IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 46–61
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL705

Estimation of irregular probability densities*

Lieven Desmet[1,†], **Irène Gijbels**[2,†] **and Alexandre Lambert**[3]

Katholieke Universiteit Leuven and Université catholique de Louvain

Abstract: This paper deals with nonparametric estimation of an unknown density function which possibly is discontinuous or non-differentiable in an unknown finite number of points. Estimation of such irregular densities is accomplished by viewing the problem as a regression problem and applying recent techniques for estimation of irregular regression curves. Moreover, the method can deal with estimation of densities that have an irregularity at the endpoint(s) of their support. A simulation study compares the performance of the proposed method with those of other methods available in the literature. A further illustration on real data is provided.

1. Introduction

Consider a random variable X with unknown density function f_X. Based on an i.i.d. sample $X_1, X_2, \cdots, X_n$ from X a well-known nonparametric estimator for f_X is the kernel density estimator

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) , \tag{1}$$

with K a kernel function and $h > 0$ a bandwidth parameter. When $f_X(\cdot)$ is continuous at x, then $f_n(x)$ is a consistent estimator of $f_X(x)$. By contrast, in points of discontinuity the estimate will typically smooth out the discontinuous behaviour and will not be consistent (see e. g. [20] and [27]). A particular example here is the case of a density with support $[0, +\infty[$ (for example an exponential density) which is discontinuous at the endpoint 0 of its support. See for example [11]. Several approaches for obtaining consistent estimates of densities at such discontinuous endpoints or boundary points have been proposed in the literature: a reflection method of [25], transformation methods as in [21] and kernel methods with specially

*This research was supported by the IAP research network P6/03, Federal Science Policy, Belgium.

†The first and second author gratefully acknowledge financial support from the GOA/07/04-project of the Research Fund KULeuven.

[1] This work was part of the doctoral research of the first author carried out at the Katholieke Universiteit Leuven; e-mail: lieven.desmet@village.uunet.be

[2] Katholieke Universiteit Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Box 2400, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium; e-mail: Irene.Gijbels@wis.kuleuven.be

[3] This work was initiated during the doctoral research of the third author carried out while at the Institut de Statistique, Université catholique de Louvain; e-mail: alexandrelambert@yahoo.fr

AMS 2000 subject classifications: Primary 62G07; secondary 62G08.

Keywords and phrases: density estimation, irregularities, local linear fitting, variance stabilization.

adapted kernels for the boundary points, as in [17]. There is also a vast literature on detection of locations of discontinuity points in density or regression functions (see e. g. [6], [28], [12], among others, and references therein).

An important issue in kernel density estimation is the choice of the bandwidth. Global and local bandwidth selection procedures have been studied. See [27] and references therein. Papers on local bandwidth selection in kernel density estimation include [23], [24] and [18], among others. See [5] for a comparative study on bandwidth selectors.

In this paper we consider the more general problem of estimating f_X when this function possibly exhibits discontinuities, at the function itself or in its derivative, at certain (unknown) locations at the interior or at the boundary of its support. If the density is continuous but not differentiable at a point x, then the estimate (1) will be consistent but the rate of convergence is slower than at points of continuity. To deal with estimation of densities that possibly show irregularities of the jump type (i. e. discontinuity in the function itself) or of the peak type (i. e. discontinuity in the derivative) we first view the density estimation problem as a regression problem and then apply the technique developed by [10] for regression functions with jump and/or peak irregularities to the resulting regression problem. Of importance is to link the density estimation problem with the regression problem to see how properties of the regression estimation context lead to properties of the resulting density estimator. Viewing density estimation as a regression problem is not new, and has been used in for example [7] and [19] for respectively estimation of densities at boundaries and densities at points of discontinuity. The contribution of this paper consists of dealing with estimation of irregular densities showing jump or peak irregularities at unknown locations. The proposed method also leads to consistent estimation at (discontinuous) boundary points. The method relies on local linear fits. The merits of techniques based on local linear fitting for estimating regression curves and surfaces with irregularities have been largely proven in [13], [14], [11], [9] and [8].

The paper is organized as follows. In Section 2 we recall how binning of the data leads to a regression problem, and we briefly discuss important properties of this regression problem. Section 3 provides insights in how irregularities in the density f_X have an impact on the regression problem. The proposed estimation procedure is discussed in Section 4. The finite sample performance of the method is investigated via a simulation study in Section 5, which includes also comparisons with existing methods, and a real data example.

2. Density estimation formulated in a regression context

2.1. Data binning

Define an interval $[a, b]$ such that essentially no data point X_i fall outside it. Partition the interval $[a, b]$ into N subintervals $\{I_k; k = 1, \cdots, N\}$ of equal length $(b-a)/N$. More precisely, let $I_k = [a + (k-1)\frac{b-a}{N}, a + k\frac{b-a}{N}[$, for $k = 1, \cdots, N-1$, and let the last bin be $I_N = [a + \frac{N-1}{N}(b-a), b]$. Denote by C_k the number of observations in the bin I_k, $k = 1, \cdots, N$. The bin counts $(C_1, \ldots, C_N)$ behave like a multinomial distribution with n trials and probabilities $(\beta_1/N, \ldots, \beta_N/N)$ where $\beta_k := N\int_{a+(b-a)\frac{k-1}{N}}^{a+(b-a)\frac{k}{N}} f_X(x)\,\mathrm{d}x$, $k = 1, \cdots, N$. Denote by $x_k = a + \frac{b-a}{N}(k - \frac{1}{2})$ the center of the bin I_k, $k = 1, \cdots, N$.

Then, asymptotically, for $N = N(n)$ tending to infinity with n, we have that $\beta_k \approx (b-a)f_X(x_k)$. Since the counts $C_k \sim \text{Binomial}(n, \beta_k/N)$, it holds that $\text{E}(C_k) = n\beta_k/N = m\beta_k$, with $m = n/N$, and $\text{Var}(C_k) = n\beta_k/N(1-\beta_k/N) = m\beta_k(1-\beta_k/N)$, and hence asymptotically, as N tends to infinity, $\text{E}\{C_k/((b-a)m)\} \approx f_X(x_k)$ and $\text{Var}\{C_k/((b-a)m)\} \approx f_X(x_k)/((b-a)m)$. Estimating $f_X(x)$ can thus be viewed as a heteroscedastic nonparametric regression problem where the regression curve (the mean regression function) is $f_X(x)$ and the conditional variance function $\sigma^2(x) \approx f_X(x)/m$ with data set $\{(x_k, C_k/((b-a)m), k = 1, \cdots N\}$ as the sample.

We will assume that $m \to \infty$ as $n \to \infty$, meaning that the number of data per bin also increases as the total number of data increases.

For future developments it is convenient to treat the bin counts as Poisson variables. Indeed, the variables $C_k \sim \text{Binomial}(n, \beta_k/N)$ behave asymptotically like Poisson variables with parameter $m\beta_k$ (recall, as $n \to \infty$, we have that $N \to \infty$).

A widely used approach to diminish heteroscedasticity is to apply a variance-stabilizing transformation to the bin counts, which in some sense normalizes their variance to a constant value.

Strictly speaking, the local linear fitting procedure does not require the conditional variance to be constant but its consistency properties are established under continuity. This is not guaranteed when starting from densities with jumps as these will show up in the conditional variance. Due to the variance stabilizing transformation we need however not to worry about this. See Section 3.

2.2. Variance stabilizing transformations

It was suggested already by [2] that the square root of a Poisson variable (say $X \sim \text{Poisson}(\lambda)$ with $\lambda > 0$) has a distribution that is closer to the normal distribution than the original variable. The variance is approximately $1/4$ when λ is large. This idea was further explored in [1], in particular by considering transformations of the type $\sqrt{X+c}$ with $c \geq 0$.

The behaviour of the expectation and the variance of the transformed $\sqrt{X+c}$ Poisson random variable X, for $\lambda \to \infty$, can be obtained via Taylor expansion. The following result can be found in for example [4].

Lemma 1 *Assume $X \sim Poisson(\lambda)$ and $c \geq 0$ is a constant. Then it holds:*

$$\begin{aligned} E(\sqrt{X+c}) &= \lambda^{\frac{1}{2}} + \frac{4c-1}{8}\lambda^{-\frac{1}{2}} - \frac{16c^2-24c+7}{128}\lambda^{-\frac{3}{2}} + O(\lambda^{-\frac{5}{2}}) \\ Var(\sqrt{X+c}) &= \frac{1}{4} + \frac{3-8c}{32}\lambda^{-1} + \frac{32c^2-52c+17}{128}\lambda^{-2} + O(\lambda^{-3}) \,. \end{aligned}$$

In [1] it was proposed to take $c = 3/8$ in order to get a constant variance and nearly constant bias but [4] argue that the choice $c = 1/4$ is better for minimizing the first order bias $E(\sqrt{X+c}) - \sqrt{\lambda}$ while still stabilizing the variance equally well (for λ large enough). In this paper we opt for the choice $c = 1/4$.

2.3. Asymptotic properties of the transformed bin counts

In [4] the behaviour of the transformed bin counts as stochastic variables was studied in detail. That paper establishes an explicit decomposition of the transformed bin counts in a deterministic term directly related to the (square root of the) density in the corresponding grid points, a deterministic $o(1)$ term and a stochastically small random variable. This result extends Lemma 1 and applies it to the binning case where the bin counts C_k are assumed Poisson variables with parameter $m\beta_k$.

Proposition 1 *With notations as before,* $\widetilde{Y}_k = \sqrt{C_k + \frac{1}{4}}$, *we have*

$$\widetilde{Y}_k = \sqrt{m\beta_k} + \varepsilon_k + \frac{1}{2} Z_k + \xi_k, \qquad k = 1, 2, \ldots, N \,, \tag{2}$$

where the Z_k *are i.i.d.* $N(0,1)$ *variables, the* ε_k *are constants that are* $O((m\beta_k)^{-\frac{3}{2}})$, *the quantity* $\sum_{k=1}^{N} \varepsilon_k^2$ *is* $O(1)$ *and the* ξ_k *are independent and stochastically small variables. More, precisely we have:* $E|\xi_k|^\ell \leq c_\ell (m\beta_k)^{-\frac{\ell}{2}}$ *and* $P(|\xi_k| > \alpha) \leq (\alpha^2 m\beta_k)^{-\frac{\ell}{2}}$ *where* $\ell > 0, \alpha > 0$ *and* $c_\ell > 0$ *is a constant (depending on* ℓ *only).*

The authors in [4] rely on this regression model to estimate $f_X(\cdot)$ using wavelet block thresholding techniques. The simulation study in Section 5 includes a comparison with this method.

The above result is of course an asymptotic result requiring that $m\beta_k \to \infty$. Note that then the ε_k are $o(1)$ quantities and the ξ_k are $o_P(1)$. It is thus important that $\beta_k > 0$ while $m \to \infty$. In other words, the result is not applicable for $\beta_k = 0$ as the parameter of a Poisson variable cannot be 0. Consequently, the finite sample behaviour of any estimate using this model could be bad in regions where the true density is zero or close to zero. Therefore, we need to assume, on the domain under consideration, that $\inf f_X(x) > 0$.

3. Variance stabilization and irregularities

We now turn to the situation that the unknown density f_X is continuous and twice differentiable except at a finite (unknown) number of points in which the density function itself or its derivative is discontinuous. A point s is called a jump irregularity when $f_X(s+) = f_X(s-) + d$ with $f_X(s-) > 0$, $f_X(s+) > 0$, and $d \neq 0$. A point s is called a peak irregularity when $f'_X(s+) = f'_X(s-) + d^*$, with $d^* \neq 0$ and $f_X(s+) = f_X(s-) > 0$, where f'_X denotes the first derivative of f_X. We assume that the second order derivatives of f_X at all regular points (i. e. points at which f_X is continuous and twice differentiable) are uniformly bounded. We now investigate what is the impact of such irregularities on the regression problem related to the transformed counts.

The following result shows how the asymptotic variance changes with the grid point x_k. It is an immediate consequence of Lemma 1.

Corollary 1 *Let* C_k *be the bin counts and suppose that* x_k *and* x_{k+1} *are in the interior of the support of* f_X*. Then the asymptotic difference in variance over these neighbouring grid points behaves like:*

$$\Delta\, Var_k := Var(\sqrt{C_k + \frac{1}{4}}) - Var(\sqrt{C_{k+1} + \frac{1}{4}}) = \frac{1}{m} \frac{3 - 8c}{32} (\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}}) + o(1/m) \,.$$

Proof. Apply Lemma 1 to the variables C_k that are distributed as Poisson variables with parameter $m\beta_k$. Then we have $\mathrm{Var}(\sqrt{C_k + c}) = \frac{1}{4} + \frac{3-8c}{32} \frac{1}{m\beta_k} + o(1/m)$. The result follows by rewriting this equation in the neighbouring point with index $k+1$ and taking the difference. □

From the result in Corollary 1 we get insight into the effect of the variance stabilisation on the behaviour of the conditional variance function in the regression problem, and more particularly on how this variance changes with the x-coordinate. We first study $\Delta\mathrm{Var}_k$, with x_k and x_{k+1} interior points of the support of f_X, for different situations, namely that the interval $]x_k, x_{k+1}[$: (S1) does not contain any irregularity point; (S2) contains a jump irregularity point s; and (S3) contains a peak irregularity point s. The findings can be summarized as follows:

> **(S1)**. We have that $|f_X(x_k) - f_X(x_{k+1})| = O(1/N)$ and since asymptotically $\beta_k \to (b-a)f_X(x_k)$ we have that $(\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}}) \to 0$ as well as $1/m \to 0$. Therefore $\Delta\mathrm{Var}_k$ vanishes asymptotically, or in other words the variance in smooth regions of the density f_X tends to behave like a constant.
> **(S2)**. In this situation we have $f_X(x_k) = f_X(s-) + O(1/N)$ and $f_X(x_{k+1}) = f_X(s-) + d + O(1/N)$ and a first order approximation of the $(\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}})$ term is given by $d[(b-a)\, f_X(s-)(f_X(s-)+d)]^{-1}$. However, since $1/m \to 0$, the quantity $\Delta\mathrm{Var}_k$ will converge to zero although slower than in situation **(S1)** (and (**S3**)).
> **(S3)**. In this case, an analysis similar to the one in **(S1)** applies, and the difference in variance $\Delta\mathrm{Var}_k$ vanishes asymptotically.

The case when the unknown density shows a jump discontinuity at an endpoint of its support is discussed in Section 4.2.

4. Proposed estimation procedure

4.1. Jumps and peaks preserving fit

In [10] a nonparametric method for estimating regression curves with jump and/or peak irregularities using local linear fitting was proposed. The aim is to apply this method to the regression model obtained from the binned and transformed data.

The requirement of homoscedastic errors in [10] can be relaxed since it is sufficient to have a continuous (locally constant) conditional variance for the method to work. From the result in (2) and the discussion in Section 3 we know that the regression model for $(x_k, \widetilde{Y}_k)$ has a less heteroscedastic conditional variance, and the effect of irregularities in the interior of the support vanishes asymptotically.

We need to assume that $m = n/N \to \infty$; thus the number of observations per bin grows as the number of observations grows.

From the transformed bin counts $\widetilde{Y}_k = \sqrt{C_k + \frac{1}{4}}$ we can effectively estimate the function $g(\cdot)$ that relates to the original density $f_X(\cdot)$ as follows: $g(x) \approx \sqrt{m(b-a)f_X(x) + \frac{1}{4}}$. Once an estimate for $g(\cdot)$ is obtained we recover an estimate for $f_X(\cdot)$ by applying an inverse transformation.

In summary, the estimation procedure reads as follows:

- Step 1. Binning step: set up the grid of N equal-length intervals and calculate the bin counts C_k, $k = 1, \ldots, N$.
- Step 2. Root transform: put $\widetilde{Y}_k = \sqrt{C_k + \frac{1}{4}}$ and treat $(x_k, \widetilde{Y}_k)$, $k = 1, \ldots, N$ as the new equispaced sample for a nonparametric regression problem.
- Step 3. Apply the jump and peak preserving local linear fit of [10] to obtain an estimate $\widehat{g}(\cdot)$ of $g(\cdot)$.
- Step 4. Perform an inverse transformation and renormalization

$$\widehat{f}_X(.) = S\,(\widehat{g}^2(.) - \frac{1}{4})_+\,, \tag{3}$$

where $z_+ = \max(z, 0)$ and S is a normalization constant.

The jump and peak preserving local linear fitting method of [10] consists of fitting three local linear models, using observations in a centered, a right and a left neighbourhood of the point. In the presence of a jump or peak irregularity, one of the three fits will outperform the other two, and this fit is selected in a data driven way using an appropriate diagnostic quantity. We now provide details of this estimation algorithm in Step 3. Let K_c be a bounded symmetric kernel density function supported on the interval $[-1/2, 1/2]$, and let $h > 0$ be the bandwidth parameter. The (conventional) local linear estimate for $g(x)$ is obtained by weighted least-squares minimization:

$$(4) \qquad (\widehat{a}_{c,0}(x), \widehat{a}_{c,1}(x)) = \arg\min_{a_0, a_1} \sum_{k=1}^{N} \left[\widetilde{Y}_k - a_0 - a_1(x_k - x)\right]^2 K_c\left(\frac{x_k - x}{h}\right) .$$

Starting from this conventional kernel K_c one then considers one-sided versions $K_\ell(x) = K_c(x)\, I\{x \in [-1/2, 0\,[\,\}$ and $K_r(x) = K_c(x)\, I\{x \in [0, 1/2]\,\}$ which via a weighted least-squares minimization as in (4) but with $K = K_\ell$, respectively $K = K_r$, leads to the left local linear estimate, respectively the right local linear estimate, denoted by $(\widehat{a}_{j,0}(x), \widehat{a}_{j,1}(x))$ with $j = \ell, r$ respectively.

Consider the Residual Sum of Squares (RSS) of the three fits, defined as:

$$(5) \quad \mathrm{RSS}_j(x) = \sum_{k=1}^{N} \left[\widetilde{Y}_k - \widehat{a}_{j,0}(x) - \widehat{a}_{j,1}(x)(x_k - x)\right]^2 K_j\left(\frac{x_k - x}{h}\right) , \quad j = c, \ell, r .$$

Then an important diagnostic quantity is

$$(6) \qquad \mathrm{diff}(x) = \max\left(\frac{\mathrm{RSS}_c(x)}{w_c(x)} - \frac{\mathrm{RSS}_\ell(x)}{w_\ell(x)}, \frac{\mathrm{RSS}_c(x)}{w_c(x)} - \frac{\mathrm{RSS}_r(x)}{w_r(x)}\right) ,$$

where $w_j(x) = \sum_{k=1}^{N} K_j\left(\frac{x_k - x}{h}\right)$, for $j = c, \ell, r$. The peak and jump preserving local linear regression estimator is then given by

$$(7) \quad \widehat{g}(x) = \begin{cases} \widehat{a}_{c,0}(x) & \text{if} \quad \mathrm{diff}(x) < u \\ \widehat{a}_{r,0}(x) & \text{if} \quad \mathrm{diff}(x) \geq u \text{ and } \frac{RSS_r(x)}{w_r(x)} < \frac{RSS_\ell(x)}{w_\ell(x)} \\ \widehat{a}_{\ell,0}(x) & \text{if} \quad \mathrm{diff}(x) \geq u \text{ and } \frac{RSS_r(x)}{w_r(x)} > \frac{RSS_\ell(x)}{w_\ell(x)} \\ (\widehat{a}_{\ell,0}(x) + \widehat{a}_{r,0}(x))/2 & \text{if} \quad \mathrm{diff}(x) \geq u \text{ and } \frac{RSS_r(x)}{w_r(x)} = \frac{RSS_\ell(x)}{w_\ell(x)} , \end{cases}$$

where $u > 0$ is a suitably chosen threshold value.

Together with good choices of the parameters h and u involved, this leads to the following practical estimation algorithm:

Consider a grid of bandwidths $h_{\text{grid}} := (h_1, \ldots, h_M)$.
Iterate over these bandwidths and put $h := h_q$, $q = 1, \ldots, M$.
For this bandwidth:

- ◇ Calculate estimates $\widehat{a}_{j,0}(x)$ and $\widehat{a}_{j,1}(x)$ for $j = \ell, r, c$.
- ◇ Obtain $\widehat{d} := \sup_x |\widehat{a}_{r,0}(x) - \widehat{a}_{\ell,0}(x)|$ and $\widehat{d^*} := \sup_x |\widehat{a}_{r,1}(x) - \widehat{a}_{\ell,1}(x)|$.

- ◇ Put $u_{\max} := \frac{1}{2}\left(\widehat{d}^2\, \frac{C_0^c(0)}{v_{0,c}} + \widehat{d}^{*\,2}\, \frac{C_2^c(0)}{v_{0,c}}\, h^2\right)$, with $v_{0,c} = \int_{-1/2}^{1/2} K_c(t)\,\mathrm{d}t$ and with $C_0^c(0)$ and $C_2^c(0)$ constants that only depend on K (see [10] for details).
- ◇ Put $u_{\text{grid}} := (0.001u_{\max}, 0.01u_{\max}, 0.1u_{\max}, u_{\max})$.
  Now iterate over the threshold values and put $u := u_p$, $p = 1, \ldots, 4$.
  - ∗ For the combination of h and u values at hand, calculate $\widehat{g}^{-k}(x_k)$ as in (7), but leaving out the k-th observation itself.
  - ∗ Calculate $\sum_{k=1}^{n}[\widetilde{Y_k} - \widehat{g}^{-k}(x_k)]^2$.
- ◇ Retain the value of u that yields the minimum for the sum in the former step and associate with h_q by putting it $\widetilde{u}_q$.

Repeat the above procedure for each bandwidth and look for the bandwidth h_q (and associated threshold $\widetilde{u}_q$) that yields the lowest value for the sum.
Calculate the final estimate with (7) from the couple (h, u) obtained as above.

For a detailed study of this jump and peak preserving estimator, in a general regression context, see [10]. From this and previous studies we need to impose conditions on how the bandwidth decreases as $N \to \infty$. More precisely, we need to impose that $h \sim (\log N)^{2/5} N^{-1/5}$, which can be translated to a condition on n depending on the relation between N and n.

From the discussion in Section 3 it is already clear that the above estimation procedure can deal with estimation of irregular densities at the interior of their support. We now show that the method can also handle a non-smooth behaviour of the density at an unknown boundary.

4.2. Densities with discontinuity at the boundary

As mentioned before a boundary point can be seen as a potential jump in the regression function to be estimated with the jump and peak preserving local linear fit of Section 4.1. In practice, we take a large enough binning interval (extending to the left of the smallest and to the right of the largest observation) and consider the unknown density as a function defined on this whole interval (coinciding with the density on its support and with value zero outside of the support).

Let s be a boundary point of the support of f_X, and suppose that $f_X(\cdot)$ is discontinuous in s, i. e. $f_X(s-) = 0$, and $f_X(s+) = d_{\mathrm{B}} > 0$, and we have uniformly bounded derivatives up to the second order outside of s. Then to the left of s the bin counts have variance zero (since they remain zero themselves) and to the right of s we see the variance converging to 1/4. Therefore, asymptotically, the jump discontinuity in the variance cannot be resolved by a variance stabilizing transformation.

The proposed method however can deal with this situation in an automatic way. The jump and peak preserving estimator from Section 4.1 will select the suitable one-sided local linear fit in the neighbourhood of the boundary, and hence will estimate the jump correctly. The argumentation for this is in two steps: first we analyse this problem in the regression context in Lemma 2 and then we apply this to the density estimation setting.

Lemma 2 *Consider a regression model $Y_i = m(x_i) + \varepsilon_i$ where $m(\cdot)$ is an unknown function such that $m(x) = 0$ for $x < s$ and $m(s+) = d > 0$ (and m has continuous second order derivatives outside of s), the errors have constant variance σ^2 for*

$x_i > s$ (and are 0 for $x_i < s$), with $E\varepsilon^4 < \infty$. Assume the kernel K is uniform Lipschitz continuous and $h \to 0$, $\frac{nh}{\log n} \to \infty$ as $n \to \infty$. Then asymptotically, we have the following behaviour of the residual sum of squares quantities, in points $x = s + \tau h$ near the jump point s.

	$-1/2 < \tau \leq 0$	$0 < \tau < 1/2$
$\frac{\mathrm{RSS}_c(x)}{w_c(x)}$	$d^2 \frac{C_0^c(\tau)}{v_{0,c}} + \frac{v_{0,c}^{\tau,+}}{v_{0,c}} \sigma^2 + o(1)\,a.s.$	$d^2 \frac{C_0^c(\tau)}{v_{0,c}} + \frac{v_{0,c}^{\tau,+}}{v_{0,c}} \sigma^2 + o(1)\,a.s.$
$\frac{\mathrm{RSS}_r(x)}{w_r(x)}$	$d^2 \frac{C_0^r(\tau)}{v_{0,r}} + \frac{v_{0,r}^{\tau,+}}{v_{0,r}} \sigma^2 + o(1)\,a.s.$	$\sigma^2 + o(1)\,a.s.$
$\frac{\mathrm{RSS}_\ell(x)}{w_\ell(x)}$	$o(1)\,a.s.$	$d^2 \frac{C_0^\ell(\tau)}{v_{0,\ell}} + \frac{v_{0,\ell}^{\tau,+}}{v_{0,\ell}} \sigma^2 + o(1)\,a.s.$

where asymptotic remainder terms are uniform in x and with $v_{0,j}^{\tau,+} := \int 1_{-\tau}^{1/2} K_j(t)\,\mathrm{d}t$ for $j = \ell, r, c$.

The proof of Lemma 2 is omitted here, and can be found in [8].

We cannot immediately apply this result to our density estimation setting where responses $\widetilde{Y}_k$ are obtained from transformed bin counts. However, asymptotically we do have conditions as in the lemma: for $x_k < s$ we have $C_k = 0$, $\widetilde{Y}_k = 0.5$ and $\mathrm{Var}\widetilde{Y}_k = 0$, whereas for $x_k > s$, asymptotically $\widetilde{Y}_k = \sqrt{m\beta_k} + \frac{1}{2} Z_k + o_P(1)$, with Z_k standard normal variables as in (2).

The jump d and the quantity σ^2 in the lemma then correspond to $(\sqrt{m(b-a)d_{\mathrm{B}}} - 0.5)$, respectively $1/4$ in our setting. Asymptotically, as $m \to \infty$, the contribution of the jump increases unboundedly. Therefore, considering (7) and the definition of $\mathrm{diff}(x)$ in (6), we have for $-1/2 < \tau \leq 0$, $\mathrm{diff}(x) = \frac{\mathrm{RSS}_c(x)}{w_c(x)} - \frac{\mathrm{RSS}_\ell(x)}{w_\ell(x)}$ which increases asymptotically above threshold values and clearly $\frac{\mathrm{RSS}_r(x)}{w_r(x)} > \frac{\mathrm{RSS}_\ell(x)}{w_\ell(x)}$ so the left estimate will be selected. Now for $0 < \tau < 1/2$ we will see $\mathrm{diff}(x) = \frac{\mathrm{RSS}_c(x)}{w_c(x)} - \frac{\mathrm{RSS}_r(x)}{w_r(x)}$ increase above threshold and since $\frac{\mathrm{RSS}_\ell(x)}{w_\ell(x)} > \frac{\mathrm{RSS}_r(x)}{w_r(x)}$, the right estimate will be selected.

5. Numerical analysis

5.1. Simulation study

The proposed estimation method is applied to five test densities with jump and/or peak irregularities in the interior or with discontinuous boundary.

Model (a) is a discontinuous density defined from two different exponential densities.

$$f_X(x) = 0.5\exp(x)\, I\{x < 0\} + 5\exp(-10\,x)\, I\{x \geq 0\}\,.$$

Model (b) is a discontinuous density which is a mixture of two different normal densities and was considered in [15]:

$$f_X(x) = 0.5 f_{\mathrm{N}(0,(\frac{10}{3})^2)}\, I\{x < 0\} + 0.5 f_{\mathrm{N}(0,(\frac{32}{3})^2)}\, I\{x \geq 0\}\,.$$

Model (c) is the claw density defined in [22] (their model #10). It can be seen as a convex combination of normal densities:

$$\begin{aligned} f_X(x) \;&=\; \frac{1}{2} f_{\mathrm{N}(0,1)}(x) + \frac{1}{10}\Big(f_{\mathrm{N}(-1,(\frac{1}{10})^2)}(x) + f_{\mathrm{N}(-\frac{1}{2},(\frac{1}{10})^2)}(x) + f_{\mathrm{N}(0,(\frac{1}{10})^2)}(x) \\ &\quad + f_{\mathrm{N}(\frac{1}{2},(\frac{1}{10})^2)}(x) + f_{\mathrm{N}(1,(\frac{1}{10})^2)}(x)\Big)\,. \end{aligned}$$

Strictly speaking this is a smooth model but it is challenging.
Model (d) is the standard exponential density (so with a discontinuity at the boundary).
Model (e) is a density with discontinuity in the first derivative.

$$f_X(x) = 5\exp(-|10x|) \ .$$

All these models have unbounded support (on at least one side), and are shown in Figure 1.

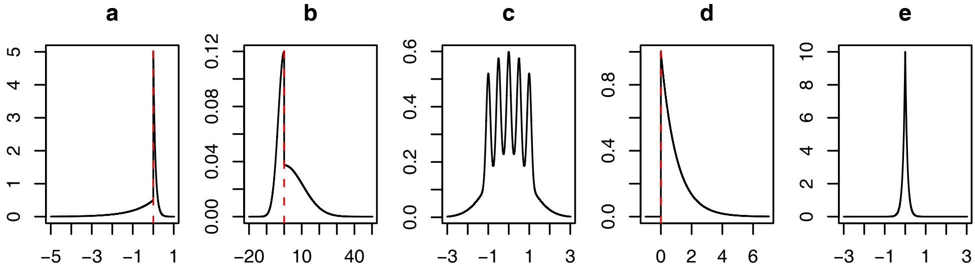

FIG. 1. The five test models.

An illustration of the effect of the variance stabilization is provided in Figure 2. Hundred samples of size $n = 16384$ are generated from each model. Each sample is binned into $N = 256$ bins. In each gridpoint we thus have a sample of bin counts and transformed bin counts of size 100 (from the 100 repetitions), from which the sample standard deviations are then calculated.

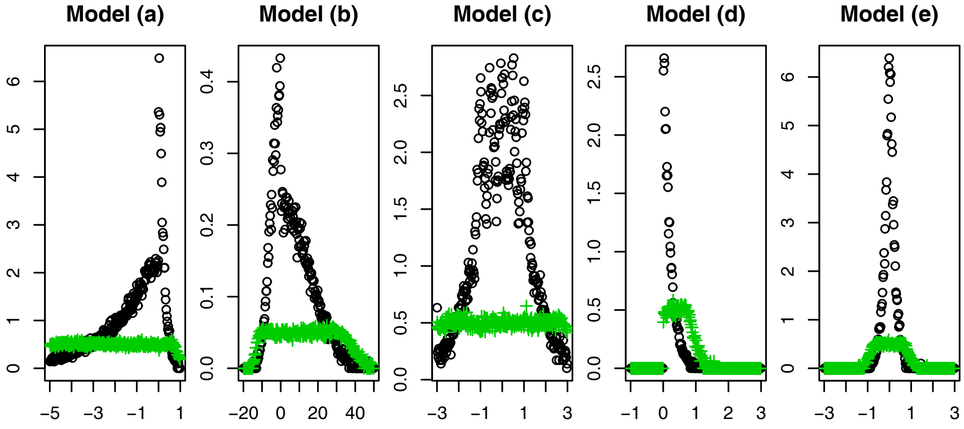

FIG. 2. Variance stabilization in each model. Black circles indicate standard deviations based on bin counts C_k, grey crosses show standard deviations based on transformed bin counts $\sqrt{C_k + \frac{1}{4}}$ for a large value $m = 64$.

As can be seen from Figure 2 the original bin counts are strongly heteroscedastic and the standard deviations follow the shape of the density itself, as explained in Section 2.1. This greatly improves when taking a transformation: peaks get largely suppressed and discontinuities also diminish. However, in those regions where the density was already small (near zero) we still have small values after transformation and hence standard deviation is still far from the theoretical value. In addition, a discontinuity at the boundary still gives rise to a discontinuity of magnitude 0.5 in the standard deviation of the transformed values (see Model (d)). However this does not cause any problem, as explained in Section 4.2.

In this simulation study we include a comparison with a variety of other methods, such as standard kernel density estimation methods (see the estimator in (1)) with different bandwidth selection strategies as well as methods developed for densities with irregularities such as wavelet thresholding and a histogram method combined with a suitable selection of the number of bins. An overview of the considered estimators and their short notation is given in Table 1.

Table 1
Overview of estimators

Name	Method	Input data	Main smoothing parameter
$\widehat{f}_1$	proposed estimator	binned transformed	global bandwidth (cross-validation)
	kernel	raw data	global bandwidth:
$\widehat{f}_2$			Sheather-Jones solve-the-equation (ste)
$\widehat{f}_3$			Sheather-Jones direct plug-in (dpi)
$\widehat{f}_4$	conventional local linear	binned transformed	local bandwidth
$\widehat{f}_5$	wavelet	raw data	thresholding
$\widehat{f}_6$	wavelet	binned transformed	blocked thresholding
	histogram	raw data	number of bins:
$\widehat{f}_7$			penalized max likelihood (Hellinger distance)
$\widehat{f}_8$			penalized max likelihood (L_2 distance)

Details about the methods are provided below.

- The proposed estimator is denoted by $\widehat{f}_1 = \widehat{f}_X$, defined in (3), and is obtained via the fully automatic procedure described in Section 4.1. We use an equispaced grid of bandwidth values $h_{\text{grid}} = \{0.02 + (q-1)0.01; q = 1, \cdots, 13\}$.
- Methods $\widehat{f}_2$ and $\widehat{f}_3$ are kernel density estimators based on the Sheather-Jones bandwidth selectors, respectively with direct plug-in and solve-the-equation strategies, as implemented in `R: stats` package. See [26].
- Method $\widehat{f}_4$ is the estimate obtained from local kernel regression with a variable bandwidth as in [16] (package `R: lokern`).
- Estimator $\widehat{f}_5$ is a recent wavelet thresholding method of [15]. As recommended in that paper we use the Haar wavelet basis for which the theory was developed as well as the guidelines on the finest resolution level. An important procedure parameter is then still the p-value for the testing procedure, for which no guidelines are given. The results reported here are for $p = 0.05$ (which gave the best performance in the majority of cases).
- In $\widehat{f}_6$, binned and transformed data are used as a model for $\sqrt{f}$ (up to a scaling factor). The block thresholding wavelet method yields $\widehat{\sqrt{f}}$ and the final estimate is obtained by squaring and renormalization (see [4]). The parameter λ_* in the James-Stein shrinkage formula regulates thresholding. The standard value of 4.50524 recommended in the paper gives only small amount of smoothing (visually the estimates were quite wiggly), therefore simulations

were also done for 10 and 100 times this value. The reported results are for $\lambda_* = 10 \times 4.50524$.

- Methods $\widehat{f}_7$ and $\widehat{f}_8$ are histogram methods developed by [3] where the number of bins is selected by maximization of a maximum likelihood criterium (respectively based on Hellinger or L_2 distance) over a grid of values namely from 10 to 100 (steps of 2) or from 100 to 800 (steps of 10).

In the simulation study hundred replications were performed and in one replication a sample of size n was generated from the given distribution. For the methods that are based on regression, data were binned over a number N of bins: for samples sizes n 2048, 1024 and 512, the number of bins N are respectively 512, 256 and 128.

We now summarize the simulation results. For saving space we only present plots for Model (a) and sample size $n = 1024$. These pictures provide information on the performance of each method, including its variability. For each method we present pointwise 10% and 90% quantiles and median values calculated from the 100 estimation values. For increasing the visibility at the peak irregularity at the point zero, we add short horizontal segments at that location.

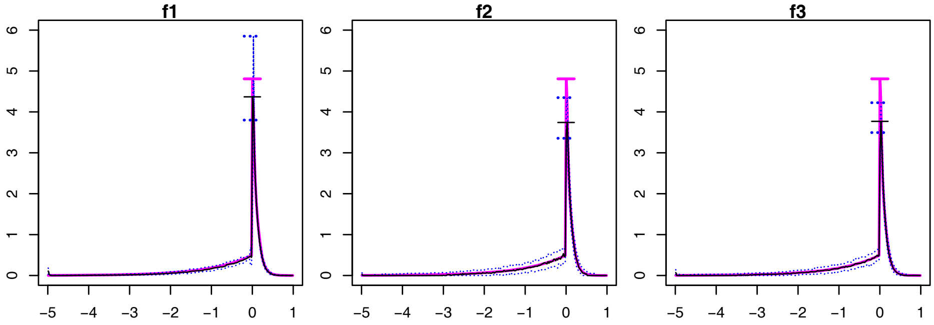

FIG. 3. 10% and 90% percentiles (dotted lines), median (black solid line) and true model (thick grey line). Left panel: $\widehat{f}_1$, middle panel: $\widehat{f}_2$ and right panel: $\widehat{f}_3$.

Figure 3 presents the results for the proposed jump and peak preserving local linear method ($\widehat{f}_1$) and for the global bandwidth kernel methods ($\widehat{f}_2$ and $\widehat{f}_3$). From this figure it can be seen that $\widehat{f}_1$ shows reasonably low bias and low variance (except near the irregularity where the gap between quantiles is larger). The estimates $\widehat{f}_2$, $\widehat{f}_3$ have higher variance in the smooth regions and both underestimate the irregularity (unlike for $\widehat{f}_1$, the true model value falls outside the 10% to 90% quantile interval). In general we noticed that the cross-validation procedure selects significantly larger bandwidths than the Sheather-Jones bandwidth selectors. However, bias is still reduced thanks to one-sided estimation in the jump and peak preserving procedure. Outside of the irregularities, variance is kept low thanks to the larger bandwidth.

Using a local bandwidth parameter (estimate $\widehat{f}_4$) introduces some artifacts as can be seen from Figure 4. This happens in all models except in Model (c). The artifacts are related to jumps in the local bandwidth selection taking place in the transition from flat regions (large selected bandwidth) to regions with higher density values (more reasonable smaller bandwidth values are selected). Local bandwidth selection around irregularities behaves as one would expect as can be seen in the right panel of Figure 4. Across all models, the variance of $\widehat{f}_4$ is comparable to that of $\widehat{f}_1$ or slightly larger. In Model (a) the variance is larger for $\widehat{f}_4$ than for $\widehat{f}_1$. The

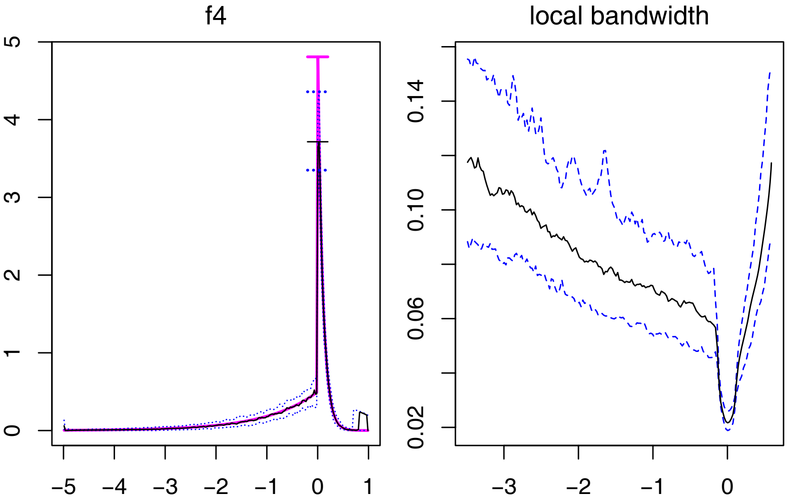

FIG. 4. Left panel: 10% and 90% quantiles (dotted lines), median (black solid line) and true model (thick grey line) for $\widehat{f}_4$. Right panel: selected local bandwidth 10%, 50% and 90% quantiles.

bias for $\widehat{f}_4$ is comparable with that of $\widehat{f}_2$ and $\widehat{f}_3$.

For results on Model (a) for the wavelet threshold method (estimate $\widehat{f}_5$ of [15]), see Figure 5 (left panel). In terms of bias this wavelet method does a rather poor job, in particular in Models (a), (c), (d) and (e), where the true model values at irregular points fall outside of the band delimited by 10% and 90% quantiles (not all plots are shown here). The variability is also quite large in certain models.

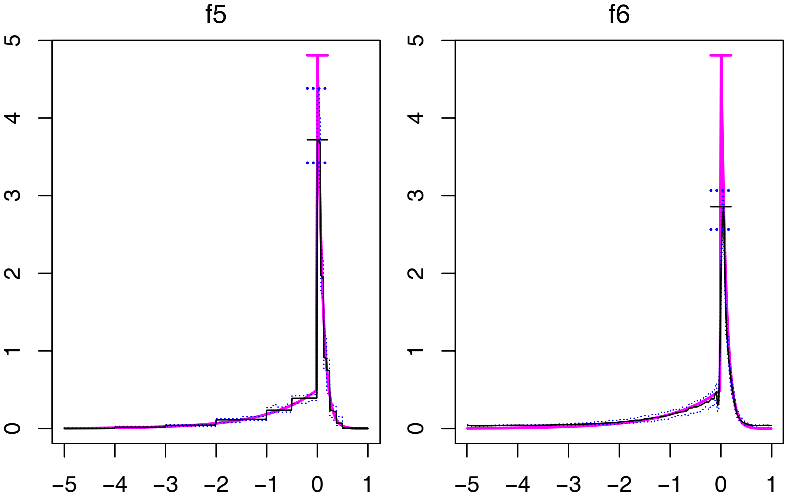

FIG. 5. Left panel: 10% and 90% quantiles (dotted lines), median (black solid line) and true model (thick grey line) for $\widehat{f}_5$. Right panel: same for $\widehat{f}_6$.

The blocked wavelet thresholding estimate $\widehat{f}_6$ is based on squaring the estimate obtained from the binned transformed data. This approach introduces a systematic bias in the baseline (bin counts of zero are transformed to a value of 0.5, squaring and rescaling still yields a non-zero value). Especially in Models (b), (c) and (d) this effect was visible (due to the scale of these models). In general the performance of this blocked wavelet estimate $\widehat{f}_6$ was rather poor. A possible explanation is again the bias in the baseline, which in turn causes bias in other regions when doing the normalization step.

The histogram methods of [3] (estimates $\widehat{f}_7$ and $\widehat{f}_8$) perform quite well. See Figure 6 for results for Model (a), showing a better performance for $\widehat{f}_8$ than for $\widehat{f}_7$ at the discontinuity location. In general the variant $\widehat{f}_8$ based on an L_2 measure, selected a larger number of bins (resulting into better bias properties but a larger variance). Except for Model (e) the bias is indeed quite good. For these models, the method based on L_2 outperforms the recommended one, both in terms of bias and MISE (see also Table 2).

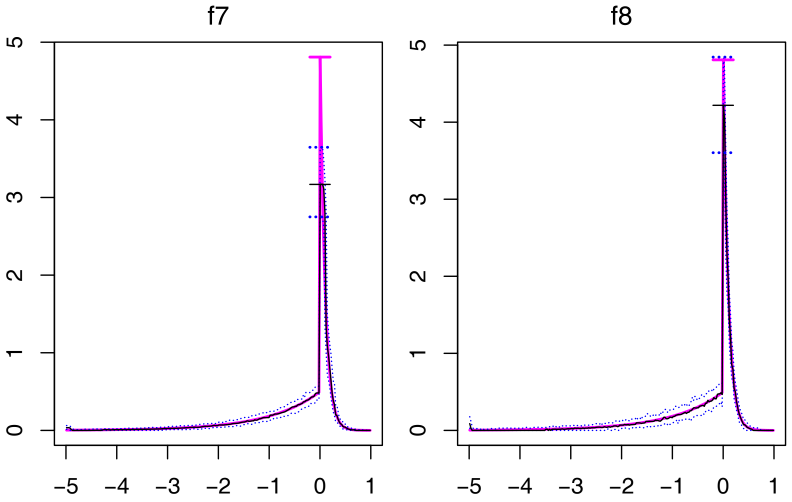

Fig. 6. Left panel: 10% and 90% quantiles (dotted lines), median (black solid line) and true model (thick grey line) for $\widehat{f}_7$. Right panel: same for $\widehat{f}_8$.

Table 2
MISE values for n=2048 and n=512.

	Model (a).		Model (b).		Model (c).		Model (d).		Model (e).	
n	2048	512	2048	512	2048	512	2048	512	2048	512
$\widehat{f}_1$	0.01929	0.0973	0.001084	0.002156	0.006986	0.01169	0.00593	0.01375	2.122	2.407
$\widehat{f}_2$	0.05620	0.08888	(0.02926)	(0.1216)	0.01032	0.05013	0.01254	0.02253	2.474	2.516
$\widehat{f}_3$	0.05609	0.1273	0.001887	0.002853	0.008586	0.01456	0.01153	0.01961	2.472	2.501
$\widehat{f}_4$	0.04827	0.08178	0.001222	0.002719	0.006762	0.02012	0.02916	0.01974	2.558	2.4851
$\widehat{f}_5$	0.04907	0.1257	0.0009915	0.002741	0.016394	0.04696	0.009218	0.06274	2.598	2.827
$\widehat{f}_6$	0.07452	0.1041	0.002454	0.003847	0.01740	0.01826	0.02551	0.03288	3.407	3.503
$\hat{f}_7$	0.06238	0.1448	0.001023	0.003562	0.01195	0.03666	0.007062	0.01451	2.812	3.124
$\widehat{f}_8$	0.02697	0.09854	0.0007281	0.002595	0.01055	0.02696	0.005029	0.01128	2.521	2.412

In Table 2 we provide the MISE (Mean Integrated Squared Error) values for all models for sample sizes $n = 2048$ and $n = 512$. From this table it is seen that $\widehat{f}_1$

has the best performance in many models (for example in the challenging Model (e)) or it has very competitive performance. If it is outperformed, then this is by $\widehat{f}_8$. The latter estimate has good to very good performance in Models (a), (b) and (d). The proposed estimate $\widehat{f}_1$ is doing quite well overall, far better than $\widehat{f}_3$, $\widehat{f}_5$ and $\widehat{f}_6$.

As for specific methods: among the Sheather-Jones global bandwidth methods, $\widehat{f}_3$ (direct plug-in, with larger selected bandwidths) shows better MISE (some values for $\widehat{f}_2$ were unreliable due to convergence problems and therefore put between parentheses). It is not surprising that $\widehat{f}_3$ is doing well in smooth models such as model (c), however from pictures its inconsistency at jumps and unsatisfactory behaviour at peaks is clearly visible (see Figure 3 for Model (a)). For the local bandwidth type kernel estimate $\widehat{f}_4$, note the low value for Model (c) and the high value for Model (d) ($n = 2048$), probably due to the artifacts mentioned before. Finally, in the histogram methods $\widehat{f}_8$ outperforms $\widehat{f}_7$ also in terms of MISE (the former method selects generally a larger number of bins).

The effect of sample size is also clearly visible: MISE values are generally larger for the smaller sample size, in line with a general decline in variance and bias performance noticed for smaller sample size.

5.2. Data example: call center data

The data example concerns data gathered between January 1st and December 31st of 1999 in the call-center of "Anonymous Bank" in Israel. We gratefully acknowledge Prof. Avisham Mandelbaum and Dr. Ilan Guedj from Technion University at Haifa for making the data freely accessible.

The dataset, organized per month, contains some 20000–40000 records on phone calls made to the call center. Among many other features recorded we focus on the time the call entered the system. We use data for the month of May, concerning 39553 phonecalls.

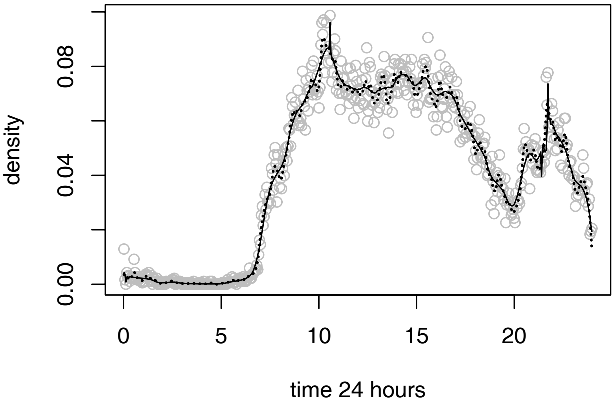

FIG. 7. Black solid line: proposed estimate $\hat{f}_1$, black dotted line: kernel estimate $\widehat{f}_3$.

In Figure 7 data are plotted together with two density estimates: the proposed estimator $\widehat{f}_1$ and the kernel density estimate $\widehat{f}_3$ based on Sheather-Jones direct plug-in bandwidth (of value 0.264; the solve-the-equation bandwidth yields a bandwidth of 0.297 and $\widehat{f}_2$ is very similar to $\widehat{f}_3$). The bandwidth selected in the cross-validation procedure was 0.72. This results in a smooth curve except for some peak features. In contrast, the estimate $\widehat{f}_3$ based on a smaller bandwidth produces a rather wiggly curve (probably too wiggly to reflect the true underlying density). The estimate $\widehat{f}_1$ shows a smoothly ascending curve (starting shortly after 7am, time at which the call center begins to be staffed), leading to a peak between 10 and 11 when people seem to be most keen on thinking about banking. After the peak, the density decreases to a plateau in the early afternoon and then descends further to reach a minimum around 8pm. After this, the density increases again peaking around 10pm, which may be related to phone rates in Israel which change at that time. The call center stops being staffed at midnight.

References

[1] Anscombe, F. J. (1948). The transformation of Poisson, Binomial and Negative-Binomial data. *Biometrika* **35** 246–254.

[2] Bartlett, M. S. (1936). The square root transformation in the analysis of variance. *Journal of the Royal Statistical Society, Supplement* **3** 68.

[3] Birgé, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM Probability and Statistics* **10** 24–45.

[4] Brown L., Cai T., Zhang R., Zhao L., and Zhou H. (2010). The Root-Unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields* **146** 401–433.

[5] Cao, R., Cuevas, A., and González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis* **17** 153–176.

[6] Couallier, V. (1999). Estimation non paramétrique d'une discontinuité dans une densité. *C.R. Acad. Sci. Paris* **329** 633–636.

[7] Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics* **25** 1691–1708.

[8] Desmet, L. (2009). Local linear estimation of irregular curves with applications. *Doctoral Dissertation*, Statistics Section, Department of Mathematics, Katholieke Universiteit Leuven, Belgium.

[9] Desmet, L. and Gijbels, I. (2009). Local linear fitting and improved estimation near peaks. *The Canadian Journal of Statistics* **37** 453–475.

[10] Desmet, L. and Gijbels, I. (2009). Curve fitting under jump and peak irregularities using local linear regression. *Communications in Statistics–Theory and Methods*, to appear.

[11] Gijbels, I. (2008). Smoothing and preservation of irregularities using local linear fitting. *Applications of Mathematics* **53** 177–194.

[12] Gijbels, I. and Goderniaux, A.-C. (2004). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics* **46** 76–86.

[13] Gijbels, I., Lambert, A., and Qiu, P. (2006). Edge-preserving image denoising and estimation of discontinuous surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** 1075–1087.

[14] Gijbels, I., Lambert, A., and Qiu, P. (2007). Jump-preserving regression and smoothing using local linear fitting: a compromise. *The Annals of the In-*

stitute of Statistical Mathematics **59** 235–272.

[15] Herrick, D. R. M., Nason, G. P., and Silverman, B. W. (2001). Some new methods for wavelet density estimation. *Sankhyā Series A* **63** 394–411.

[16] Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* **6** 35–54.

[17] Jones, M. C. and Foster, P. J. (1993). Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics* **3** 81–94.

[18] Park, B. U., Jeong, S.-O., Jones, M. C., and Kang, K. H. (2003). Adaptive variable location kernel density estimators with good-performance at boundaries. *Journal of Nonparametric Statistics* **15** 61–75.

[19] Lambert, A. (2005). Nonparametric estimations of discontinuous curves and surfaces. *Doctoral dissertation*, Institut de Statistique, Université catholique de Louvain, Louvain-La-Neuve, Belgium.

[20] Leibscher, E. (1990). Kernel estimators for probability densities with discontinuities. *Statistics* **21** 185–196.

[21] Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, Series B* **56** 653–671.

[22] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* **20** 712–736.

[23] Mielniczuk, J., Sarda, P. and Vieu, P. (1989). Local data-driven bandwidth choice for density estimation. *Journal of Statistical Planning and Inference*, **23**, 53–69.

[24] Schucany, W. R. (1989). Locally optimal window widths for kernel density estimation with large samples. *Statistics & Probability Letters* **7** 401-405.

[25] Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics–Theory and Methods* **14** 1123–1136.

[26] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* **53** 683–690.

[27] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

[28] Wu, J. S. and Chu, C. K. (1993). Kernel type estimators of jump points and values of regression function. *The Annals of Statistics* **21** 1545–1566.