# Simple and Effective Boundary Correction for Kernel Densities and Regression with an Application to the World Income and Engel Curve Estimation

Dai, J; Sperlich, Stefan Andréas

# Simple and Effective Boundary Correction for Kernel Densities and Regression with an Application to the World Income and Engel Curve Estimation

**Jing Dai** and **Stefan Sperlich**[*]

**Abstract**

In both nonparametric density estimation and regression, the so-called boundary effects, i.e. the bias and variance increase due to one sided data information, can be quite serious. For estimation performed on transformed variables this problem can easily get boosted and may distort substantially the final estimates, and consequently the conclusions. After a brief review of some existing methods a new, straightforward and most simple boundary correction is proposed, applying local bandwidth variation at the boundaries. The statistical behavior is discussed and the performance for density and regression estimation is studied for small and moderate sample sizes. In a simulation study this method is shown to perform very well. This method exhibits to be excellent for estimating the world income distribution, and Engel curves in economics.

**Keywords:** boundary correction, kernel density estimation, kernel regression, local bandwidth.

[*]Both authors: Georg-August Universität Göttingen, Institut für Statistik und Ökonometrie, Platz der Göttinger Sieben 5, D - 37073 Göttingen, Germany. Corresponding author: stefan.sperlich@wiwi.uni-goettingen.de, tel: +49 551 397350. The authors gratefully acknowledge financial support from the DFG FOR916, and the Spanish MTM2008-03010. We thank three anonymous referees for helpful comments which improved this paper significantly. We further appreciated helpful discussion with Walter Zucchini.

# 1 Introduction

Boundary effects are a well known problem in nonparametric estimation, no matter if we think of density estimation or regression. Moreover, if the estimation has been performed on transformed covariates as recommended in the literature, see Wand et al. (1991), Ruppert and Cline (1994), Yang and Marron (1999), this problem may become boosted in two ways. Following these articles, a most appropriate transformation is the assignment $x_i \to \int_{-\infty}^{x_i} p(x)dx$ with $p$ being a parametric prior (maybe with estimated parameters) of the density of $X$. First, after such a transformation we definitely face boundaries (here 0 and 1) with especially heavy tails. Second, what is just a boundary effect for the transformed data may then affect big and essential parts of the untransformed model. But also when we estimate an untransformed model directly, "boundaries" are not necessarily small nor they are mostly of minor interest. The larger the noise to sample size ratio or the smoother the function, the larger is the bandwidth and thus the affected boundary region. Furthermore, often right the boundaries are of special interest; for example in poverty analysis, it is necessary to have reliable estimates of the income distribution at the left side "close" to the natural boundary 0. Similarly, when using nonparametric regression in econometrics, spill over effects, flexible returns to scale or multiple (dynamic) equilibria can typically, if at all, only be detected at or close to the boundaries. To conclude, if we are interested in risk, in poverty and inequality, look at the performance of especially young or old people, highly or lowly educated, compare large with small companies, etc. we always focus (also) at boundaries. In this article we will be confronted with boundary problems when studying the world income distribution, and when estimating the Engel curve for food expenditures in a poor country (Indonesia in our case).

As can be seen from these examples, we are concerned about boundary correction method for both kernel density and kernel regression estimation. A quick internet search reveals that seemingly many boundary correction methods exist already, many are referred to the linear correction for density estimation, see Jones (1993), and can be considered as modifications of this method. A quite comprehensive discussion of boundary correction methods for density estimation is given in Cheng et al. (1997). In general, the existing methods can be divided in following groups:

The method of modifying the kernel. The majority goes for this option, including Gasser et al. (1985), Jones (1993) and the local polynomial approaches (Cheng et al. 1997). Referring to the argument that local polynomial estimation would automatically correct for boundary effects in regression (see for example Fan and Gijbels, 1992) they apply this idea in density estimation. Effectively, however, a boundary correction takes only place if the polynomial is of the "correct" order; else it can even aggravate the boundary effect. In density estimation the use of local polynomial fitting has not prevailed although Zhang and Karunamuni (1998, 2000) extended this method to the case of density estimation in combination with a bandwidth-variation function. Nevertheless, in many situations local polynomials are certainly an attractive remedy for boundary effects in regression though the optimal weighting introduced by Cheng et al. (1997) has not been applied (much) until now.

The second set of boundary correction methods modifies the bandwidth toward the boundaries. This group is much smaller and less known. Among them, Rice (1984), Gasser et al. (1985) and Müller (1991), see also Hall and Wehrly (1991), are maybe the most practical ones. They consider the regression context and suggest to fix the window size inside the support of the covariates. Somehow similar to this idea, the loess and lowess smoother of Cleveland (1979, 1981) implemented in R and S, uses a fixed span addressing automatically the boundary effects, see also Cleveland et al. (1992).

A quite old idea is the reflection method, introduced by Schuster (1985) and Silverman (1986), and later on extended by Cline and Hart (1991). A further development of it are the more recent methods of creating pseudo data to correct for edges, see Cowling and Hall (1996). This method is more adaptive than the common data reflection approach in the sense that it corrects also for discontinuities in derivatives of the density. Zhang et al. (1999) suggested a method of generating pseudo data combining the transformation and reflection methods. In some sense one could also add here the idea of Hall and Park (2002). They proposed an empirical translation of the argument of the kernels and a bootstrap method to translate the boundary estimate towards the body of the data set.

Finally we should mention again the transformation methods, see for example Wand et al. (1991), Ruppert and Marron (1994), and Yang (2000).

It is surprising that in spite of their importance in practice and the considerable (though not enormous) amount of theoretical studies, boundary correction methods are hardly used neither in density estimation nor in regression. One obvious reason is the lack of implementations in statistical and econometric software, another could be a disappointingly small performance improvement when using them. Finally, practitioners are often not willing to apply complex, sometimes seemingly non-intuitive methods.

For this reason we will concentrate mainly on comparing ours with Jones (1993) but also compare with fixed window size, the pseudo data approach (in particular Cowling and Hall, 1996) for densities, local linear for regression, and data transformation (in an application). However, to the best of our knowledge, even the quite well known and also reasonably well working method of Jones is neither much used nor implemented in standard software packages. Beside the lack of software, another reason for the scarce usage could be its complexity compared to the visible improvement in the final estimate. As one will see, our method is much less complex and requires hardly more computational effort than the estimation without boundary correction does.

Summarizing, we are looking for a quick and easy boundary correction method that can at least compete with Jones (1993) and local polynomials in both, density and regression problems. As usually, our method is driven by the idea of substantial bias reduction, c.f. Hall and Park (2002). Although the simpleness of our method allows for a (serious) variance increase, in sum the boundary estimates improve in mean squared error. We are not asking which method handles uniformly best the probability mass at or near the boundaries. We introduce a new simple and practical method, give asymptotic insight, a comprehensive simulation study, a comparison with existing methods, and 2 applications.

## 2 Kernel Estimators and Boundary Correction

Suppose we want to estimate a probability density $f$ nonparametrically based on a random sample $\{X_1, X_2, \ldots, X_n\}$, $X_i \in R^d$. For the ease of presentation we restrict in the following notation to univariate models ($d = 1$) in both density estimation and regression. The extensions to multivariate density and regression estimation is straight forward. Then the standard kernel density estimator of $f(x)$ is given by

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x), \tag{1}$$

where $K_h(\bullet) = \frac{1}{h} K(\bullet/h)$ could be any common symmetric kernel with support $[-1, 1]$, satisfying $\mu_0(K) = 1$, $\mu_1(K) = 0$, $\mu_2(K) < \infty$, with $\mu_l(K) = \int_{-1}^{1} u^l K(u) du$ ($l = 0, 1, 2$; $u = \frac{X-x}{h}$) and $h$ denoting the bandwidth. For such a kernel method to make sense, $f$ is supposed to be smooth, typically expressed in the assumption of an existing second derivative $f''$.

However, if the support of $f$ is bounded and has no exponentially falling tails, this estimator is well known to suffer from the so-called "boundary effects". This means, for all points $x$ being closer to the boundary than $h$, (1) underestimates (strongly) $f(x)$ since the kernel searches erroneously for information outside the support of $f$.

Now imagine we consider a random sample $\{(Y_i, X_i)\}_{i=1}^n$ for the regression model

$$Y_i = m(X_i) + \epsilon_i, \tag{2}$$

where $\epsilon_i$ are random errors with expectation zero and finite variance $\sigma_i^2$, and a smooth regression function $m(\bullet)$ that is supposed to have second derivatives. Then, the local polynomial estimator of degree $\alpha$ can be expressed as

$$\widehat{m}^{(v)}(x) = (v!) e_v^T (Z^T W Z)^{-1} Z^T W Y \ , \tag{3}$$

where $m^{(v)}$ denotes the $v \leq \alpha$ derivative of $m$, $Z$ is a $(n \times (\alpha + 1))$ matrix with elements $Z_{ik} = (X_i - x)^{k-1}$, $Y = (Y_1, \ldots, Y_n)$, $W = \text{diag}\{K_h(X_i - x)\}_{i=1}^n$, and $e_v$ is a vector of zeros with a 1 at position $(v + 1)$. For $v = 0$ and $\alpha = 0$ we get the popular and simple Nadaraya-Watson estimator (Nadaraya, 1964). Also in this regression case, the problem of boundary effects is well known and can become quite serious in practice.

To avoid confusion we shall assume (at least in the notation) that global bandwidths $h_{global}$ were used if not stated differently, especially for the estimation at all interior points. Henceforth, the lower boundary - if exists - is called $a$, and the upper boundary - if exists - is denoted by $c$. In other words, the interior region is $[a + h_{global}, c - h_{global}]$ while $B_l = \{x : a \leq x < (a + h_{global})\}$ and $B_r = \{x : (c - h_{global}) < x \leq c\}$ are the left and right boundary regions.

Many methods have been proposed to correct for boundary effects, see Section 1. The probably most popular one is the method of Gasser and Müller (1979), revitalized by and named after Jones (1993), the local linear estimation. Jones (1993) proposed to borrow more strength from inside of the support. More specific, if $f$ is supported on $[a, c]$, then the used kernel is given by

$$K^*(u) = \frac{w_3 - w_2 u}{w_1 w_3 - (w_2)^2} K(u) \mathbb{1}_{[c_2, c_1]}, \tag{4}$$

where the re-normalizing moments $w_j$ are defined by

$$w_j = \int_{c_2}^{c_1} \left( \frac{t - x}{h_{global}} \right)^{j-1} K\left( \frac{t - x}{h_{global}} \right) dt,$$

with $c_1 = min(c, x + h_{global})$ and $c_2 = max(a, x - h_{global})$. Then the density estimate applying his linear boundary corrector is $\hat{f}$ in (1) but with the linearly corrected kernel $K^*(u)$. Similarly, for the regression estimator (3) we would use $K^*(u)$ in the definition of $W$.

An alternative is to choose local bandwidths in the boundary area. Typically, one would say it is obvious that larger bandwidths should be used there. Rice (1984) and Gasser et al. (1985) suggested choosing a bandwidth that keeps the window width fixed at the boundary, see also Hall and Wehrly (1991). To reach this we simply use for all boundary points a local bandwidth defined by

$$h_x = \begin{cases} 2h_{global} - (x-a) & \text{for } a < x < (a + h_{global}), \\ 2h_{global} - (c-x) & \text{for } (c - h_{global}) < x < c, \\ h_{global} & \text{otherwise.} \end{cases} \tag{5}$$

Hall and Wehrly (1991) extended this idea to first generate pseudo-data (with a kind of extrapolating bootstrap) and then estimate in the boundary region using the set of real and pseudo data. In the context of estimating a regression function $m(\bullet)$, Rice (1984) used a kind of Richardson extrapolation proposing a linear combination of not corrected estimators $\hat{m}_{h_{global}}$ and corrected estimators $\hat{m}_{h_x}$. I.e. for all boundary points $x = a + h\rho$, $\rho < 1$ he set

$$\tilde{m}(x) = (1 + \beta_\rho)\hat{m}_{h_{global}}(x) - \beta_\rho \hat{m}_{h_x}(x), \tag{6}$$

with $\hat{m}$ as in (3) with $\alpha = 0$, $h_x$ as in (5), and

$$\beta_\rho = \frac{w_1(\rho) w_0^{-1}(\rho)}{(2-\rho) w_1\left(\frac{\rho}{2-\rho}\right) w_0^{-1}\left(\frac{\rho}{2-\rho}\right) - w_1 w_0^{-1}} \qquad \text{for } w_k(v) = \int_{-1}^{v} u^k K(u) du.$$

In contrast to the idea of enlarging the bandwidth at the boundary, we suggest to reduce the bandwidth in the boundary regions. Our local bandwidth $h_x$ for $a \leq x \leq c$

can be indicated by

$$h_x = \begin{cases} max(x-a,\varepsilon) & \text{if } a \le x < (h_{global}+a), \\ max(c-x,\varepsilon) & \text{if } (c-h_{global}) < x \le c, \\ h_{global} & \text{otherwise.} \end{cases} \tag{7}$$

where $\varepsilon > 0$ is just added for numerical reasons going to zero for $n \to \infty$. For theoretical discussion one could even skip $\varepsilon$ and define $h_x$ only for $a < x < c$ such that the density or regression estimator is not defined at the boundaries but arbitrarily close to them.

Inserting $h_x$, no matter whether (5) or (7) into (1), we have

$$\hat{f}_{h_x}(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_x}\right) \tag{8}$$

for the kernel density estimator. As we can see, the local bandwidths $h_x$ are adjusted within the boundary region while $\hat{f}_{h_x}(x)$ is identical to the usual kernel density estimator (1) if $x$ is in the interior region. This also corresponds to Jones' method. It should be emphasized that the index $x$ of $h_x$ refers to a given point at which we wish to estimate the density or regression function. When we insert $h_x$ into the regression estimator (3), we adjust only weight $W$. However, Jones' method is identical to (3) with global bandwidth inside the interior region but using $K^*$ in $W$.

We concentrate here on the situation where the boundary is (naturally) given, see also our applications in Section 4. For given boundaries and $x$ the bandwidths $h_x$ are neither in the interior nor at the boundary random. Therefore the statistical behavior of our resulting estimators is as simple as the method is. One might also imagine situations where the boundary is unknown and has to be estimated. Sometimes in the literature, people set the boundaries equal to the smallest and largest observation. Especially for density estimation, however, this is a quite questionable procedure to estimate the boundaries. In those cases the statistical behavior of our final estimate (density and regression) complicates a lot for the reason of having then a random bandwidth. One would first have to establish assumptions and conditions on the boundary estimates etc. For simpler situations there can be found some discussions about random bandwidths e.g. in Abramson (1982), Hall (1983) or Hall and Marron (1988).

Recall that in our notation, a point $x$ belongs to the boundary region when its distance to the boundary is smaller than $h_{global}$. In asymptotic theory a boundary point is a point $x$ being closer to the boundary than the bandwidth used to estimate $f(x)$ or $m(x)$ respectively. In this sense, our method turns all support points into interior points and the asymptotics therefore remain unchanged. This was also the original idea of the reflection and the pseudo data approach; they changed therefore artificially the support, we change the bandwidth. Then, for the kernel density estimator (1) one obtains

$$Bias\{\hat{f}_{h_x}(x)\} = \frac{h_x^2}{2} f''(x)\mu_2(K) + o_p(h_x^2), \tag{9}$$

with $\mu_2(K) = \int_{-1}^{1} u^2 K(u)du$, and

$$Var\{\hat{f}_{h_x}(x)\} = \frac{1}{nh_x} f(x)\|K\|_2^2 + o_p(\frac{1}{nh_x}), \tag{10}$$

with $\|K\|_2^2 = \int K^2(u)du$. For the regression (3) one obtains

$$Bias\{\hat{m}_{h_x}(x)\} = \frac{h_x^2}{2}\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\}\mu_2(K) + o_p(h_x^2) \tag{11}$$

for the Nadaraya-Watson estimator with $\alpha = 0$, and

$$Bias\{\hat{m}_{h_x}(x)\} = \frac{h_x^2}{2} m''(x)\mu_2(K) + o_p(h_x^2) \tag{12}$$

for the local linear estimator with $\alpha = 1$, both with

$$Var\{\hat{m}_{h_x}(x)\} = \frac{1}{nh_x}\frac{\sigma^2(x)}{f(x)}\|K\|_2^2 + o_p(\frac{1}{nh_x}). \tag{13}$$

For consistency one needs $h_x \to 0$ and $nh_x \to \infty$ for $n \to \infty$. It is clear that our proposal of $h_x$, given in (7), gives full preference to the bias reduction at the cost of the variance. This becomes evident when we compare it with the methods of Jones (1993) and fixed window sizes. Nevertheless, in sum this can easily yield a mean squared error decrease, cf. with our simulations in the next section. The pseudo data approach is constructed to control for both bias and variance at the edges.

Let us consider the asymptotics of a kernel density estimator when the method of Jones (1993) is applied. Without loss of generality imagine we have a lower bound $a$. Recall that we consider kernels bounded on $[-1, 1]$. We skip the index *global* of bandwidth $h$ and define implicitly a scalar $p$ depending on $x$ and $a$ via $x = p(a + h)$. Then, for $a_l(p) = \int_{-1}^{\min\{1,p\}} u^l K(u)du$ and $b(p) = \int_{-1}^{\min\{1,p\}} K^2(u)du$ the asymptotics can be approximated by

$$Bias\{\hat{f}_h(x)\} \simeq f(x)(a_0(p) - 1) - ha_1(p)f'(x) + \frac{h^2}{2} f''(x)a_2(p), \tag{14}$$

with

$$Var\{\hat{f}_h(x)\} \simeq \frac{1}{nh} f(x)b(p) \ . \tag{15}$$

Note that for all interior points, the asymptotics coincide with the common expressions (9) and (10) respectively. In order to achieve a bias of order $h^2$ near the boundary as well as in the interior, Jones (1993) defined a linear combination of $K$ and a closely related function to obtain boundary kernel (4) such that $a_0(p) = \int_{-1}^{\min\{1,p\}} K^*(u)du = 1$ and $a_1(p) = \int_{-1}^{\min\{1,p\}} uK^*(u)du = 0$. Similar observations can be made for regression and the other boundary correcting methods.

This however, are asymptotic statements. In the next section we will study how these methods compare for finite samples of different sizes. We should emphasize once again that in the past it has been stressed a lot that local polynomial estimators do automatically correct for boundary effects. We mentioned already in Section 1 that this is only true if the order of polynomials is chosen accordingly. We should further remark that local polynomial estimators (in practice and theory) need larger bandwidths for increasing degrees. In boundary regions where data are sparse, it can even be recommendable to choose degree $\leq 1$, i.e. to use the Nadaraya-Watson or local linear estimator. Applying Jones' or our method for local linear smoothers yielded pretty bad numerical performance and is therefore skipped in the simulation section. The proposal of Cheng et al. (1997) extending the local polynomial estimator by an additional weighting, turns out to be rather complex in practice and still needs a reasonable amount of data.

We will also compare these simple methods with the reflection or pseudo data approach of Cowling and Hall (1996). Note, however, that this is by no means an easy to use, intuitive method. In fact, the practitioner has to chose two further parameters which are essential for the success of the method. Cowling and Hall (1996) defined the density estimator at the boundaries as

$$\hat{f}(x) = \frac{1}{nh}\left\{ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^{m} K\left(\frac{x - X_{-i}}{h}\right)\right\}, \tag{16}$$

where $m$ is such that $O(nh) < O(m) < O(n)$, and $X_{-i}$ are pseudo data. More specific, for positive constants $A_1, \ldots, A_s$, $s \geq r$, where $r$ is related to the smoothness of the quantile function of $X$ at the considered edge, and real numbers $a_1, \ldots, a_s$ they define

$$X_{-i} = \sum_{j=1}^{s} a_j X_{A_j i}, \quad 1 \leq i \leq \frac{n}{\max\{A_i\}}, \quad \text{s. th. } \sum_{k=1}^{s} a_k A_k^j = (-1)^j, \quad 1 \leq j \leq r. \tag{17}$$

For example, in their article they recommend the so called best three-point rule $X_{-i} = -5X_{(i/3)} - 4X_{(2i/3)} + 10/3X_{(i)}$, $i = 1, 2, \ldots, n$ with $X_{(i)}$ indicating order statistics. Unfortunately, in Cowling and Hall (1996) there is nothing said about the choice of $m$ neither in general nor in their simulations. For more details and asymptotic behavior we refer to the paper of Cowling and Hall (1996).

Finally we would like to mention that there exist certainly many other methods for nonparametric regression estimation like different versions of splines, Fourier series, wavelets, etc. All these suffer a different kind of boundary effect. Fortunately, for our approach it is clear how it can be applied / extended to these other methods.

# 3 Finite Sample Comparison

We separate the simulation study into two parts: a more detailed one for density estimation, and a smaller study for regression. The reason is that in regression, the boundary performance depends on too many factors to provide a really comprehensive study; in fact, it depends on the distribution of the covariate(s), the functional form of the conditional mean of the response, on the degree of the (local) polynomial, and even on the heteroscedasticity. Therefore, the regression part of our simulation study has rather an illustrative character. In our simulations we set $\varepsilon = 0.001$ in (7).

## 3.1 Density Estimation

To assess the effect of the correction methods near the boundaries, the following six models are investigated:

1. uniform distribution on $[0, 1]$;
2. gamma distribution $Gamma(2.25, 1.5)$ applied on $5x$;
3. log-normal distribution with $\mu = 0$ and $\sigma = 1$;
4. log-normal distribution with $\mu = 0$ and $\sigma = 1.5$;
5. log-normal distribution with $\mu = 0$ and $\sigma = 2$;
6. exponential distribution with $\lambda = \mu = 5$.

The density estimator was defined as in (1) with the Epanechnikov kernel $K(u) = 3/4(1 - u^2)\mathbb{1}\{|u| < 1\}$. For illustration issues we chose $h_{global} = 0.3$ provoking thereby serious boundary effects. We estimated $f(\bullet)$ on a grid of 25 equidistant points $x_1 < x_2 < \ldots < x_{25}$, where $x_1 = 0$ and $x_{25} = 1$. Then, the first 8 points lie in the left boundary region. The simulated sample sizes were $n = 50$, $n = 100$ (not shown for brevity) and $n = 200$. All results were calculated from 1000 simulation runs.

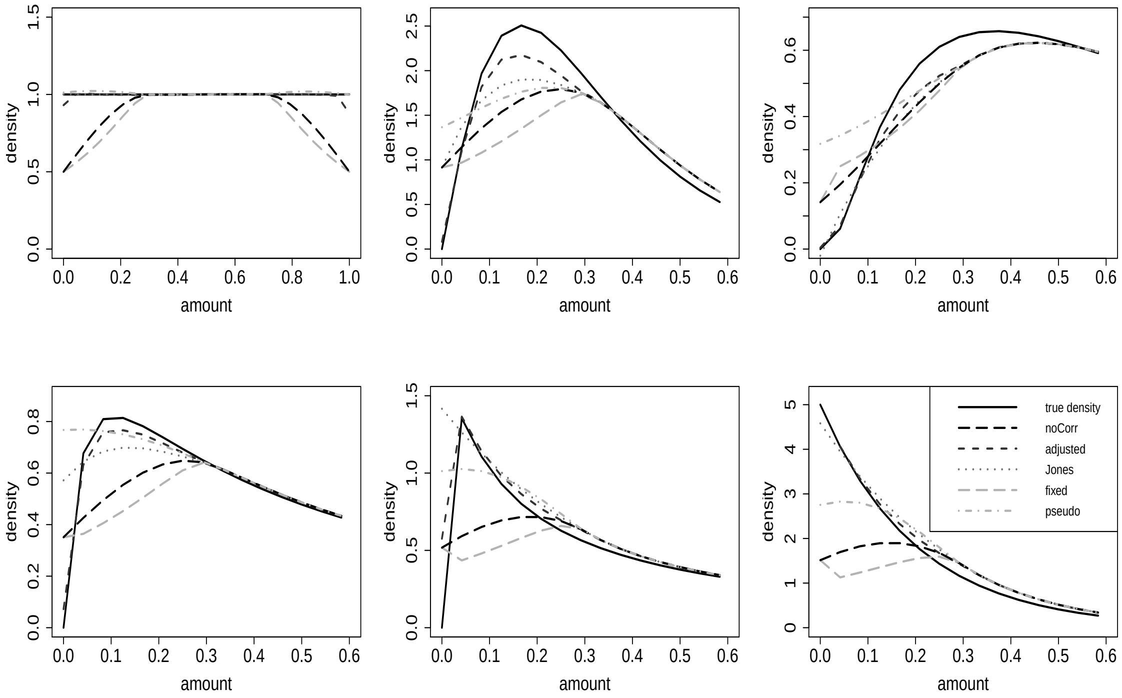

Figure 1: The estimates for the six densities (upper left to the lower right) for n=50. Black line is the true density, black long dashes indicate the density estimate without boundary correction, grey long dashed is the method with fixed window size (5), black short dashed is our adjusted window method (7), grey dashed & dotted is the pseudo data method (16), and grey dotted line is Jones' estimate.

In Figures 1 and 2 are given the true density and the expectation of its kernel estimates, i.e. the averages over 1000 simulation runs. To highlight the behavior in the boundary region, we plotted the estimates in $[0, 0.6]$ for designs 2 to 5, and in $[0, 1]$ for design 1. Maybe not surprisingly, see discussion in Section 2, our new method has the smallest bias and reflects best the true boundary behavior of the underlying densities. For both, moderate sample size ($n = 50$) and relatively large samples ($n = 200$) our method outperforms the others, while Jones' method seems to be uniformly the second best. It should be remarked that Jones' estimator shows exactly the behavior indicated in

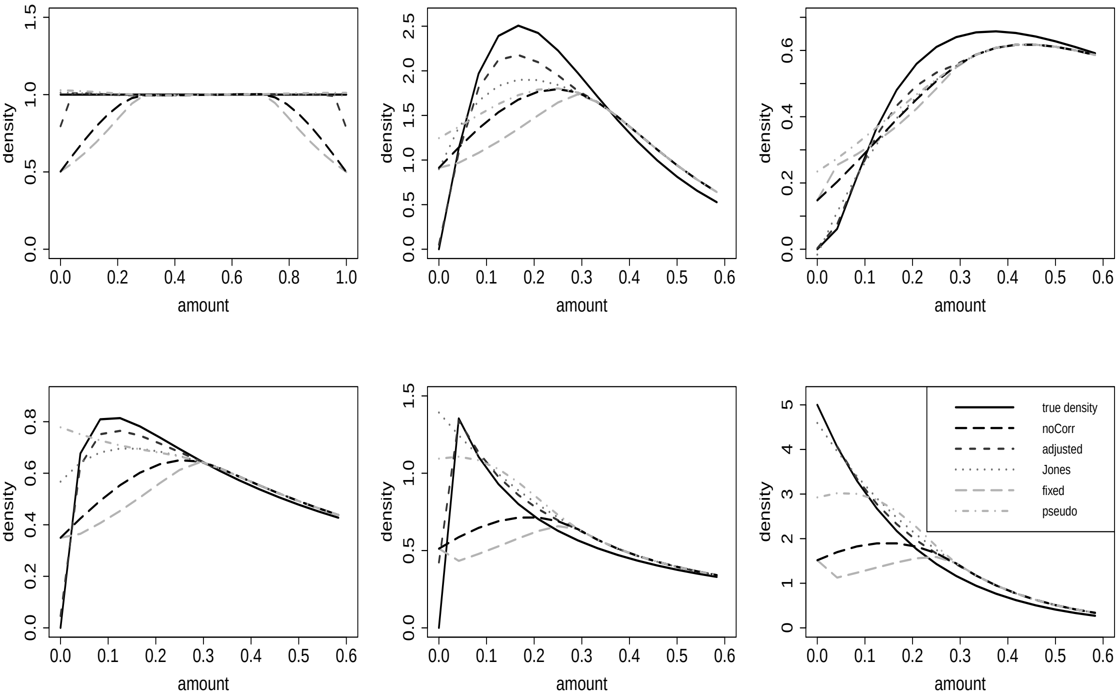

Figure 2: The estimates for the six densities (upper left to the lower right) for n=200. Black line is the true density, black long dashes indicate the density estimate without boundary correction, grey long dashed is the method with fixed window size (5), black short dashed is our adjusted window method (7), grey dashed & dotted is the pseudo data method (16), and grey dotted line is Jones' estimate.

(14), it strongly underestimates the curvature e.g. for design 2 and 4. The method with fixed window size is even worse than not correcting at all. As indicated, for the density estimation at the boundary we tried also the method of Cowling and Hall (1996) with the best three-point rule and the maximal possible $m$ resulting out of it. This maximal number seems to be $n-1$, but it turned out that the performance improves (except for density 6) when we ignore all pseudo data $X_{-i}$ lying in the support of X, cf. p. 555 of Cowling and Hall (1996). We also tried other choices like $m = n^{9/10}$ but got worse results. Apart from the choice of pseudo generator and $m$ it is computationally easy but its performance can only compete with Jones' or ours when the original data are uniformly distributed.

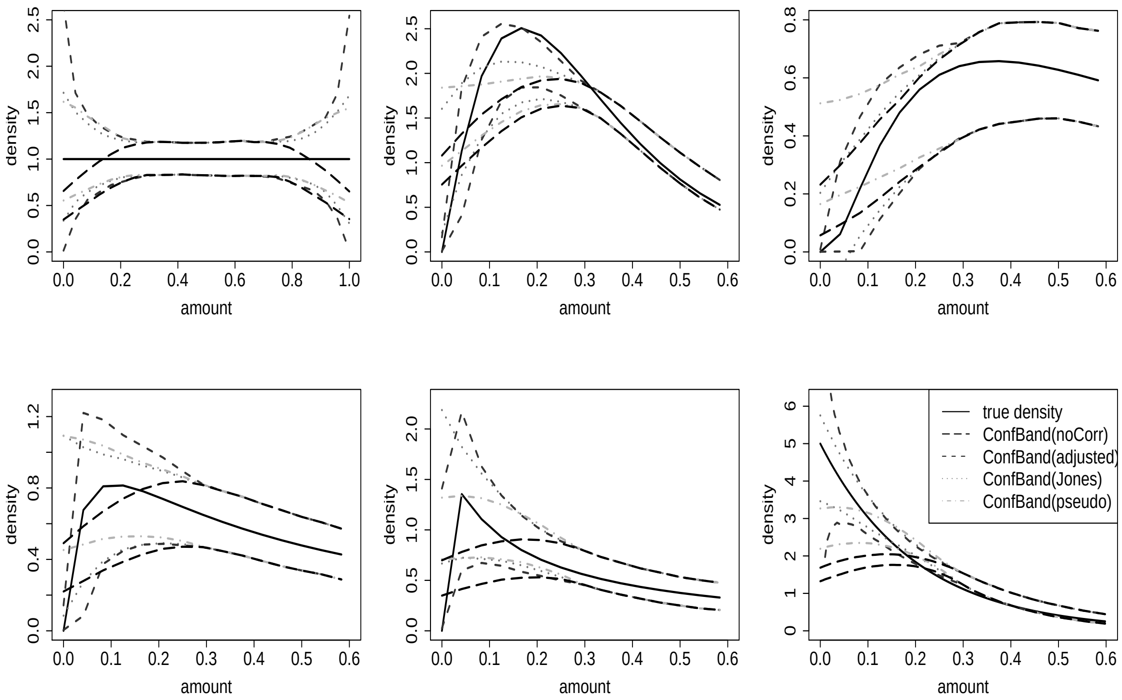

Figure 3: The simulated confidence bands corresponding to Figure 1 with coverage probability of 80%.

Clearly, as said in Section 2, our method is tailored to reduce bias but may have awful large variance. If so, it can not really be considered as an improvement since the outcome would be rather random. To check this we constructed - again out of our 1000 simulation runs - point wise confidence bands with a coverage probability of 80%. These bands are given in Figures 3 and 4. First, we have to admit that at the boundaries our method has often the widest intervals. A closer look, however, reveals that they are not much wider and sometimes even tighter than the bands corresponding to Jones' method; and they are the only confidence bands including always the true function, except for design 2. For $n = 200$ the widths of all the confidence bands is almost the same for our and Jones' method.

To better quantify the gain in bias and mean squared error, we calculated the absolute bias and mean squared error averaged over the grid of 8 equidistant points $x_l$ over the

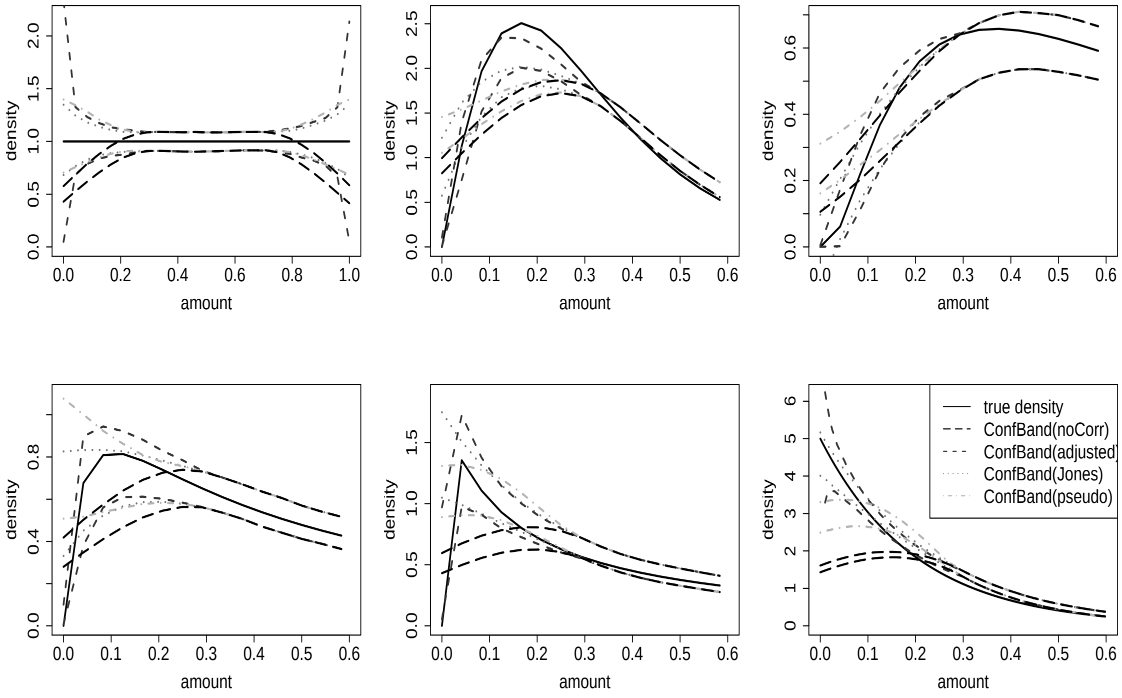

Figure 4: The confidence bands corresponding to Figure 2 with coverage probability of 80%.

left boundary region, i.e. we calculated

$$|Bias\{\hat{f}_h(x)\}| \;=\; \frac{1}{8}\sum_{l=1}^{8}\left|\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{f}_h^M(x_l) - f(x_l)\right)\right|, \tag{18}$$

$$\text{and MSE}\{\hat{f}_h(x)\} \;=\; \frac{1}{8}\sum_{l=1}^{8}\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{f}_h^M(x_l) - f(x_l)\right)^2 . \tag{19}$$

The results are displayed in Table 1. It can be seen from this table that, as expected, our method beats by far the competitors when looking at the bias. For the variance this is different, at least for small sample sizes (except for the $U[0,1]$ design). For $n = 100$ (not shown) the mean squared error is about the same for our and Jones' method; for $n = 200$ our new method outperforms all (except for the $U[0,1]$ and $Log - N(0,2)$ design when comparing with Jones).

Before coming to the regression we should briefly summarize. We have looked for a rather simple, easy to implement and to interpret method for mitigating the boundary

| $n$ | | | $M1$ | $M2$ | $M3$ | $M4$ | $M5$ | $M6$ |
|---|---|---|---|---|---|---|---|---|
| 50 | $\|Bias\|$ | *no correction* | .2022 | .5673 | .0975 | .1895 | .2405 | .7830 |
| | | *Jones* | .0013 | .4696 | .0678 | .1280 | .0765 | .2370 |
| | | *adjusted* | .0105 | .2093 | .0474 | .0365 | .0518 | .1436 |
| | | *fixed* | .2577 | .7572 | .1158 | .2477 | .3352 | 1.096 |
| | | *pseudo* | .0147 | .6003 | .1373 | .1345 | .1266 | .4355 |
| | $MSE$ | *no correction* | .0596 | .3384 | .0208 | .0527 | .0776 | .6261 |
| | | *Jones* | .0835 | .3053 | .0227 | .0693 | .0681 | .1421 |
| | | *adjusted* | .6816 | .1446 | .0236 | .0718 | .1037 | .2082 |
| | | *fixed* | .0776 | .5811 | .0215 | .0723 | .1239 | 1.206 |
| | | *pseudo* | .0708 | .4053 | .0354 | .0522 | .0516 | .2512 |
| 200 | $\|Bias\|$ | *no correction* | .2014 | .5665 | .0951 | .1888 | .2428 | .7811 |
| | | *Jones* | .0053 | .4668 | .0618 | .1278 | .0745 | .2360 |
| | | *adjusted* | .0295 | .2080 | .0392 | .0341 | .0500 | .1440 |
| | | *fixed* | .2575 | .7563 | .1142 | .2460 | .3365 | 1.096 |
| | | *pseudo* | .0142 | .5962 | .1109 | .1528 | .1168 | .4498 |
| | $MSE$ | *no correction* | .0450 | .3251 | .0120 | .0399 | .0639 | .6136 |
| | | *Jones* | .0195 | .2383 | .0085 | .0292 | .0203 | .0780 |
| | | *adjusted* | .1484 | .0670 | .0071 | .0158 | .0259 | .0678 |
| | | *fixed* | .0691 | .5739 | .0151 | .0633 | .1161 | 1.202 |
| | | *pseudo* | .0281 | .3647 | .0162 | .0405 | .0271 | .2371 |

Table 1: Absolute bias and MSE of density estimates in left boundary region for sample size $n = 50$ and $n = 200$, based on 1000 repetitions: *adjusted* refers to our method (7); *fixed* refers to a fixed window size (5); *pseudo* refers to (16).

effects which in practice can cause rather serious problems and nuisance. As has been shown in Section 2, equations (7), our method complies with these requirements. Among all methods we have seen it is even the one with the simplest implementation. The ease of interpretation comes along with the insight that the statistical behavior is the same as for the interior points; it is a local bandwidth which - this we admit - can become rather small numerically although not in its rate. Fortunately, it has turned out in our simulation study that this method is not just the simplest one but also shows an excellent performance. In fact, it outperforms even the popular method of Jones. The further discussed alternatives seem not to work in our density examples.

## 3.2 Regression Estimation

We recommend our new method not only for density estimation but also for kernel regression. As mentioned above, due to the fact that the boundary effects depend on too many factors, we have limited the following study to a brief illustrative simulation with only one design for the one dimensional covariate $X$, and a simple cubic polynomial for the regression function. That is, we consider random samples $\{(Y_i, X_i)\}_{i=1}^n$ from the nonlinear model

$$Y_i = m(X_i) + \epsilon_i \ , \text{ where } \ m(x) = -(10/3)x^3 + 5x^2 - 1.275x \tag{20}$$

is a smooth regression function, $X \sim U[0,1]$ i.i.d. and $\epsilon \sim N(0, 0.1)$ i.i.d. We estimated $m(\bullet)$ with the Nadaraya-Watson and the local linear estimator, i.e. (3) with $\alpha = 0$ or $\alpha = 1$ respectively. We used the Quartic kernel $K(u) = 15/16(1-u^2)^2 \mathbb{1}\{|u| < 1\}$ on a grid of 25 equidistant points $x_1 < x_2 < \ldots < x_{25}$, where $x_1 = 0$ and $x_{25} = 1$, as we did above. Then again, for a global bandwidth of $h_{global} = 0.3$ the first 8 points form an equidistant grid in the left boundary region. Note that the design choice favors now Jones' method, recall the results of Section 3.1. Like before, we did simulations for sample sizes $n = 50$ and $n = 200$.

As it was true for the density estimation context, a most serious problem is the bias at the boundary, and this is exactly what our method tries to mitigate. It can be seen from Figure 5, that the bias is corrected best by our method. Jones' method improves the Nadaraya-Watson but not the local linear estimator (not shown). It turned out that also our method can cause problems in combination with the local linear estimator (not shown), see our discussion about local polynomial estimation when data are sparse. Again, the method with fixed window size performs worst. We also tried Rice' (1984) more complex procedure, see (6), and found that it could not uniformly compete with simple local linear nor with ours. Additionally, it does not really count to the set of "simple and practical" methods. The local linear estimator turned out to be the

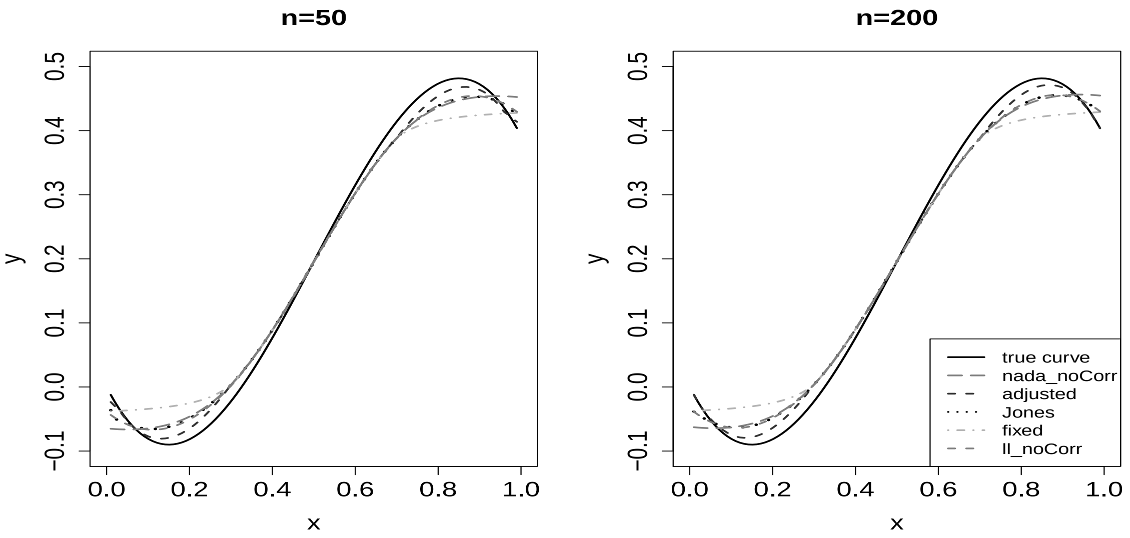

Figure 5: Comparison of regression estimates: black line is true curve, grey long dashed is Nadaraya-Watson estimate without boundary correction, black short dashed is our (adjusted) method, black dotted line is Jones' estimate, grey dashed and dotted is the estimate with fixed window size, and grey short dashed is the local linear estimator.

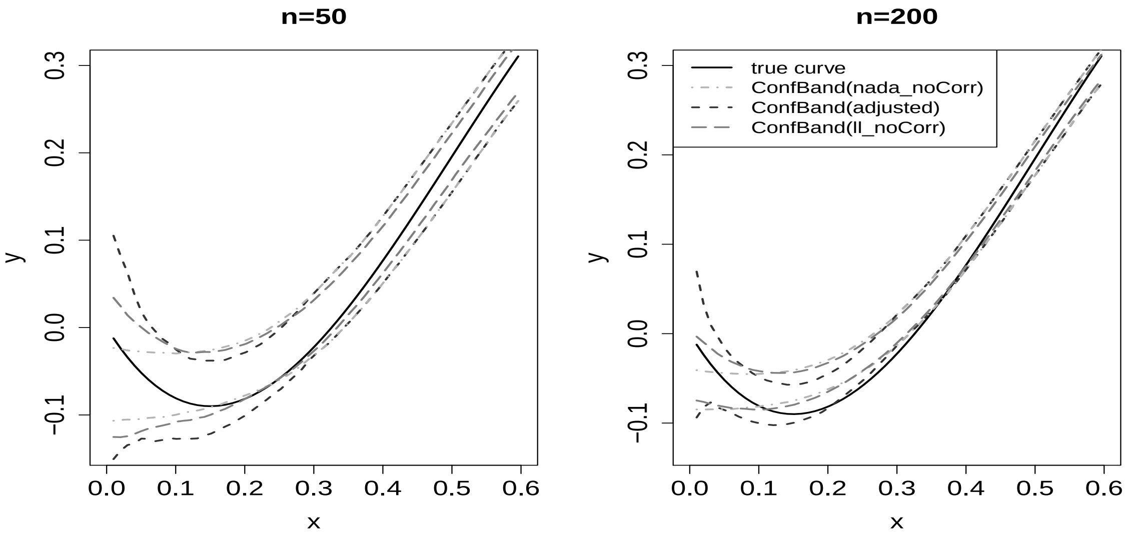

Figure 6: The confidence bands for the left boundary, corresponding to Figure 5 for the non corrected Nadaraya-Watson, the non corrected local linear, and our method.

strongest competitor compared to our method.

To have an idea about the variance of the estimators, we again constructed point wise confidence bands with an 80% coverage probability, see Figure 6. As for the density estimation, the bands for our corrector are getting wider at the boundaries than for the other methods. This time, the confidence bands are still much wider when increasing the sample size from $n = 50$ to $n = 200$. However, again it is only our method that really captures the curvature of the true data generating function such that the true function is almost always inside the 80% point wise confidence bands, especially in the boundary region.

Our simulations conclude with Table 2 showing the average absolute biases and mean squared errors of the left boundary region. As we did for density estimation, we calculated

$$\begin{aligned} |Bias\{\hat{m}_h(x)\}| &= \frac{1}{8}\sum_{l=1}^{8}\left|\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{m}_h^M(x_l) - m(x_l)\right)\right|, & (21) \\ \text{and } \ \mathrm{MSE}\{\hat{m}_h(x)\} &= \frac{1}{8}\sum_{l=1}^{8}\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{m}_h^M(x_l) - m(x_l)\right)^2. & (22) \end{aligned}$$

The results confirm what we have seen in Figures 5 and 6. Our method beats the others by far concerning bias reduction at the boundary. Due to its large variance, however, its mean squared errors (in average) are clearly larger than for all in the small sample $n = 50$ and is still larger than others with sample size $n = 200$.

| | $\|Bias\|$ | | $MSE$ | |
|---|---|---|---|---|
| | $n = 50$ | $n = 200$ | $n = 50$ | $n = 200$ |
| *adjusted* | .0146 | .0125 | .0028 | .0016 |
| *NW(no correction)* | .0317 | .0308 | .0018 | .0011 |
| *Jones* | .0272 | .0247 | .0408 | .0010 |
| *LL(no correction)* | .0259 | .0246 | .0022 | .0009 |
| *fixed* | .0447 | .0435 | .0027 | .0021 |

Table 2: Absolute bias and MSE of regression estimates in left boundary region for sample size 50 and 200 based on 1000 repetitions.

# 4 World Income Distribution and Engel Curve Estimation

The potential of our method and the need of boundary correction can easily be seen in the two following applications. First we estimate the world income distribution, and second we estimate the Engel regression curves for food expenditure in Indonesia.

The world income distribution is an ongoing concern for economists and scholars worldwide, see e.g. Acemoglu and Ventura (2002) and Sala-I-Martin (2006). The discussion of a two or even three mode shape (cf. Holzmann et al. 2007) of the world income distribution has been challenging the conventional findings of growth empirics. As a consequence, for example, the convergence literature established divergence among countries but found different convergence clubs. Further, from this world income distribution one can obtain measures for global inequality and poverty as well as global growth incidence curves.

An often discussed question is how many convergence clubs do we find world wide, what should be certainly reflected in the shape of the income density function. The typical problem here is the one of a proper modeling, for example should one use a normal mixture or a log-normal mixture, and how should we bound the number of components (from above) or the variances (from below). This problem even appears in nonparametrics: when Holzmann et al. (2007) used the income, they had serious problems at the left boundary; when they considered log-income, the 'convergence club' of the rich countries (i.e. a bump on the right) was no longer visible. This can be seen quite well in our application in Figure 7. It shows kernel density estimates based on all available worlds real PPP GDP per capita for the year 2003 from the Penn World Table, Version 6.2. The available and here used income data comprise 174 countries. In this analysis we estimate density $f(\bullet)$ with lower bound $a = 0$ on a grid of 200 equidistant points.

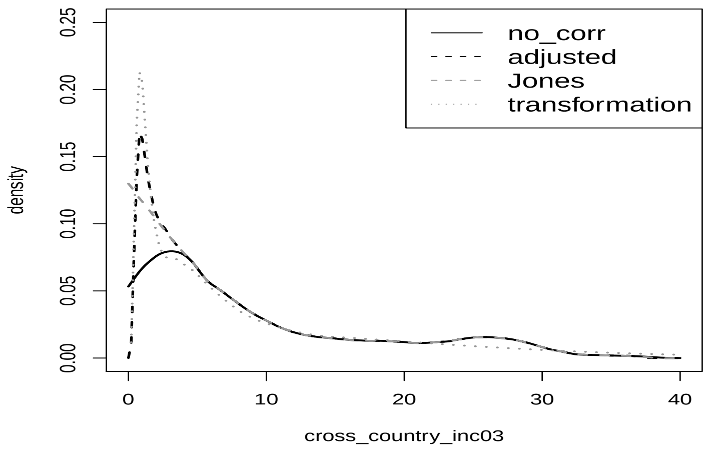

Figure 7: Comparison of kernel density estimates for cross-country income distribution in 2003 with $h_{global} = 5.37$ (1.5 times Silverman's rule-of-thumb bandwidth): black solid line is kernel density estimate without boundary correction, black dashed is our method ('adjusted'), grey dashed is Jones' estimate, and grey dotted line is kernel estimate on log-transformed data with $h_{global} = 0.76$ (1.5 times Silverman's rule-of-thumb bandwidth). Scale: x-axis $10^3$, y-axis: $10^{-3}$.

The black line is the usual kernel estimate without boundary correction. The comparison with all other methods shows that we face a serious boundary problem at 0. The global bandwidth has been chosen such that we could replicate the graphical results of Holzmann et al. (2007) where the bandwidth choice is not mentioned. The density estimation based on the log incomes and therefore facing no boundaries nicely resolves the very sharp peak at $income \approx 880$ (very poor countries) and also makes visible a second convergence club of developing countries showing a plateau and a flat slope (to the left) at around $income \approx 3500$. However, it does not exhibit the mode on the heavy right tail, i.e. the rich countries' mode. Jones method linearizes the slope until zero (from the right) what causes several problems in practice (density does not start at zero in zero nor it does exhibit the two first convergence clubs). Our simple boundary correction method is the only one that allows the estimator to reveal all interesting characteristics of this density. We refer to Vollmer (2009) for more discussion on the behavior of the cross-country income distribution.

The second application requires a nonparametric but boundary corrected regression. Since almost the beginning of econometrics, the specification and estimation of Engel curves has attracted the attention of many economists and applied econometricians. A detailed discussion and review of the parametric approaches to these problems are given in Deaton and Muellbauer (1980); an analysis of the cross-sectional consumer behavior in the context of fully nonparametric models can be found in Bierens and Pott-Buter (1990) or Engel and Kneip (1996). Still today, Engel curves are of special interest in welfare analysis, and especially affected by a boundary problem at the left in poor countries like Indonesia. In Figure 8, we see $n = 6242$ observations of household annual food and total consumption expenditures per capita for the whole country (left), and among them $n = 502$ observations for the province North Sumatra (right). The source of these data is the second wave of Indonesia Family Life Survey (IFLS) in 1997. Graphically, we fit an Engel curve to the left scatter plot of food versus total expenditure on a grid of 200 equidistant points with a natural left boundary at $a = 0$. Certainly, for poverty, welfare, and development analysis we pay special attention to the poorest, and these are exactly at the boundary. Again, the usefulness of our correction method is evident but one might argue that the local linear estimator does as well. There are several pro and cons and we do not want to enter the question which estimator has to be preferred. What we can say is that it seems that with our boundary correction the Nadaraya-Watson can compete with local linear estimation.

# References

Abramson, I.S., 1982. On bandwidth variation in kernel estimates-a square root law. The Annals of Statistics 10(4), 1217-1223.

Acemoglu, D. and Ventura, J., 2002. The world income distribution. The Quarterly Journal of Economics 117(2), 659-694.

Bierens, H.J., and Pott-Buter, H.A., 1990. Specification of household engel curves by

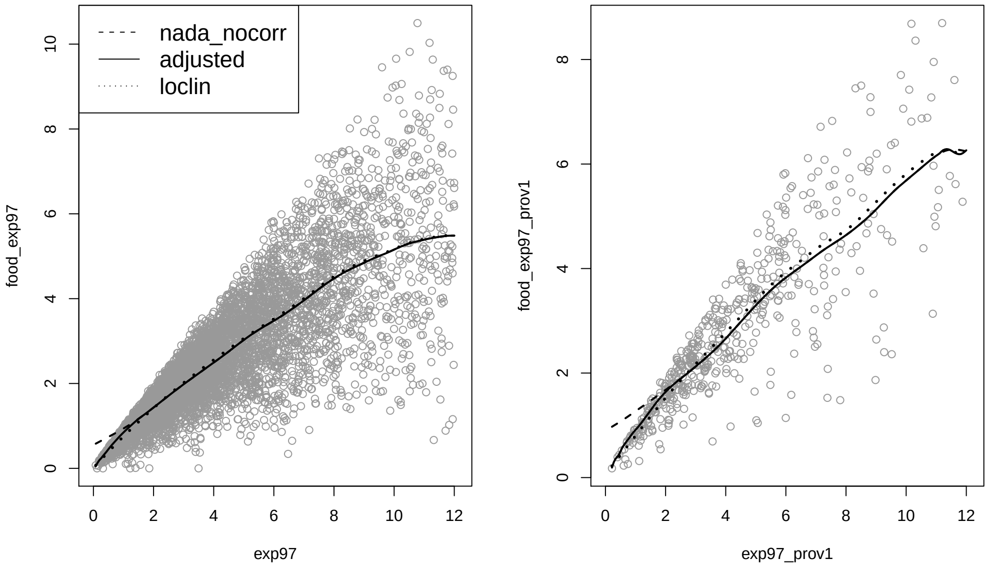

Figure 8: Comparison of Engel curve estimates in 1997 with $h_{global} = 1.5$ (left, for Indonesia) and $h_{global} = 2.4$ (right, for North Sumatra) which is Silverman's rule-of-thumb times 3. Black dashed line is Nadaraya-Watson estimate without boundary correction, black solid is our method, and dotted line is local linear. Scale: x-axis $10^6$, y-axis: $10^6$.

non-parametric regression. Econometric Reviews 9(2), 123-184.

Cheng, M.Y., Fan, J.Q and Marron, J.S., 1997. On automatic boundary corrections. The Annals of Statistics 25(4), 1691-1708.

Cline, D.B.H. and Hart, J.D., 1991. Kernel estimation of densities with discontinuities or discontinuous derivatives. Statistics 22, 69-84.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatter plots. J. Amer. Statist. Assoc. 74, 829-836.

Cleveland, W.S., 1981. LOWESS: A program for smoothing scatter plots by robust locally weighted regression. The American Statistician 35, 54.

Cleveland, W.S., Grosse, E. and Shyu, W.M., 1992. Local regression models, in: Chambers, J.M. and Hastie, T.J. (Eds.), Chapter 8 of Statistical Models in S. Wadsworth & Brooks/Cole.

Cowling, A., and Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society, Series B 58(3), 551-563.

Deaton, A., and Muellbauer, J., 1980. Economics and Consumer Behavior, Cambridge University Press, Cambridge.

Engel, J., and Kneip, A., 1996. Recent approaches to estimating engel curves. Journal of Economics 63(2), 187-212.

Fan, J.Q. and Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. The Annals of Statistics 20(4), 2008-2036.

Gasser, T. and Müller, H.-G., 1979. Kernel estimation of regression functions, in: Gasser, T. and Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation (Lecture Notes in Mathematics 757). Springer Verlag, Berlin, pp. 23-68.

Gasser, T., Müller, H.-G. and Mammitzsch, V., 1985. Kernels for nonparametric curve estimation, Journal of the Royal Statistical Society, Series B 47(2), 238-252.

Hall, P., 1983. On near neighbour estimates of a multivariate density. Journal of Multivariate Analysis 13, 24-39.

Hall, P. and Marron, J.S., 1988. Variable window width kernel estimates of probability densities. Probability Theory and Related Fields 80, 37-49.

Hall, P. and Wehrly, T. E., 1991. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. Journal of the American Statistical Association 86(415), 665-672.

Hall, P. and Park, B.U., 2002. New methods for bias correction at endpoints and boundaries. The Annals of Statistics 30(5), 1460-1479.

Holzmann, H., Vollmer, S. and Weisbrod, J., 2007. Perspectives on the world income distribution - beyond twin peaks towards welfare conclusions, in: Proceedings of the German Development Economics Conference, Göttingen.

Jones, M.C., 1993. Simple boundary correction in kernel density estimation. Statistics and Computing 3, 135-146.

Müller, H.-G., 1991. Smooth optimum kernel estimators near endpoints. Biometrika 78(3), 521-530.

Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and Its Applications 10, 186-190.

Rice, J., 1984. Boundary modification for kernel regression. Communications in Statistics - Theory and Methods 13(7), 893-900.

Ruppert, D. and Cline, D.B.H., 1994. Bias reduction in kernel density estimation by smoothed empirical transformations. The Annals of Statistics 22(1), 185-210.

Ruppert, D. and Marron, J.S., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society, Series B(Methodological) 56(4), 653-671.

Sala-I-Martin, X., 2006. The world distribution of income: falling poverty and ... convergence, period. The Quarterly Journal of Economics 121(2), 351-397.

Schuster, E.F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods 14(5), 1123-1136.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Vollmer, S., 2009. A Contribution to the Empirics of Economic and Human Development, Ph.D. Dissertation, Georg-August-University Göttingen. Peter Lang, Frankfurt am Main.

Wand, M.P., Marron, J. S. and Ruppert, D., 1991. Transformations in density estimation. Journal of the American Statistical Association 86(414), 343-353.

Yang, L. and Marron, S., 1999. Iterated transformation-kernel density estimation. Journal of the American Statistical Association 94(446), 580-589.

Yang, L., 2000. Root-n convergent transformation-kernel density estimation. Journal of Nonparametric Statistics 12(4), 447-474.

Zhang, S. and Karunamuni, R.J, 1998. On kernel density estimation near endpoints. Journal of Statistical Planning and Inference 70, 301-316.

Zhang, S., Karunamuni, R.J and Jones, M.C. 1999. An improved estimator of the density function at the boundary. Journal of the American Statistical Association 94(448), 1231-1241.

Zhang, S. and Karunamuni, R.J, 2000. On nonparametric density estimation at the boundary. Journal of Nonparametric Statistics 12(2), 197-221.