Tilburg University

The half-half plot

Einmahl, J.H.J.; Gantner, M.

Published in:
Technometrics

DOI:
[10.1080/07350015.2012.657102](https://doi.org/10.1080/07350015.2012.657102)

Publication date:
2012

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Einmahl, J. H. J., & Gantner, M. (2012). The half-half plot. *Technometrics*, 54(2), 138-146.
<https://doi.org/10.1080/07350015.2012.657102>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This article was downloaded by: [Tilburg University]

On: 26 May 2012, At: 14:10

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Technometrics

Publication details, including instructions for authors and subscription information:

<http://amstat.tandfonline.com/loi/utch20>

The Half-Half Plot

John H.J. Einmahl^a & Maria Gantner^a

^a Department of Econometrics & OR and CentEr, Tilburg University, Tilburg, LE, 5000, The Netherlands

Available online: 17 Jan 2012

To cite this article: John H.J. Einmahl & Maria Gantner (2012): The Half-Half Plot, Technometrics, 54:2, 138-146

To link to this article: <http://dx.doi.org/10.1080/07350015.2012.657102>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://amstat.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

The Half-Half Plot

John H.J. EINMAHL

Department of Econometrics & OR and
CentEr, Tilburg University
5000 LE Tilburg, The Netherlands
(j.h.j.einmahl@uvt.nl)

Maria GANTNER

Department of Econometrics & OR and
CentEr, Tilburg University
5000 LE Tilburg, The Netherlands
(mariagantner@gmail.com)

The Half-Half (HH) plot is a new graphical method to investigate qualitatively the shape of a regression curve. The empirical HH-plot counts observations in the lower and upper quarter of a strip that moves horizontally over the scatterplot. The plot displays jumps clearly and reveals further features of the regression curve. We prove a functional central limit theorem for the empirical HH-plot, with rate of convergence $1/\sqrt{n}$. In a simulation study, the good performance of the plot is demonstrated. The method is also applied to two case studies. The proofs and one more case study are deferred to a supplement, which is available online.

KEY WORDS: Data analysis; Functional central limit theorem; Graphical methods; Jump detection; Nonparametric regression.

1. INTRODUCTION

Assume the pairs $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$, $n \in \mathbb{N}$, are independent and identically distributed (iid) with bivariate distribution function (df) F . From the regression perspective, we can define $\varepsilon = Y - m(X)$ where m , the nonparametric regression function, is a location functional (like the median, mean, or mode) applied to the conditional distribution of $Y|X = x$. Equivalently, this can be written in the “standard form”:

$$Y = m(X) + \varepsilon.$$

A first step in data analysis is exploratory diagnostics. Using a good graphical representation of the data sheds light on their main features. The standard procedure for depicting m is kernel estimation, which produces a smooth regression function. When there are sudden changes in m , such as jumps, those are forced into the smoothed picture and therefore ignored. More refined procedures that allow for and detect jumps of m have been introduced and will be discussed later.

In this article, we present a novel, nonparametric, computationally fast method for exploring regression curves. It displays important features of a regression curve, such as jumps and increases or decreases. If it is a goal to search for jumps, then an ad hoc estimator (introduced in Section 4.2) can be used to find their locations. These procedures impose no particular model on the data; also the regression function m need not be estimated.

Let F be absolutely continuous with Lebesgue density f . Denote the corresponding probability measure with P . Write F_1, F_2 for the marginals of F and Q_1, Q_2 for their (left-continuous) inverse or quantile functions. The empirical counterparts of these functions are denoted with a subscript n , in particular, F_n denotes the empirical df of the (X_i, Y_i) , $i = 1, \dots, n$, that is,

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x] \times (-\infty, y]}(X_i, Y_i), \quad -\infty < x, y \leq \infty.$$

Fix $\alpha \in (0, \frac{1}{2})$. For $x \in (Q_1(\alpha), Q_1(1 - \alpha))$, define the vertical α -strip centered at x by

$$S_\alpha(x) := \{(u, v) \in \mathbb{R}^2 : Q_1(F_1(x) - \alpha) \leq u \leq Q_1(F_1(x) + \alpha)\}.$$

Consider the univariate df on this strip

$$G_{x,\alpha}(y) := \frac{1}{2\alpha} [F(Q_1(F_1(x) + \alpha), y) - F(Q_1(F_1(x) - \alpha), y)], \quad y \in \mathbb{R}.$$

The corresponding quantile function is denoted by $Q_{x,\alpha}$. For convenience, we write in the sequel $x^- = Q_1(F_1(x) - \alpha)$ and $x^+ = Q_1(F_1(x) + \alpha)$.

Definition 1. Let the coverage $\alpha \in (0, \frac{1}{2})$ be fixed. The HH-value for $x \in (Q_1(\alpha), Q_1(1 - \alpha))$ is defined by

$$H_\alpha(x) = F\left(x, Q_{x,\alpha}\left(\frac{1}{4}\right)\right) - F\left(x^-, Q_{x,\alpha}\left(\frac{1}{4}\right)\right) - \left(F\left(x^+, Q_{x,\alpha}\left(\frac{3}{4}\right)\right) - F\left(x, Q_{x,\alpha}\left(\frac{3}{4}\right)\right)\right) + \frac{1}{2}\alpha.$$

The HH-value is obtained by first vertically dividing the strip $S_\alpha(x)$, as depicted in Figure 1, into two halves of equal mass: the middle half (C) and the outer half [lower ($D \cup E$) and upper ($A \cup B$) quarter]. Then the mass in B is added to the mass in D . To standardize the statistic, $\frac{1}{2}\alpha$ is subtracted; this is the mass in $B \cup D$ in case X and Y are independent. Hence, in that case, $H_\alpha \equiv 0$. In general, $H_\alpha \in [-\frac{1}{2}\alpha, \frac{1}{2}\alpha]$.

Note that

$$H_\alpha(x) = F\left(x, Q_{x,\alpha}\left(\frac{1}{4}\right)\right) - F\left(x^-, Q_{x,\alpha}\left(\frac{1}{4}\right)\right) + F\left(x, Q_{x,\alpha}\left(\frac{3}{4}\right)\right) - F\left(x^-, Q_{x,\alpha}\left(\frac{3}{4}\right)\right) - \alpha.$$

Figure 1. Subsets of $S_\alpha(x)$, needed for calculation of $H_\alpha(x)$.

Definition 2. Let $\alpha \in (0, \frac{1}{2})$. The HH-plot is the graph of the function

$$x \mapsto H_\alpha(x), \quad x \in (Q_1(\alpha), Q_1(1 - \alpha)).$$

Define for $x \in (Q_{n,1}(\alpha), Q_{n,1}(1 - \alpha))$ the empirical counterpart of $G_{x,\alpha}$ by

$$G_{n,x,\alpha}(y) = \frac{1}{2\alpha} \left[F_n(Q_{n,1}(F_{n,1}(x) + \alpha), y) - F_n\left(Q_{n,1}\left(F_{n,1}(x) - \frac{[n\alpha]}{n}\right), y\right) \right], \quad y \in \mathbb{R},$$

where $[\cdot]$ denotes the ceiling function. Write $X_{[1]} \leq X_{[2]} \leq \dots \leq X_{[n]}$ for the order statistics of the $X_i, i = 1, \dots, n$. It is easily seen that

$$G_{n,x,\alpha}(y) = \frac{1}{2\alpha} [F_n(X_{[i+[n\alpha]]}, y) - F_n(X_{[i-[n\alpha]]}, y)],$$

where i is such that $X_{[i]} \leq x < X_{[i+1]}$. Denote the quantile function of $G_{n,x,\alpha}$ with $Q_{n,x,\alpha}$. The HH-statistic is for $x \in (Q_{n,1}(\alpha), Q_{n,1}(1 - \alpha))$ defined to be the empirical version of the HH-value:

$$H_{n,\alpha}(x) = F_n\left(x, Q_{n,x,\alpha}\left(\frac{1}{4}\right)\right) - F_n\left(Q_{n,1}\left(F_{n,1}(x) - \frac{[n\alpha]}{n}\right), Q_{n,x,\alpha}\left(\frac{1}{4}\right)\right)$$

$$\begin{aligned} &+ F_n\left(x, Q_{n,x,\alpha}\left(\frac{3}{4}\right)\right) \\ &- F_n\left(Q_{n,1}\left(F_{n,1}(x) - \frac{[n\alpha]}{n}\right), Q_{n,x,\alpha}\left(\frac{3}{4}\right)\right) - \alpha \\ &= F_n\left(X_{[i]}, Q_{n,x,\alpha}\left(\frac{1}{4}\right)\right) - F_n\left(X_{[i-[n\alpha]]}, Q_{n,x,\alpha}\left(\frac{1}{4}\right)\right) \\ &+ F_n\left(X_{[i]}, Q_{n,x,\alpha}\left(\frac{3}{4}\right)\right) \\ &- F_n\left(X_{[i-[n\alpha]]}, Q_{n,x,\alpha}\left(\frac{3}{4}\right)\right) - \alpha. \end{aligned}$$

The empirical HH-plot is defined to be the graph of the function

$$x \mapsto H_{n,\alpha}(x), \quad x \in (Q_{n,1}(\alpha), Q_{n,1}(1 - \alpha)).$$

The article is organized as follows. Features of the HH-plot are discussed in Section 2. The asymptotic behavior of the HH-statistic is expounded in Section 3, where it is shown that the rate of convergence of the empirical HH-statistic to the theoretical one is $1/\sqrt{n}$, uniformly in x . The limit theorem is used to formalize the detection of jumps. In Section 4, three simulation studies are presented and an ad hoc estimator for a jump location is introduced in Section 4.2. Finally, two real data applications can be found in Section 5.

2. FEATURES OF THE HH-PLOT

Jumps of m are depicted in the HH-plot by a high positive (jump up) or negative (jump down) value. In the first example, we consider $m_1(x) = \mathbb{1}_{[0.5, \infty)}(x)$, with X uniformly distributed on $(0, 1)$ and ε standard normal and independent of X , see Figure 2. The jump is clearly indicated by the large HH-value and HH-statistic at $x = 0.5$. Additionally, the HH-plot reveals that the regression curve is constant before and after the jump, which is indicated by values of $H_\alpha(x)$ and $H_{n,\alpha}(x)$ close to 0.

Another feature of the HH-plot can be observed in Figure 3. Let X and ε be as above, but $m_2(x) = x - \mathbb{1}_{[0.5, \infty)}(x)$. The jump down is indicated by large negative values of $H_\alpha(0.5)$ and $H_{n,\alpha}(0.5)$, whereas the positive values suggest that m_2 is increasing; for the empirical HH-plot; however, small positive (or

Figure 2. (a) Theoretical and (c) empirical HH-plot with $\alpha = 0.1$ (gray line) and $\alpha = 0.2$ (black line) of (b) a sample of size $n = 500$ of m_1 with ε standard normal and $X \sim UN(0, 1)$.

Figure 3. (a) Theoretical and (c) empirical HH-plot with $\alpha = 0.1$ (gray line) and $\alpha = 0.2$ (black line) of (b) a sample of size $n = 500$ of m_2 with ε standard normal and $X \sim UN(0, 1)$.

negative) values of $H_{n,\alpha}(x)$ can also be due to random fluctuations.

The HH-plot is an appropriate tool when there is a single dominant predictor X ; in case of more than one relevant predictor, the present version of the plot does not apply. Due to the discrete character of $H_{n,\alpha}$, the use of the HH-plot is limited to sample sizes of at least 50, but $n \geq 100$ is recommended. Another feature of $H_{n,\alpha}$ is that it is only defined for $x \in (Q_{n,1}(\alpha), Q_{n,1}(1 - \alpha))$. Therefore, a part of the x -values is not covered by the HH-plot; this part “decreases” with α . A jump location that is not covered by the plot cannot be detected by it. In Section 4.4, we delineate some other limitations of the HH-plot.

Clearly, the HH-plot and its performance depend on α . A good value of α is one where the empirical HH-plot provides insight into features of the regression curve m . In case there is a jump, we want the probability of detecting it to be as large as possible (see Section 3.2 for the “critical values” used for this detection). Such an optimal α depends very much on the true, unknown m and not so much on the sample size n . Typically a relatively large α works well for jump detection (e.g., for m_1 above, the probability of detection increases with α irrespective of n), but recall that a smaller α has the advantage that the HH-plot covers a larger range of x -values. In general, $\alpha = 0.25$ is a natural maximum for α , otherwise the range of x -values is less than “half.” Due to the aforementioned discrete character of the plot, about $20/n$ is a lower bound for α . Let us consider sample sizes ranging from 100 (see above) to 1000. If $100 \leq n < 200$, there is not much flexibility and we recommend to take $\alpha = 0.20$. For $200 \leq n \leq 1000$, it is advised to combine the information provided by the HH-plots for two α -values. We recommend to take the larger $\alpha = 0.2$ again, and to take the smaller $\alpha = 0.1$. The latter α makes “half” of the x -values that we do not see for $\alpha = 0.2$ visible.

There exists a large body of literature on the specific topic of estimating jump points in nonparametric regression. A good overview can, for example, be found in Gijbels, Lambert, and Qiu (2007). The most common approach is to compare left- and right-sided estimators of the regression function at point x . These estimators are attained by kernel methods [among others, Hall and Titterton (1992), Müller (1992), Wu and Chu (1993), and Joo and Qiu (2009)] or local polynomial

regression [among others, McDonald and Owen (1986), Loader (1996), Grégoire and Hamrouni (2002), Horváth and Kokoszka (2002), and Gijbels et al. (2007)]. Some articles propose a two-step procedure, where in a first step, the jump location is estimated and in a second step the curve is fitted [cf. Gijbels, Hall, and Kneip (1999), Sánchez-Borrego, Martínez-Miranda, and González-Carmona (2006)]. Wavelet methods are, among others, employed in Wang (1995), Antoniadis and Gijbels (2002), and Park and Kim (2006). In Kim and Marron (2006), another graphical method, the SiZer for jump detection, is introduced. The SiZer is a kernel-based approach that combines multiple bandwidths in one plot. Its main goal is to get a first, general overview of the regression curve.

In the articles mentioned above, the specific assumptions on the regression curve and the errors play an important role. In contrast, we have no direct assumptions on m or ε . In Dempfle and Stute (2002), an approach based on U-statistics is used. As in the present article, this approach requires minimal assumptions on the model and avoids smoothing for estimating the location of the jump.

3. ASYMPTOTIC RESULTS AND INFERENCE BAND

3.1 Functional Central Limit Theorem

Define for $n \in \mathbb{N}$ the empirical process indexed by points as

$$U_n(x, y) := n^{\frac{1}{2}} \{F_n(x, y) - F(x, y)\}, \quad -\infty < x, y \leq \infty.$$

Furthermore, let B_F be a bounded, mean zero Gaussian process on $(-\infty, \infty]^2$ that is uniformly continuous and has covariance function $F(x_1 \wedge x_2, y_1 \wedge y_2) - F(x_1, y_1)F(x_2, y_2)$, $(x_1, y_1), (x_2, y_2) \in (-\infty, \infty]^2$. Then, by the functional central limit theorem for U_n and the Skorohod representation theorem, there exist $\tilde{B}_F \stackrel{d}{=} B_F$ and a sequence $\tilde{U}_n \stackrel{d}{=} U_n$, $n \in \mathbb{N}$, such that

$$\sup_{-\infty < x, y \leq \infty} |\tilde{U}_n(x, y) - \tilde{B}_F(x, y)| \rightarrow 0 \quad \text{a.s. as } n \rightarrow \infty. \quad (1)$$

Henceforth we will use (1), but drop the tildes from the notation.

We will use the following assumptions:

(A) Let $\mathcal{S} := \{(x, y) \in \mathbb{R}^2 : f(x, y) > 0\} = (x_*, x^*) \times \mathbb{R}$ for some $-\infty \leq x_* < x^* \leq \infty$ and F be of class C^2 on $\mathcal{S}$ and f_2 be bounded.

Hence, the function $G_{x,\alpha}(\cdot)$ is increasing and continuous on $(Q_1(\alpha), Q_1(1-\alpha))$; also we can write $F'_x(x, y) := \frac{\partial}{\partial x} F(x, y) = \int_{-\infty}^y f(x, v) dv$ and $F'_y(x, y) := \frac{\partial}{\partial y} F(x, y) = \int_{-\infty}^x f(u, y) du$.

For the second assumption, we first introduce some more notation: let $\tilde{F}(u, y) := F(Q_1(u), y)$, $(u, y) \in (0, 1) \times \mathbb{R}$, denote the df after using the quantile transformation to achieve a uniform marginal for the first coordinate, and write $\tilde{F}'_x(u, y) := \frac{\partial}{\partial u} \tilde{F}(u, y) = \frac{F'_x(Q_1(u), y)}{f_1(Q_1(u))}$ and $\tilde{F}'_y(u, y) := \frac{\partial}{\partial y} \tilde{F}(u, y) = F'_y(Q_1(u), y)$.

(B) $\tilde{F}'_x$ and $\tilde{F}'_y$ are uniformly continuous on $(0, 1) \times \mathbb{R}$.

Let $\alpha \in (0, \frac{1}{2})$ and define $I_n := (Q_1(\alpha), Q_1(1-\alpha)) \cap (Q_{n,1}(\alpha), Q_{n,1}(1-\alpha))$ and $I_0 := (Q_1(\alpha), Q_1(1-\alpha))$. Denote the HH-process with

$$V_{n,\alpha}(x) := n^{\frac{1}{2}}(H_{n,\alpha}(x) - H_\alpha(x)), \quad x \in I_n,$$

and for $s \in \{\frac{1}{4}, \frac{3}{4}\}$ define

$$\begin{aligned} B_{\alpha,s}(x) = & B_F(x, Q_{x,\alpha}(s)) - B_F(x^-, Q_{x,\alpha}(s)) \\ & - \frac{F'_x(x^-, Q_{x,\alpha}(s))}{f_1(x^-)} [B_F(x, \infty) - B_F(x^-, \infty)] \\ & + \frac{F'_y(x^-, Q_{x,\alpha}(s)) - F'_y(x, Q_{x,\alpha}(s))}{F'_y(x^+, Q_{x,\alpha}(s)) - F'_y(x^-, Q_{x,\alpha}(s))} \\ & \times \left\{ B_F(x^+, Q_{x,\alpha}(s)) - B_F(x^-, Q_{x,\alpha}(s)) \right. \\ & + \frac{F'_x(x^+, Q_{x,\alpha}(s))}{f_1(x^+)} [B_F(x, \infty) - B_F(x^+, \infty)] \\ & \left. - \frac{F'_x(x^-, Q_{x,\alpha}(s))}{f_1(x^-)} [B_F(x, \infty) - B_F(x^-, \infty)] \right\}, \\ & x \in I_0. \end{aligned} \quad (2)$$

Theorem 1. Under the assumptions (A) and (B), we have for all $\alpha \in (0, \frac{1}{2})$, on the probability space of (1),

$$\sup_{x \in I_n} |V_{n,\alpha}(x) - (B_{\alpha,1/4}(x) + B_{\alpha,3/4}(x))| \rightarrow 0 \quad \text{a.s. as } n \rightarrow \infty.$$

Remark 1. For X and Y independent (m is constant), the distribution of the limiting process from Theorem 1 for fixed $x \in I_0$ simplifies to a $N(0, \frac{1}{4}\alpha)$ -distribution.

The proofs of Theorem 1 and Remark 1 are available online in the supplementary files for this article.

3.2 Inference Band

For an easier interpretation of the HH-plot, we add a horizontal band to the picture. If the empirical HH-plot escapes the band, this indicates jumps or steep increases or decreases of the regression curve. In addition, this band gives a standard to assess the relative magnitude of the HH-statistic. The band is obtained by calculating high quantiles of

$$\sup_{x \in I_0} |B_{\alpha,1/4}(x) + B_{\alpha,3/4}(x)| \quad (3)$$

Table 1. The 0.90, 0.95, and 0.99 quantiles of the random variable in (3), for $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$

α	0.05	0.1	0.15	0.2	0.25
$q_\alpha(0.9)$	0.40	0.53	0.61	0.67	0.72
$q_\alpha(0.95)$	0.42	0.56	0.66	0.73	0.79
$q_\alpha(0.99)$	0.47	0.64	0.75	0.84	0.91

in case X and Y are independent. These quantiles are denoted as q_α (see Table 1) and have to be divided by $\sqrt{n}$ and $-\sqrt{n}$, respectively, in order to obtain the upper and lower boundary of the band.

4. SIMULATION STUDY

4.1 X and Y Independent

First, we consider the case of X, Y being independent, with sample sizes of $n = 250$ and 500 for coverage levels $\alpha = 0.1$ and 0.2 , respectively. For each n , 10,000 samples are taken. Table 2 provides the fraction of exceedances of $q_\alpha(0.95)/\sqrt{n}$ by $\max\{|H_{n,\alpha}(x)| : Q_{n,1}(\alpha) < x < Q_{n,1}(1-\alpha)\}$. We see that these numbers are close to, but somewhat lower than, $1 - 0.95 = 0.05$. This might be due to the fact that for fixed x , the effective sample size (i.e., the number of observations in the lower and upper quarter of the strip) is $n\alpha$, which can be as small as 25. Indeed, for $n = 500$ and $\alpha = 0.2$, we see that the asymptotic quantiles are quite accurate.

4.2 Regression Functions m_1 and m_2 of Section 2

Next, we consider m_1 and m_2 as in Section 2. We simulate from the models presented there, that is, $X \sim UN(0, 1)$ and $\varepsilon \sim N(0, 1)$, and in addition take centered normal errors with $\sigma = 0.1$ and $\sigma = 0.5$. Furthermore, we consider $\varepsilon = \varepsilon_1/\sqrt{12}$, where ε_1 is t_3 -distributed, and $\varepsilon = 3\varepsilon_2/\sqrt{2}$, with $\varepsilon_2 \sim \text{Beta}(1, 2)$. Observe that the standard deviation $\sigma = 0.5$ for both ε 's, which allows us to compare the simulation results directly with those of $\varepsilon \sim N(0, 0.5^2)$. We only regard sample size $n = 250$, but keep the number of replications 10,000. As before, the upwards (m_1) and downwards (m_2) fraction of exceedances of $q_\alpha/\sqrt{n}$ and $-q_\alpha/\sqrt{n}$, respectively, is presented. In addition, the fraction of how often there is indeed a jump up of m_1 detected at $x = 0.5$, thus how often the empirical HH-plot at $x = 0.5$ exceeds $q_\alpha(0.95)/\sqrt{n}$, is reported. Similarly, this fraction is also given for the jump down of m_2 .

Most exceedance rates in Table 3 are close to 100% and show that the HH-plot is an effective tool to detect jumps. Only at

Table 2. Simulated Type I error probabilities for X and Y independent

n	$\max(H_{n,\alpha}(x)) \geq q_\alpha(0.95)/\sqrt{n}$	
	$\alpha = 0.1$	$\alpha = 0.2$
250	0.025	0.032
500	0.037	0.044

Table 3. Simulation study as described in Section 4.2 for the regression functions m_1 and m_2

α	0.1	0.2	0.1	0.2	0.1	0.2
m_1	$\max(H_{n,\alpha}(x)) \geq$		$H_{n,\alpha}(0.5) \geq$		$\text{sd}(\hat{\theta})$	
ε	$q_\alpha(0.95)/\sqrt{n}$		$q_\alpha(0.95)/\sqrt{n}$			
$N(0, 0.1^2)$	1.000	1.000	1.000	1.000	0.007	0.007
$N(0, 0.5^2)$	0.993	1.000	0.979	1.000	0.011	0.011
$N(0, 1)$	0.412	0.957	0.243	0.858	0.069	0.038
t_3 with $\sigma = 0.5$	0.994	1.000	0.985	1.000	0.009	0.008
Beta(1, 2) with $\sigma = 0.5$	0.995	1.000	0.983	1.000	0.008	0.006
m_2	$\min(H_{n,\alpha}(x)) \leq$		$H_{n,\alpha}(0.5) \leq$			
ε	$-q_\alpha(0.95)/\sqrt{n}$		$-q_\alpha(0.95)/\sqrt{n}$			
$N(0, 0.1^2)$	1.000	1.000	1.000	1.000	0.005	0.004
$N(0, 0.5^2)$	0.965	1.000	0.931	1.000	0.012	0.012
$N(0, 1)$	0.289	0.737	0.159	0.569	0.070	0.038
t_3 with $\sigma = 0.5$	0.985	1.000	0.970	1.000	0.009	0.008
Beta(1, 2) with $\sigma = 0.5$	0.977	1.000	0.950	1.000	0.009	0.008

$\sigma = 1$, when the picture is very blurry, the effective sample size of $n\alpha = 25$ ($n = 250$, $\alpha = 0.1$) is too small and the HH-plot has less power.

Finally, an ad hoc estimator $\hat{\theta}$ of the jump location $\theta = 0.5$ is given, by taking in every simulation the mean of $\text{argmax}\{H_{n,\alpha}(x) : x \in (Q_{n,1}(\alpha), Q_{n,1}(1-\alpha))\}$ (and, for m_2 , similarly the mean of the argmin). Due to symmetry, this estimator is unbiased. The standard deviations, presented in the last two columns of Table 3, show that this approach of averaging the locations of the extrema leads to a good estimator of the jump location.

In all three aspects, the type of the distribution of ε plays a minor role, whereas the standard deviation is decisive.

4.3 Regression Function as in Gijbels et al. (1999)

The final simulation is an example from Gijbels et al. (1999), with regression function

$$m_3(x) = \begin{cases} \exp\{-2(x-0.35)\} - 1 & \text{if } x \in [0, 0.35) \\ \exp\{-2(x-0.35)\} & \text{if } x \in [0.35, 0.65) \\ \exp\{2(x-0.65)\} + \exp\{-0.6\} - 2 & \text{if } x \in [0.65, 1]. \end{cases} \quad (4)$$

It is depicted in Figure 4(a); a scatterplot with $X \sim UN(0, 1)$, ε standard normal and independent of X , and $n = 250$ is given in Figure 4(b).

In Figure 5, the theoretical HH-plots for m_3 with $X \sim UN(0, 1)$ and $\varepsilon \sim N(0, \sigma^2)$, $\sigma \in \{0.1, 0.5, 1\}$, for both $\alpha = 0.1$ and $\alpha = 0.2$, are given. The features of the HH-plot are clearly shown in Figure 5(a). The possible range of $H_{0,1}$ is $[-0.05, 0.05]$, and that of $H_{0,2}$ is $[-0.10, 0.10]$, and in both cases it is fully used. The negative slope of the first part of m_3 is depicted in the negative, almost horizontal line of the HH-plot for $\alpha = 0.1$ and the negative values of $H_{0,2}$. The latter one starts at $x = 0.2$ and hence already takes the jump upwards into account ($2\alpha = 0.4 > 0.35$), which is depicted by the increasing HH-values. For $\alpha = 0.1$, this increase starts at $x = 0.35 - \alpha = 0.25$. The jump point at $x = 0.35$ is for both coverage levels α depicted by a distinct local maximum. The regression curve after this jump is again decreasing, hence the HH-values become negative, even before $x^- > 0.35$. For $x \in (0.45, 0.55)$, the plot of $H_{0,1}$ only depicts the negative slope, and hence almost reaches its minimal possible value. The jump at $x = 0.65$ is then difficult to see from the HH-plot, because the HH-values for both $\alpha = 0.1$ and $\alpha = 0.2$ only change slightly in a region left of the jump point. The last part of m_3 has a positive slope, hence the HH-values become positive again.

Figures 5(b) and (c) bear similar features as (a). The jumps are indicated by the extrema of the HH-plot, the steep increases and decreases by positive/negative HH-values. But in contrast to panel (a), the jumps down of the regression curve are depicted by distinct minima of the HH-plot. Also observe that the range of the HH-values is getting smaller when σ increases.

In Table 4, the outcomes of the simulations are presented, again for sample size $n = 250$, with columns as described for m_1 and m_2 . The standard deviations of the estimators of the two jump locations, defined in the following, are moved to Table 5. The outcomes of the simulations presented in Table 4 are similar

Figure 4. Panel (a) shows m_3 as given in (4) and panel (b) one simulation of m_3 with ε standard normal, $X \sim UN(0, 1)$, and $n = 250$.

Figure 5. The theoretical HH-plot for m_3 is shown for various standard deviations [(a) $\sigma = 0.1$, (b) $\sigma = 0.5$, and (c) $\sigma = 1$] of the normal error, for $\alpha = 0.1$ (gray line) and $\alpha = 0.2$ (black line).

to the outcomes of the simulations for m_1 and m_2 ; in general, the HH-plot depicts the main features of the regression curve well, but again the combination of a smaller $n\alpha$ with a larger σ results in a lower power.

In Table 5, we compare the biases and standard deviations of the location of the extrema of the HH-statistic with the two estimators of the jump locations in Gijbels et al. (1999). The estimator of the location ($\theta_{\max} = 0.35$) of the jump up is denoted with $\hat{\theta}_{\max}$ and that of the location ($\theta_{\min} = 0.65$) of the jump down with $\hat{\theta}_{\min}$. [Here we have restricted $\operatorname{argmax} H_{n,\alpha}$ in the definition of $\hat{\theta}_{\max}$ to $(Q_{n,1}(\alpha), \frac{1}{2}]$ (and similarly $\operatorname{argmin} H_{n,\alpha}$ to $[\frac{1}{2}, Q_{n,1}(1 - \alpha))$) in order not to confuse the steep increase for large x with the jump at $x = 0.35$.] Similarly, the two estimators in Gijbels et al. (1999) are denoted with $\hat{\theta}_{\max}^{G1}$ and $\hat{\theta}_{\max}^{G2}$, for the jump up, and $\hat{\theta}_{\min}^{G1}$ and $\hat{\theta}_{\min}^{G2}$, for the jump down. These estimators are based on relatively complicated iterative procedures based on a kernel-type diagnostic (and for “G2” additionally local least-squares). For the comparison, we also include $n = 100$.

Although estimation of the jump location is not our main goal, our estimator compares well with Gijbels et al. (1999). For the jump up, for all four estimators, the biases are very small and the standard deviations are somewhat comparable, although for $\sigma = 0.5$, our estimators perform better. For the jump down, σ plays a more prominent role. For $\sigma = 0.1$, “G1” and “G2” are somewhat better than our estimators, mainly due to “smaller” biases. For $\sigma = 0.5$ (a statistically more difficult situation), our estimators perform somewhat better, mainly due to lower standard deviations. We see from this example that our ad hoc estimator is a competing method to detect jump locations, especially when the picture is noisy.

4.4 Limitations of the Method

We conclude this section by sketching two examples that show some limitations of the HH-plot, see Section 2. For convenience, we assume for both examples that the error ε is symmetric and independent of the conditional median (or mean) m .

For the first example, consider a continuous, increasing function m , which on a certain interval increases rather strongly, whereas it is more flat outside this interval. Now let the variance of ε be rather small. The combination of such an m and ε might lead to a false alarm, that is, the empirical HH-plot detects a nonexistent jump up. This is explained by the fact that the empirical HH-plot only counts the *number* of data points in the region $B \cup D$ in Figure 1. It does not take into account the actual Y -values of these data. The just-described example typically generates (too) many observations in $B \cup D$. A small simulation study confirms that in this case, the HH-plot has incorrectly a high probability of detecting jumps. In Section 4.3, essentially the same phenomenon takes place. The function m_3 decreases strongly on $[0.35, 0.65]$ and there is a jump down at 0.65. When σ is as small as 0.1, this leads to very low values of the empirical HH-plot for x around 0.55 and hence the probability of “detecting” the true jump down becomes artificially larger, but the estimator of the jump location suffers from this phenomenon by exhibiting a substantial, negative bias. (If m_3 would decrease on $(0.65, 1]$, then this bias would be reduced, but the standard deviation would be inflated.) The false alarm around 0.55 is due to the fact that the trend of m_3 (to the left of the jump) and the jump have the same direction. For the first jump at $x = 0.35$ (and for m_2 in Section 4.2), our method is

Table 4. Simulation study as described in Section 4.3 for the regression function m_3

σ	$\max(H_{n,\alpha}(x)) \geq q_\alpha(0.95)/\sqrt{n}$		$H_{n,\alpha}(0.35) \geq q_\alpha(0.95)/\sqrt{n}$		$\min(H_{n,\alpha}(x)) \leq -q_\alpha(0.95)/\sqrt{n}$		$H_{n,\alpha}(0.65) \leq -q_\alpha(0.95)/\sqrt{n}$	
	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$
0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.911	0.978	0.821	0.949	0.991	1.000	0.960	1.000
1	0.210	0.378	0.091	0.228	0.389	0.911	0.198	0.726

Table 5. Simulation study as described in Section 4.3 for the regression function m_3

σ	n	$\alpha = 0.1$	$\alpha = 0.2$	$b(\hat{\theta}_{\max}^{G1})$	$b(\hat{\theta}_{\max}^{G2})$	$\alpha = 0.1$	$\alpha = 0.2$	$sd(\hat{\theta}_{\max}^{G1})$	$sd(\hat{\theta}_{\max}^{G2})$
		$b(\hat{\theta}_{\max})$	$b(\hat{\theta}_{\max})$			$sd(\hat{\theta}_{\max})$	$sd(\hat{\theta}_{\max})$		
0.1	100	0.000	0.000	0.003	0.001	0.009	0.008	0.018	0.007
	250	0.000	0.000	0.000	0.000	0.004	0.003	0.009	0.002
	500	0.000	0.000	0.000	0.000	0.002	0.001	0.006	0.001
	100	0.002	0.006	0.008	0.002	0.025	0.026	0.042	0.039
0.5	250	0.001	0.003	0.001	0.002	0.011	0.012	0.027	0.027
	500	0.000	0.002	0.001	0.000	0.006	0.007	0.011	0.008
		$b(\hat{\theta}_{\min})$	$b(\hat{\theta}_{\min})$	$b(\hat{\theta}_{\min}^{G1})$	$b(\hat{\theta}_{\min}^{G2})$	$sd(\hat{\theta}_{\min})$	$sd(\hat{\theta}_{\min})$	$sd(\hat{\theta}_{\min}^{G1})$	$sd(\hat{\theta}_{\min}^{G2})$
0.1	100	-0.042	-0.049	-0.008	-0.002	0.025	0.020	0.029	0.021
	250	-0.019	-0.043	0.000	0.000	0.022	0.020	0.009	0.003
	500	-0.006	-0.034	0.000	0.000	0.011	0.022	0.005	0.001
	100	-0.012	-0.022	-0.017	-0.013	0.028	0.029	0.043	0.040
0.5	250	-0.005	-0.015	-0.009	-0.009	0.013	0.020	0.038	0.036
	500	-0.003	-0.009	-0.002	-0.002	0.007	0.014	0.016	0.016

much more effective since the trend of the regression curve on both sides of the jump is opposite to the sign of the jump.

The second example is a function m with two jumps up that are close together. For simplicity, assume that X is uniformly $(0, 1)$ distributed and that m is piecewise constant, more precisely $m = 0$ on $(0, b)$, 1 on $[b, 1 - b]$, and 2 on $(1 - b, 1)$; here $0 < b < 0.5$. Now a small simulation study shows that in the difficult case that b is quite close to 0.5 (like $b = 0.45$), it is often not possible to distinguish the two jumps, that is, we might detect only one jump. This happens in particular if the variance of ε is small, since in that case the number of data points in the region $B \cup D$ in Figure 1 is close to the maximum.

5. REAL DATA APPLICATIONS

In the following, we use the HH-plot to analyze two datasets. The first dataset has a random regressor, whereas in the second

example, the regressor comes from a fixed design. (The latter situation is typical for the applications presented in literature.) Another real data application is available online in the supplement of this article.

5.1 Random Design: Relation Between Premium and Loss Ratio in Health Insurance

The dataset used in this example consists of premiums (X) and loss ratios (Y) of a large private health insurance portfolio in Germany. There are $n = 192$ different policies, and for every policy, an annual gross premium has been determined and an average loss ratio (loss divided by gross premium) is computed from the realized losses. Clearly, the functional relationship between X and Y is of interest.

Figure 6. Relation between premium and loss ratio in health insurance: (a) HH-plot for $\alpha = 0.2$, (b) scatterplot of loss ratios against premiums; m estimated with locally weighted linear regression (fitted left and right of the jump).

Before estimating this relation, we detect jumps with the HH-plot. Figure 6(a) shows a small region where the plot is below $-q_{0.2}(0.95)/\sqrt{n}$, thus revealing a jump down. More precisely, we find with our ad hoc estimator a value of 1913.5 Euros for the jump location. In addition, the high values of the HH-plot for premiums up till about 600 Euros suggest a steep increase of the regression function on that region.

A locally weighted linear regression fitted left and right of 1913.5 Euros confirms this jump as well as the steep increase, see Figure 6(b). Apart from the very low premium policies, we can split up the policies in “high loss” (left of the jump) and “low loss” (right of the jump). In the first group, there are many policy holders who “use up” the relatively low premium entirely, whereas for the more expensive policies, which also offer a lot, the premium is not used up, since, for example, the covered treatments might not be available when needed. For the insurance company, this analysis yields useful, precise information for future premium setting. This information would be blurred by a continuous estimate of the regression curve.

5.2 Fixed Design: Prague Temperatures

As a second application, we consider the average annual temperatures in Prague from 1775 through 1989 ($n = 215$). These data have a fixed, equidistant design, rather than a random regressor X . In such a case, we observe

$$Y_{n,i} = m\left(\frac{i}{n+1}\right) + \varepsilon_{n,i}, \quad \text{with independent and centered} \\ \varepsilon_{n,i} \sim F_{n,i}, i = 1, \dots, n.$$

If all the $F_{n,i}$, $i = 1, \dots, n$, are equal and sufficiently smooth, then the limiting process is as in Remark 1 and (3); therefore, Table 3.2 remains applicable, for n large enough. In the finite

Table 6. Jump points in the literature

Methods	Years of jump		
Horváth, Kokoszka, and Steinebach (1999)	1835	1893	1927
Antoniadis and Gijbels (2002)	1787	1837	–
Gijbels and Goderniaux (2004)	1786.5	1836.5	1942.5
HH-plot	1830		1932

sample case, the only minor difference between fixed and random design is the domain of $V_{n,\alpha}$, which is I_0 in the above fixed design setting.

In previous analyses of these data, it is found that the number of jump points is two or three; see Table 6. (Note that in Antoniadis and Gijbels (2002), the data are studied only up to 1902.) We depict the HH-plot for $\alpha = 0.2$; the plot for $\alpha = 0.1$ does not yield additional information. From the HH-plot, the change-point recurring in the literature at around 1835 is easily detected. The plot, see Figure 7(b), is for the years 1823–1835 below $-q_{0.2}(0.95)/\sqrt{n}$. Hence, we have found a jump down; according to the spikes, we locate it at 1830. Between 1911 and 1933, we see several exceedances of $q_{0.2}(0.95)/\sqrt{n}$, indicating a jump up of the regression curve. Using the ad hoc estimator, we can locate such a jump at the year 1932. This change is in accordance with changes found in the literature (1927 and 1942.5). The early jump found in the literature (at 1787) is difficult to assess since there are only 12 observations before 1787; this year is outside the domain of the HH-plot.

SUPPLEMENTARY MATERIALS

Proofs and Case Study: This supplement consists of proofs of Theorem 1 and Remark 1 and Nile discharges data example (pdf).

Figure 7. Average annual temperatures in Prague from 1775 to 1989: (a) scatterplot, (b) HH-plot for $\alpha = 0.2$.

ACKNOWLEDGMENTS

We thank Irène Gijbels for providing the Prague temperature data and Otilia Boldea and Günther Sawitzki for stimulating discussions. Furthermore, we are grateful to the editor, the associate editor, and three referees for a careful reading of the manuscript and for constructive criticism that led to significant improvements.

[Received October 2009. Revised December 2011.]

REFERENCES

- Antoniadis, A., and Gijbels, I. (2002), "Detecting Abrupt Changes by Wavelet Methods," *Journal of Nonparametric Statistics*, 14, 7–29. [140,145]
- Dempfle, A., and Stute, W. (2002), "Nonparametric Estimation of a Discontinuity in Regression," *Statistica Neerlandica*, 56, 233–242. [140]
- Gijbels, I., and Goderniaux, A.-C. (2004), "Bandwidth Selection for Change-point Estimation in Nonparametric Regression," *Technometrics*, 46, 76–86. [145]
- Gijbels, I., Hall, P., and Kneip, A. (1999), "On the Estimation of Jump Points in Smooth Curves," *Annals of the Institute of Statistical Mathematics*, 51, 231–251. [140,142,143]
- Gijbels, I., Lambert, A., and Qiu, P. (2007), "Jump-Preserving Regression and Smoothing Using Local Linear Fitting: A Compromise," *Annals of the Institute of Statistical Mathematics*, 59, 235–272. [140]
- Grégoire, G., and Hamrouni, Z. (2002), "Change Point Estimation by Local Linear Smoothing," *Journal of Multivariate Analysis*, 83, 56–83. [140]
- Hall, P., and Titterton, D. M. (1992), "Edge-Preserving and Peak-Preserving Smoothing," *Technometrics*, 34, 429–440. [140]
- Horváth, L., and Kokoszka, P. (2002), "Change-Point Detection With Nonparametric Regression," *Statistics*, 36, 9–31. [140]
- Horváth, L., Kokoszka, P., and Steinebach, J. (1999), "Testing for Changes in Multivariate Dependent Observations With an Application to Temperature Changes," *Journal of Multivariate Analysis*, 68, 96–119. [145]
- Joo, I., and Qiu, P. (2009), "Jump Detection in a Regression Curve and its Derivative," *Technometrics*, 51, 289–305. [140]
- Kim, C. S., and Marron, J. S. (2006), "SiZer for Jump Detection," *Journal of Nonparametric Statistics*, 18, 13–20. [140]
- Loader, C. R. (1996), "Change Point Estimation Using Nonparametric Regression," *The Annals of Statistics*, 24, 1667–1678. [140]
- McDonald, J. A., and Owen, A. B. (1986), "Smoothing With Split Linear Fits," *Technometrics*, 28, 195–208. [140]
- Müller, H.-G. (1992), "Change-Points in Nonparametric Regression Analysis," *The Annals of Statistics*, 20, 737–761. [140]
- Park, C., and Kim, W.-C. (2006), "Wavelet Estimation of a Regression Function With a Sharp Change Point in a Random Design," *Journal of Statistical Planning and Inference*, 136, 2381–2394. [140]
- Sánchez-Borrego, I. R., Martínez-Miranda, M. D., and González-Carmona, A. (2006), "Local Linear Kernel Estimation of the Discontinuous Regression Function," *Computational Statistics*, 21, 557–569. [140]
- Wang, Y. (1995), "Jump and Sharp Cusp Detection by Wavelets," *Biometrika*, 82, 385–397. [140]
- Wu, J. S., and Chu, C.K. (1993), "Kernel-Type Estimation of Jump Points and Values of a Regression Function," *The Annals of Statistics*, 21, 1545–1566. [140]