New York.

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

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. (1993a). Kernel type estimation of jump points and values of regression function. *The Annals of Statistics*, **21**, 1545–1566.

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

Wu, J.S. and Chu, C.K. (1993c). Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statistica Sinica*, **3**, 557 – 576.

Yin, Y.Q. (1988). Detection of the number, locations and magnitudes of jumps. *Communications in Statistics, Stochastic Models*, **4**, 445–455.