A REMEDY FOR KERNEL ESTIMATION UNDER RANDOM DESIGN

Alois Kneip

Institut des Statistique
Universite Catholique de Louvain
34 Voie du Roman Pays
1348 Louvain–La–Neuve
Belgium

Joachim Engel

Wirtschaftstheorie II
Universität Bonn
Lennestr. 37
53225 Bonn
Germany

August 1994

Summary

Two common kernel–based methods for non–parametric regression estimation suffer from well–known drawbacks when the design is random. The Gasser–Müller estimator is inadmissible due to its high variance while the Nadaraya–Watson estimator has zero asymptotic efficiency because of poor bias behavior. Under asymptotic considerations, the local linear estimator avoids these two drawbacks of kernel estimators and achieves minimax optimality. However, when based on compact support kernels its finite sample behavior is disappointing because sudden kinks may show up in the estimate.

This paper proposes a modification of the kernel estimator, called the binned convolution estimator leading to a fast $O(n)$ method with asymptotic properties identical with those of the local linear estimator.

Key words and phrases: kernel regression; local polynomials; smoothing; binning;

AMS 1991 Subject Classification: Primary 62G07, Secondary 62F11, 62J07, 65D10
Short title: Kernel Regression

Acknowledgement: Research supported by Sonderforschungsbereich 303 (project A 3) with funds from the Deutsche Forschungsgemeinschaft

1 Introduction

Kernel estimators are very polular methods for nonparametric regression. They have been thoroughly studied and are quite well understood. Furthermore, several other nonparametric curve estimators such as smoothing splines and local polynomial estimators are known to be equivalent to kernel estimators in an asymptotic sense .

Two competing weighting schemes for calculating kernel weights are the Nadaraya–Watson estimator (Nadaraya, 1964 and Watson, 1964; short: N–W estimator) and the Gasser–Müller estimator (1979) (short: G–M estimator). In case of the regressor variable being under control of the experimenter (fixed design) the G–M method is the established procedure because of its superior bias behavior. However, under a random design, where the regressor variable as well as the dependent variable are random observations, the situation is less clear–cut: considering bias, the advantages of the G–M method still hold, but the N–W method has smaller variance. For further discussions of the two kernel estimators, see Gasser & Engel (1990) and Chu & Marron (1991). Under asymptotic aspects, the local linear estimator is a method that combines the advantages of both kernel methods. For random design its variance equals the N–W variance while its bias is identical with that of the G–M estimator. Furthermore, the local linear estimator automatically adjusts to the boundary of the domain of estimation while the kernel methods require the introduction of particular boundary kernels for the bias at boundary points to be as low as in the interior. Moreover, Fan (1992) has shown the local linear estimator to be asymptotically minimax efficient among all linear smoothers.

For the applied statistician, however, finite sample performance may be more important than asymptotic behavior. When applied to small or moderate data sets the local linear estimator based on kernels with finite support is disappointing because then its denominator may become zero. To alleviate this draw–back, Fan's (1992) proposal to add a small term of $(1/n^2)$ into the denominator avoids zero denominators but may introduce artifical kinks in the estimate. Recently, Seifert and Gasser (1994) proposed local bandwidth enlargement to avoid this pitfall of the local linear estimator.

Besides approximating the true regression regression function as closely as possible, a second important concern in applied data analysis is speed of computation. Usually nonparametric curve estimation methods are very computer intensive, in particular when the smoothing factor has to be chosen from the data.

In this paper, we propose a modification of the kernel estimator: the binned convolution estimator (BCE). A first step of binning the data is followed by an application of the G–M kernel estimator. The BCE enjoys the same asymptotic properties as the local linear estimator. Hence it also is asymptotically minimax efficient. The BCE is closely related to the double smoothing estimator (short: DSE) recently proposed by Wu and Chu (1992). The DSE is obtained by pre–smoothing the data with the N–W method to an equidistant grid. In a second step the G–M estimator is applied to the presmoothed data. To obtain the required rate of convergence the pre–smoothing bandwith b has to be much smaller than the bandwidth h in the second smoothing step. However, Wu and Chu's paper gives no advise on how to actually choose these two bandwidths. Using the N–W method in step 1 is an unnecessary complication compared to the computationally much simpler idea of binning. Consequently, the BCE is computationally much faster then the DSE. Its computational speed is comparable with the WARPING algorithm of Härdle and Scott (1992). In section 2 we introduce our notation and review briefly the pros and cons of the kernel regression and local linear methods. Section 3 introduces the binned convolution estimator and provides mathematical results regarding its asymptotic properties. Implementation of the BCE requires the choice of a binwidth and of a bandwidth. We derive an explicit formula for the optimal relationship between these quantities in one important special case: the design is uniform. This result serves as a guide when choosing the bin size. Results of a simulation study comparing the BCE with N–W and G–M kernel estimators and with the local linear estimator are summarized in section 4.

2 Kernel Estimation and Local Linear Regression

We consider the problem of univariate regression estimation under random design. Given data $(X_i, Y_i) \in [0,1] \times \mathrm{I\!R}, \quad i = 1, \ldots, n$, the objective is to estimate the mean regression function

$$\begin{aligned} m(x) &= \mathrm{E}(Y \mid X = x) \\ &= \frac{\int y f_{X,Y}(x,y)\mathrm{d}y}{f_X(x)} \end{aligned} \tag{2.1}$$

where $f_{X,Y}$ and f_X denote joint and marginal density of the random vector (X, Y) respectively random variable X. Alternatively, the regression model can be described by

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where X_i and ϵ_j are independent with $\mathrm{E}(\epsilon_j) = 0, \quad \mathrm{Var}(\epsilon_j) = \sigma^2$. For simplicity, we consider the homoscedastic case only.

The Nadaraya–Watson estimator is obtained by replacing the densities in (2.1) with kernel density estimators (using a product kernel for estimating the bivariate density). This leads to

$$\hat{m}_{\mathrm{NW}}(x) = \frac{\frac{1}{nh^2}\int y \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-Y_i}{h}\right)\mathrm{d}y}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}$$

$$= \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{K\left(\frac{x-X_i}{h}\right)}. \tag{2.2}$$

Here the kernel function K is usually chosen as symmetric probability density and the parameter h denotes the smoothing factor or bandwidth.

The Gasser–Müller estimator can be represented as a convolution of the data with the kernel. Integration is involved to guarantee that the kernel weights add up to 1. The G–M method assumes data $(X_i, Y_i), i = 1, \ldots, n$ ordered according to X_i and is defined as

$$\hat{m}_{\mathrm{GM}}(x) = \frac{1}{h} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K\left(\frac{x-z}{h}\right) \mathrm{d}z\, Y_i. \tag{2.3}$$

Here $s_i = (X_i + X_{i+1})/2, i = 1, \ldots, n-1, s_0 = 0, s_n = 1$.

Conceptually the local polynomial estimator takes a different approach. Provided the regression function is sufficiently smooth, $m(z)$ can be approximated locally by a polynomial

$$m(z) \approx m(x) + \theta_0 + \theta_1(z-x) + \theta_2(z-x)^2 + \ldots \theta_k(z-x)^p$$

for x close to z.Then the local polynomial regression method is based on a weighted local linear regression

$$\min_{\theta_0,\ldots,\theta_p} \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1(X_i - x) - \ldots - \theta_k(X_i - x)^p)^2 K\left(\frac{x-X_i}{h}\right). \tag{2.4}$$

The regression function m is estimated at point x through

$$\hat{m}(x)_{\mathrm{locpol}} = \hat{\theta}_0. \tag{2.5}$$

Notice that with local constant approximation, i.e. $p = 0$ we obtain the Nadaraya–Watson kernel estimator. Here we take a particular interest in local linear approximation ($p = 1$). Higher order approximations correspond to the use of higher order kernels in kernel regression. Fan (1992) derives for the local linear estimator

$$\hat{m}(x) = \sum_{i=1}^{n} \frac{w_i Y_i}{\sum_{i=1}^{n} w_i + n^{-2}} \tag{2.6}$$

where the weights w_i are calculated as

$$w_i = w_i(x, X_1, \ldots, X_n, h) = K\left(\frac{x-X_i}{h}\right)[t_2 - (x - X_i)t_1]$$

and

$$t_\ell = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(x - X_i)^\ell, \ell = 1, 2.$$

Fan introduces the term n^{-2} in the denominator of (2.6) to avoid zero denominators. Due to its smoothness property the Gaussian kernel is a popular choice for the kernel function K. However, optimal in the L_2 sense is the Epanechnikov kernel (Fan et al.; 1993).

For an interior point x with $f_X(x) > 0$ asymptotic bias and variance of the estimators are given as follows:

Estimator	Bias	Variance
Gasser–Müller	$b(x)$	$1.5v(x)$
Nadaraya–Watson	$b_{\mathrm{NW}}(x)$	$v(x)$
Local linear	$b(x)$	$v(x)$

Here $b(x) = h^2/2m''(x)\sigma_K^2$, $b_{\mathrm{NW}}(x) = h^2\sigma_K^2[m''(x)/2 + m'(x)f_X'(x)/f_X(x)]$ and $v(x) = \sigma^2 R(K)/(nhf_X(x)$ where $\sigma_K^2 = \int K(u)u^2\mathrm{d}u$, $R(K) = \int K^2(u)\mathrm{d}u$.

To estimate at a point near the boundary, the kernel methods require the use of particularly taylored boundary kernels in order for the bias to stay as low as in the interior. In contrast, the local linear estimator automatically adjusts to the boundary. While we see from above table that the G–M kernel estimator is inadmissable Fan (1992) shows that the local linear estimator is asymptotically minimax efficient among all linear estimator. Furthermore, the G–M estimator has a relative efficiency of 66.7%, while the N–W estimator even has zero relative efficiency.

This superior asymptotic performance of the local linear approximation method is contrasted by drawbacks this estimator has when applied to a finite sample. As Seifert and Gasser (1993) illustrate, for regions of sparse data the local linear estimator displays a rugged local behavior. A sparse region is a smoothing interval with less then $p + 3$, i.e. 4 design points in the local linear case. For small and moderate sample size, sparse regions in the realization of the design occur rather frequently, in particular under random design. When using a kernel with compact support like the Epanechnikov kernel, then the unconditional variance of the local linear estimator becomes infinite. Under the Gaussian kernel the rugged behavior is less pronounced. As illustration consider the simulated example in Figure 1. The regression function consists of a sine wave added to a linear increasing function. The design is a realization of the Beta(2,2) density on the unit interval. To exclude particular boundary effects, we focus on the interval $[.1, .9]$. $n = 100$ data points were generated with a variance of $\sigma^2 = .5$. Figure 1 shows the true regression curve, the G–M and N–W kernel estimate and the local linear estimate (for better visual separation a constant has been subtracted). All estimates are based on a MISE–optimal bandwidth, obtained through a grid search algorithm.

put Figure 1 about here

Notice the two sudden sharp peak of the local linear estimator around $x = .18$, and $x = .21$ occuring in a region where the regression function is linear, i.e. where conceptually a local linear approximation should be particularly well suited.

3 Binned Convolution Estimator

As Figure 1 illustrates, the local linear estimator may display erratic behavior in areas of sparse data. On the other hand, as Chu and Marron (1991) have shown, the Gasser–Müller estimator has problems when the design is clustered. Then a loss of efficiency occurs resulting in an increase of variance by the factor 1.5. The conceptual idea of the binned convolution estimator is to first transform the random design data into a (basically) fixed equidistant design and then to apply the G–M kernel estimator to the transformed data. The transformation is defined through binning. Then possible clusters in the design disappear, the only exception from equidistance occuring when some bins are empty.

Definition 1

Step 1: Discretizing the data into L bins defined by

$$B(k) = [(k-1)/L, k/L), \quad k = 1, \ldots, L \tag{3.1}$$

Step 2: Defining "new"data by calculating bin averages:

$$\tilde{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n} Y_i I(X_i \in B(k)), \quad k = 1, \ldots, L \tag{3.2}$$

where $n_k = \#\{X_i \mid X_i \in B(k)\}$

Step 3: Finally, smoothing the binned data

$$\hat{m}(x) = \hat{m}_{BC}(x) = \frac{1}{h} \sum_{k=1}^{L} \int_{(k-1)/L}^{k/L} K\left(\frac{x-z}{h}\right) dz\, \tilde{Y}_k. \tag{3.3}$$

An immediate advantage of the BCE over other methods considered here is its speed and its ease of computation. The G–M estimator as well as the local linear approximation can be computed through an O(n) algorithm (Seifert, Brockmann, Engel and Gasser; 1994), i.e. above step 3 requires $O(L)$ computations. The regressogram–like binning in step 1 can be done particularly fast and simply. Hence, the number of evaluations necessary to compute the BCE certainly is linear in the number of observations n. Moreover, its computational speed is somewhat slower than that of a regressogram, but certainly faster the (unbinned) $O(n)$ algorithm of Seifert et al.

The concept of the BCE is closely related to smoothing by weighted averaging of rounded points (WARPING; Härdle and Scott, 1992). Besides relying on the N–W method instead of the G–M estimator in step 3 (which for a fixed equidistant design is irrelevant), the main difference is that the WARPING algorithm in step 3 applies weights depending on the number of observations falling into each bin while the BCE treats all bins equally. More precisely, when representing the smoothing step 3 as $\hat{m}(x) = \sum_{k=1}^{L} w^{(k)}(x)\tilde{Y}_k$ the WARPING weights are approximately

$$w_{\mathrm{WARP}}^{(k)}(x) \quad = \quad \frac{K((x-k/L)/h)}{\sum_{j=1}^{L} K((x-j/L)/h)n_j/n_k}.$$

while the BCE weights amount to

$$w_{\mathrm{BCE}}^{(k)}(x) \quad = \quad 1/h\int_{(k-1)/L}^{k/L} K((x-u)/h) \quad \mathrm{d}u \approx K((x-k/L)/h)/(hL).$$

Being a fast algorithm for approximating the N–W estimator, the asymptotics of the WARPING estimator are very similar to the Nadaraya–Watson asymptotics.

To investigate the asymptotic behavior of the BCE we make the following
Assumptions

(A 1). The support of the density f_X is a possibly unbounded subinterval of $[0,\infty)$. There exists a $\delta > 0$ such that for any $0 \le \eta \le \delta$ the set $\{x \mid f_X(x) \ge \eta\}$ is connected.

(A 2). The density f_X is at least three times continuously differentiable.

(A 3). The regression function $m(x)$ is at least twice continuously differentiable for every x in the support of f_X.

(A 4). $\inf_{z\in[0,1]} f_X(z) > 0$

(A 5). The conditional fourth moment is finite, i.e. $E(Y^4 \mid X = x) < \infty$.

Theorem 1 *Suppose that above Assumptions (A1–A4) hold. Furthermore, assume that* $b := 1/L \to 0, \log(n)^4/(bn) \to 0, h \to 0, b/h \to 0$ *as* $n \to \infty$. *Then*

(a).

$$\mathrm{Bias}_h m(x)^2 := (\mathrm{E}\hat{m}(x) - m(x))^2 = \frac{h^4}{4} c_K m''(x)^2 + o(h^4) \tag{3.4}$$

where $c_K = \int K(v)v^2 \mathrm{d}v$.

(b).

$$\mathrm{Var}\left(\hat{m}(x)\right) = \mathrm{E}\left(\hat{m}(x) - \mathrm{E}\hat{m}(x)\right)^2 = \frac{d_K \sigma^2}{nhf_X(x)} + o\left(\frac{1}{nh}\right) \tag{3.5}$$

where $d_K = \int K(v)^2 \mathrm{d}v$

Theorem 1 implies minimax efficiency for the BCE within the class of all linear smoothers provided the bin size is chosen within the correct boundaries. The following proposition represents bias and variance as function of bandwidth h and bin size b. This representation allows to derive a formula for the asymptotically optimal bin size. To facilitate computations, we first need some simplifying assumptions:

Assumptions

(B 1). The design density f_X is the uniform density.

(B 2). The regression function m is six times continuously differentiable.

(B 3). The kernel function K is the Epanechnikov kernel, i.e. $K(x) = 3/4(1-x^2), x \in [-1,1]$.

Under the above assumptions (A1–A4) and (B1–B3) let us consider estimation of the regression function $m(x)$ at the bin points $x = r/L, r = 0, \ldots, L$.

Proposition 1 *Let* $x \in \{r/L\}_{r=0,\ldots,L}, x \in [h, 1-h]$, *and in addition to above assumptions suppose that* $m''(x) \neq 0$. *Furthermore assume that* $b/h \to 0, (nb)^{-1} \to 0, h^3/b \to 0$ *as* $n \to \infty$. *Then*

(a).

$$\mathrm{E}\hat{m}(x) = m(x) + \frac{h^2}{10} m''(x) \left[1 + \frac{5}{6}\frac{b^2}{h^2} + o\left(\frac{b^2}{h^2}\right)\right] + R_h \tag{3.6}$$

where $R_h = O(h^4)$ *is independent of* b.

(b).

$$\mathrm{Var}\left(\hat{m}(x)\right) \leq \frac{3}{5}\frac{\sigma^2}{nh}\left[1 + \frac{1}{nb} + o\left(\frac{1}{nb}\right)\right] \tag{3.7}$$

The proposition implies that the mean squared error can be written in the form

$$\begin{aligned}
\mathrm{MISE}_{h,b}(x) &= \mathrm{E}\left(\hat{m}(x) - m(x)\right)^2 \\
&= \frac{h^4}{100} m''(x)^2 \left(1 + \frac{5}{6}\frac{b^2}{h^2}\right) + \frac{3}{5}\frac{\sigma^2}{nh}\left(1 + \frac{1}{nb}\right) + R_h^{(1)} + R_{h,b}^{(2)} \\
&= \mathrm{AMISE} + R_h^{(1)} + R_{h,b}^{(2)}
\end{aligned}$$

where $R_h^{(1)} = O(h^6)$ is independent of b, and $R_{h,b}^{(2)} = o(b^2 + [n^2hb]^{-1})$. Our aim is to choose h and b minimizing $\text{AMISE}_{h,b}(x)$. However, since $b/h \to 0$ and $1/(nb) \to 0$, the influence of terms depending on b is asymptotically neglible when determining an optimal bandwidth h_{opt}. Hence, we consider $\text{AMISE}_{h,b}(x)$ at the asymptotically optimal bandwidth

$$h_{\text{opt}} = \left(\frac{15\sigma^2}{m''(x)^2}\right)^{1/5} n^{-1/5}.$$

This leads to

$$\begin{aligned}\text{AMISE}_{h_{\text{opt}},b}(x) &= \left(\frac{m''(x)^2}{100}\right)^{1/5} \left(\frac{3}{5}\sigma^2\right)^{4/5} 4^{-4/5} \left(1 + \frac{5}{6}\frac{b^2}{h_{\text{opt}}^2}\right) n^{-4/5} \\ &\quad + \left(\frac{m''(x)^2}{100}\right)^{1/5} \left(\frac{3}{5}\sigma^2\right)^{4/5} 4^{1/5} \left(1 + \frac{1}{nb}\right) n^{-4/5}.\end{aligned}$$

Obviously, minimizing $\text{AMISE}_{h_{\text{opt}},b}(x)$ with respect to b is equivalent to minimizing

$$4^{-4/5}\frac{5}{6}\frac{b^2}{h_{\text{opt}}^2} + \frac{4^{1/5}}{nb}.$$

We thus obtain for the optimal bin size b_{opt}

$$\begin{aligned} b_{\text{opt}} &= \left(\frac{12}{5}\right)^{1/3} \left(\frac{h_{\text{opt}}^2}{n}\right)^{1/3} \\ &= \left(\frac{12}{5}\right)^{1/3} \left(\frac{15\sigma^2}{m''(x)^2}\right)^{2/15} n^{-7/15}.\end{aligned}$$

4 Simulations

To evaluate the practical performance of the BCE its behavior when applied to real or simulated data is of interest. For an application to a *real* data set, Engel and Kneip (1994) use the BCE to estimate cross–sectional market demand curves and compare its performance with several other curve estimators. The computational speed of the BCE allowed the authors to implement this method on a 486 PC for estimating demand curves from large household expenditure surveys (sample size around $n = 8,000$).

Here we assess the efficacy of the BCE through a simulation study. From ten different regression curves covering a broad range of situations data of a sample size between $n = 50$ to $n = 2500$ were generated. The data were perturbed by a normal error of variance .1, 1. and 2.5. The design was chosen according to a beta distributions with parameters (1,1), (2,2) and (.5,.5). Figure 2 shows four selected regression curves m_1 to m_4. Results for the remaining six curves were very similar.

Put Figure 2 about here

To summarize the results we calculated the integrated square error (ISE) for each estimation method and averaged over $nsim = 200$ simulations. Figure 3 to 5 show the average ISE as a function of the sample size for the 'true' regression curves 1,2 and 3.

Put Figure 3 – 5 about here

For a small sample size the BCE does not do very well. Binning the data when the sample size is small obviously does not make much sense. Despite its asymtotic deficiencies, for a small sample size the G–M estimator performs quite well unless the variance gets too large. With increasing sample size, not surprisingly, the BCE and the local linear estimator clearly outperform the N–W and G–M method. For a moderate sample size (100–500) and a small variance the local linear estimator seems to be slightly better, but for larger variances the BCE performs better than the local linear approximation. For high sample size this difference in performance seems to fade away, validating empirically the asymptotic equivalence of these two methods. To obtain some insight into the spread of how well the various methods under consideration perform, Figure 6 shows boxplots of ISE for $nsim = 200$ simulated data sets from regression function 4 based on a Beta(2,2) design and an error variance of $\sigma^2 = 1$. Considering the scale of regression function m_4 a variance of 1. is quite high explaining the poor performance of the G–M estimator. If any discernable difference at all, Figure 6 shows that the overall good performance of the BCE is less spread then the equally good showing of the local linear estimator.

Put Figure 6 about here

Appendix

In order to prove Theorem 1, we need the following

Lemma 1 *Let $W_1, \dots, W_n$ be i.i.d. real random variables, and let Z be a real random variable with $\mathrm{E}(Z^2) < \infty$. Furthermore, suppose that the assumptions of Theorem 1 hold, let $n_{w,r} := \max\{1, \sum_{i=1}^n I(W_i \in ((r-1)/L, r/L])\}$ and $\Pi_r := P(W_1 \in ((r-1)/L, r/L])$ for $r \in \{1, \dots, L\}$. Then there exists an $n_0 \in \mathbb{N}$, independent from the distribution of W_1, Z such that for all $n \geq n_0$*

(a).

$$P\left(\max_{r=1,\dots,L}\frac{\mid n_{w,r}-\max\{1,n\Pi_r\}\mid}{(\max\{1,n\Pi_r\})^{1/2}}\geq \log(n)^2\right)\leq 2\exp\left(-\frac{1}{5}\log(n)^2\right)$$

(b). *For all* $r\in\{1,\dots,L\}$ *and* $\alpha\in\{1,2\}$

$$\mathrm{E}\mid \frac{Z}{n_{w,r}}-\frac{Z}{\max\{1,n\Pi_r\}}\mid^{\alpha}\leq \frac{(\mathrm{E}\mid Z\mid^{\alpha})\log(n)^{2\alpha}}{(\max\{1,n\Pi_r\})^{3/2\alpha}}+R_{w,z,r}$$

where

$$R_{w,z,r}:=\begin{cases}2(\mathrm{E}Z^{2\alpha})^{1/2}\exp(-1/5\log(n)^2) & \textit{if}\quad \log(n)^2/(\max\{1,n\Pi_r\})^{1/2}\leq 1/2\\ (\mathrm{E}Z^{2\alpha})^{1/2}n^{\alpha} & \textit{else}\end{cases}$$

Proof of Lemma 1:
Let $J_r(x):=I(x\in((r-1)/L,r/L])$. The random variable $\sum_{i=1}^n J_r(W_i)$ has a binomial distribution with $\mathrm{E}(\sum_{i=1}^n J_r(W_i))=n\Pi_r$ and $\mathrm{Var}(\sum_{i=1}^n J_r(W_i))=n\Pi_r(1-\Pi_r)$. Bernstein's inequality then implies that for any $\eta>0$ and every $r\in\{1,\dots,L\}$

$$\begin{aligned}P(\mid n_{w,r}-\max\{1,n\Pi_r\}\mid\geq\eta) &\leq P\left(\mid\sum_{i=1}^n J_r(W_i)-n\Pi_r\mid\geq\eta\right)\\ &\leq 2\exp\left(-\frac{\eta^2}{2(n\Pi_r(1-\Pi_r)+\eta/3)}\right).\end{aligned}$$

It follows that

$$\begin{aligned}&P\left(\max_{r=1,\dots,L}\frac{\mid n_{w,r}-\max\{1,n\Pi_r\}\mid}{(\max\{1,n\Pi_r\})^{1/2}}\geq\log(n^2)\right)\\ &\leq\sum_{r=1}^n P\left(\mid n_{w,r}-\max\{1,n\Pi_r\}\mid\geq(\max\{1,n\Pi_r\})^{1/2}\log(n)^2\right)\\ &\leq 2\exp\left(-\frac{1}{4}\log(n)^2+\log(L)\right),\end{aligned}$$

which proves Assertion a). As regard to Assertion b), note that

$$\begin{aligned}&\mathrm{E}\mid\frac{Z}{n_{w,r}}-\frac{Z}{\max\{1,n\Pi_r\}}\mid^{\alpha}=\mathrm{E}\frac{\mid Z(\max\{1,n\Pi_r\}-n_{w,r})\mid^{\alpha}}{n_{w,r}^{\alpha}\max\{1,n\Pi_r\}^{\alpha}}\\ &\leq\frac{\mathrm{E}\mid Z\mid^{\alpha}(\max\{1,n\Pi_r\})^{\alpha/2}\log(n)^{2\alpha}}{(\max\{1,n\Pi_r\})^{2\alpha}(1/2)^{\alpha}}+\left(\mathrm{E}Z^{2\alpha}\right)^{1/2}n^{\alpha}\\ &\cdot P\left\{\frac{n_{w,r}}{\max\{1,n\Pi_r\}}<1/2\quad\text{or}\quad\mid n_{w,r}-\max\{1,n\Pi_r\}\mid\geq(\max\{1,n\Pi_r\})^{1/2}\log(n)^2\right\}\end{aligned}$$

Assertion b) now follows from Assertion a) and the fact that if $\log(n)^2/\max\{1, n\Pi_r\}^{1/2} \leq 1/2$, then $| n_{w,r} - \max\{1, \Pi_r\} | < (\max\{1, n\Pi_r\})^{1/2}\log(n)^2$ implies that $n_{w,r}/\max\{1, n\Pi_r\} \geq 1/2$□.

Proof of Theorem 1

We first prove assertion a). For $r = 0, \ldots, L$ let $s_r := (r + .5)/L$ and $n_r := \max\{1, \sum_{i=1}^n J_r(X_i)\}$, $J_r(x) = I(x \in ((r-1)/L, r/L])$. Lemma 1a) and the assumptions of the theorem imply that

$$\begin{aligned}
&\sup_{r=1,\ldots,L} \left| \mathrm{E}\left(\frac{\sum_{i=1}^n m(X_i)J_r(X_i)}{n_r}\right) - m(s_r) \right| \\
&\quad \leq \sup_{r=1,\ldots,L} \left| \mathrm{E}\left(m(s_r)\frac{\sum_{i=1}^n J_r(X_i)}{n_r}\right) - m(s_r) \right| \\
&\quad + \sup_{r=1,\ldots,L} \mathrm{E}\left(\sup_{z\in[(r-1)/L,r/L]} | m'(z) | \, b \sum_{i=1}^n J_r(X_i)/n_r\right) \\
&\quad = \sup_{z\in[0,1]} | m'(z) | \, b + o_{\mathrm{P}}\left(\frac{1}{n}\right). \qquad (4.1)
\end{aligned}$$

Since $\mathrm{E}(Y_i \mid X_i) = m(X_i)$, and since $\int_{(r-1)/L}^{r/L} K((x-u)/h)\mathrm{d}u \leq \max_{z\in[-1,1]} K(z)b$ is nonzero only for a multiple of $Lh \sim h/b$ elements, $r \in \{1, \ldots, L\}$, it follows that

$$\begin{aligned}
&\left| \mathrm{E}\hat{m}(x) - \frac{1}{h}\sum_{r=1}^L \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right)\mathrm{d}u \quad m(s_r) \right| \\
&\quad \leq \left| \frac{1}{h}\sum_{r=1}^L \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right)\mathrm{d}u \sup_{z\in[0,1]} | m'(z) | \, b \right| \\
&\quad = O(b). \qquad (4.2)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\left| \frac{1}{h}\sum_{r=1}^L \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right)(m(s_r) - m(u))\mathrm{d}u \right| \qquad (4.3) \\
&\quad \leq \left| \frac{1}{h}\sum_{r=1}^L \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right)\mathrm{d}u \sup_{z\in[0,1]} | m'(z) | \, b \right| \\
&\quad = O(b) \qquad (4.4)
\end{aligned}$$

and we obtain

$$\begin{aligned}
\mathrm{E}\hat{m}(x) &= \frac{1}{h}\int K\left(\frac{x-u}{h}\right) m(u)\mathrm{d}u + O(b) \\
&= \int K(v)m(x-hv)\mathrm{d}v + O(b).
\end{aligned}$$

A Taylor expansion of m together with $\int K(v)\mathrm{d}v = 1, \int K(v)v\mathrm{d}v = 0$ now leads to the asserted result.

It remains to prove Assertion b). Since for any function W_n of the i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ it holds

$$\begin{aligned}\mathrm{Var}(W_n) &= \mathrm{E}\left(\mathrm{E}(W_n^2 \mid X_1, \ldots, X_n) - (\mathrm{E}(W_n \mid X_1, \ldots, X_n))^2\right) \\ &\quad + \mathrm{E}(\mathrm{E}(W_n \mid X_1, \ldots, X_n))^2 - (\mathrm{E}W_n)^2,\end{aligned}$$

relation (4.2) implies the existence of a real number $c < \infty$ such that

$$\begin{aligned}\mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) &= \mathrm{E}\left(\frac{\sum_{i=1}^n \sigma^2(X_i) J_r(X_i)}{n_r^2}\right) \\ &\quad + \mathrm{E}\left(\frac{(\sum_{i=1}^n m(X_i) J_r(X_i))^2}{n_r^2}\right) - m(s_r)^2 + R_r^{(1)},\end{aligned}$$

where the remainder term $R_r^{(1)}$ can be bounded by $R_r^{(1)} \leq c_1 b$ for any $r \in \{1, \ldots, L\}$. Arguments similar to those in (4.1) now allow to conclude that

$$\mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) = \mathrm{E}\left(\frac{\sigma^2(s_r)}{n_r}\right) + R_r^{(2)}, \tag{4.5}$$

where for some $c_2 < \infty$ the remainder term is bounded by $\mid R_r^{(2)} \mid \leq c_2 b$ for all $r \in \{1, \ldots, L\}$.

If F denotes the marginal distribution function of the random variables $X_1, \ldots, X_n$, then

$$\begin{aligned}P(X_1 \in ((r-1)/L, r/L]) &= F(r/L) - F((r-1)/L) \\ &= f_X(s_r) b + \sup_{z \in [0,1]} \mid f_X(z) \mid b^2. \end{aligned}\tag{4.6}$$

By assumption $c_3 := \inf_{z \in [0,1]} f_X(x) > 0$ and $(nb)/\log(n)^4 \to \infty$ as $n \to \infty$. We thus can infer from (4.5),(4.6), and Lemma 1 that

$$\mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) = \mathrm{E}\left(\frac{\sigma^2(s_r)}{n f_X(s_r) b}\right) + R_r^{(3)}, \tag{4.7}$$

where $\sup_{r=1,\ldots,L} R_r^{(3)} = o([bn]^{-1})$. Definition of the estimator $\hat{m}(x)$ now leads to

$$\begin{aligned}\mathrm{Var}(\hat{m}(x)) &= \frac{1}{h^2} \sum_{r=1}^{L} \left(\int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right) \mathrm{d}u\right)^2 \mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) \\ &= \frac{1}{h^2} \sum_{r=1}^{L} \left(\int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right) \mathrm{d}u\right)^2 \left(\frac{\sigma^2(\xi_r)}{n f_X(s_r) \mid r/L - (r-1)/L \mid} + R_r^{(3)}\right) \\ &= \frac{1}{h^2} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right) \mathrm{d}u\, K\left(\frac{x-s_r}{h}\right) \frac{\sigma^2(\xi_r)}{n f_X(\xi_r)} + o\left(\frac{1}{nh}\right)\end{aligned}$$

for some mean values $\xi_r \in ((r-1)/L, r/L], r = 1, \ldots, L$. Arguing in a way similar to (4.4), we conclude that

$$\begin{aligned}
\operatorname{Var}(\hat{m}(x)) &= \frac{1}{nh^2} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right)^2 \frac{\sigma^2(u)}{f_X(u)} \mathrm{d}u + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{nh} \int_{(r-1)/L}^{r/L} K(u)^2 \frac{\sigma^2(x-hv)}{f_X(x-hv)} \mathrm{d}v + o\left(\frac{1}{nh}\right) \\
&= \frac{\sigma^2 \int K(v)^2 \mathrm{d}v}{nhf_X(x)} + o\left(\frac{1}{nh}\right) \square
\end{aligned}$$

Proof of Proposition 1

We first prove assertion a.). We obtain

$$\begin{aligned}
\mathrm{E}\,\hat{m}(x) &= \frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v \quad \mathrm{E}\left(\frac{\sum_{i=1}^{n} m(s_r) J_r(X_i)}{n_r}\right) \\
&= \frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v \quad m(s_r) \\
&\quad + \frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v \quad \mathrm{E}\left(\frac{\sum_{i=1}^{n} m(X_i) J_r(X_i)}{n_r} - m(s_r)\right) \\
&=: B_1 + B_2.
\end{aligned}$$

Here, $s_r = (r+.5)/L, r = 0, \ldots, L$. Obviously, $\mathrm{E}(\sum_{i=1}^{n} m(X_i) J_r(X_i)/n_r - m(s_r))$ corresponds to the bias of a Nadaraya–Watson kernel estimator of $m(s_r)$. More precisely, we are in the same situation as for the Nadaraya–Watson kernel estimator with uniform kernel $\bar{K}$ and bandwidth $b/2$. Since $\int \bar{K}(x) x^2 \mathrm{d}x = 1/3$, the well–known properties of the Nadaraya–Watson estimator imply that

$$\begin{aligned}
B_2 &= \frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-u}{h}\right) \mathrm{d}u \quad \frac{b^2}{24} m''(s_r) + o(b^2) \\
&= \frac{b^2}{24} m''(x) + o(b^2). \qquad (4.8)
\end{aligned}$$

Let us now consider B_1. First note that by the choice of x, and by the symmetry of K, we have that

$$\frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v (x-s_r)^s = 0 \quad \text{for } s = 1, 3, 5, \ldots$$

Furthermore, for any $s = 0, 2, 4, 6, \ldots$ the term

$$\frac{1}{h} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)^s$$

is bounded as $L \to \infty$. A Taylor expansion of $m(s_r), r = 1, \ldots, L$ thus leads to

$$\begin{aligned} B_1 &= m(x) + \frac{1}{2}h^2 m''(x)\left(\frac{1}{h}\sum_{r=1}^{L}\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)\mathrm{d}v\left(\frac{s_r - x}{h}\right)^2\right) \\ &\quad + \frac{1}{4}h^4 m^{(4)}(x)\left[\frac{1}{h}\sum_{r=1}^{L}\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)\mathrm{d}v\left(\frac{s_r - x}{h}\right)^4\right] + o(h^6). \qquad (4.9) \end{aligned}$$

The sum in the third term on the right hand side converges to the integral $\int K(v)v^4\mathrm{d}v$, and the difference between the sum and the integral is bounded by b^2/h^2. Furthermore, some algebra shows that

$$\begin{aligned} \frac{1}{h}\sum_{r=1}^{L}\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)\mathrm{d}v\left(\frac{s_r - x}{h}\right)^2 &= \int K(z)z^2\mathrm{d}z + \frac{b^2}{12h^2} + o\left(\frac{b^2}{h^2}\right) \\ &= \frac{1}{5} + \frac{b^2}{12h^2} + o\left(\frac{b^2}{h^2}\right). \qquad (4.10) \end{aligned}$$

Moreover, since $h^6 = o(b^2)$ we conclude that

$$B_1 = m(x) + \frac{1}{2}h^2 m''(x)\left[\frac{1}{5} + \frac{b^2}{12h^2} + o\left(\frac{b^2}{h^2}\right)\right] + R_h \quad ,$$

where $R_h = O(h^4)$ is independent of b. Together with (4.8) this proves assertion a).

It remains to consider assertion b). By our assumptions, it is immediately clear that

$$\mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) = \mathrm{E}\left(\frac{\sigma^2}{n_r}\right).$$

The random variable n_r follows a binomial distribution with $\mathrm{E}n_r = nb$ (recall that by assumption the design density f_X equals the uniform density on $[0,1]$). A Taylor expansion of $1/n_r$ and the arguments in Lemma 1 yield

$$\begin{aligned} \mathrm{E}\frac{1}{n_r} &= \frac{1}{nb} + \frac{\mathrm{E}(n_r - nb)}{(nb)^3} + R_1 \\ &= \frac{1}{nb} + \frac{nb(1-b)}{(nb)^3} + R_2 \\ &= \frac{1}{nb}\left(1 + \frac{1}{nb}\right) + R_3 \end{aligned}$$

where the remainder terms R_1, R_2, R_3 are bounded by $o([nb]^{-2})$. We thus obtain

$$\begin{aligned} \mathrm{Var}\left(\hat{m}(x)\right) &= \frac{1}{h^2}\sum_{r=1}^{L}\left(\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)\mathrm{d}v\right)^2 \mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i J_r(X_i)}{n_r}\right) \\ &= \frac{1}{h^2}\sum_{r=1}^{L}\left(\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)\mathrm{d}v\right)^2 \frac{\sigma^2}{nb}\left(1 + \frac{1}{nb} + o\left(\frac{1}{nb}\right)\right). \qquad (4.11) \end{aligned}$$

It follows from the results of Gasser and Müller (1979) that for large $L (b \to 0)$

$$Q_{h,b} := \frac{1}{h^2} \sum_{r=1}^{L} \left(\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v \right)^2 \frac{\sigma^2}{nb}$$

converges to

$$\frac{\sigma^2}{nh} \int K(x)^2 \mathrm{d}x = \frac{3}{5} \frac{\sigma^2}{nh}.$$

Yet the Cauchy–Schwarz inequality implies that for any r

$$\begin{aligned} \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right) \mathrm{d}v &\leq \left[\int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)^2 \mathrm{d}v \int_{(r-1)/L}^{r/L} \mathrm{d}v \right]^{1/2} \\ &= \left[b \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)^2 \mathrm{d}v \right]^{1/2}. \end{aligned}$$

It follows that

$$\begin{aligned} \frac{1}{h^2} \sum_{r=1}^{L} \left(\int_{(r-1)/L}^{t} {}_r K\left(\frac{x-v}{h}\right) \mathrm{d}v \right)^2 \frac{\sigma^2}{nb} &\leq \frac{1}{h^2} \sum_{r=1}^{L} b \int_{(r-1)/L}^{r/L} K\left(\frac{x-v}{h}\right)^2 \mathrm{d}v \quad \frac{\sigma^2}{nb} \\ &= \frac{\sigma^2}{nh} \sum_{r=1}^{L} \int_{(r-1)/L}^{r/L} \frac{1}{h} K\left(\frac{x-v}{h}\right)^2 \mathrm{d}v = \frac{3}{5} \frac{\sigma^2}{nh}, \end{aligned}$$

which together with (4.11) proves the assertion b). $\square$

References:

Chu, C.K.; J.S. Marron (1991): Choosing a kernel regression estimator. *Statistical Science* **6**, 404–433

Engel, J.; A. Kneip (1994): Recent approaches of estimating income distributions, Engel curves and related functions. *under consideration*

Fan, J. (1993): Local linear regression smoothers and their mini–max efficiencies. *Ann Statist* **21**, 196–216

Fan, J.; T. Gasser; I. Gijbels; M. Brockmann; J. Engel (1993): Local polynomial fitting: a standard for nonparametric regression. *under consideration*

Gasser, T.; J. Engel (1990): Choice of weights in kernel regression. *Biometrika* **77**, 377–381

Gasser, T.; H.G. Müller (1979): Kernel estimation of regression functions. In: Th. Gasser and M. Rosenblatt, eds. *Smoothing Techniques for Curve Estimation*, Lecture Notes in Math., Vol. 757, Springer, New York, 23–58

Gasser, T.; L. Sroka; C. Jennen–Steinmetz (1986): Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633

Härdle, W.; D. Scott (1992): Smoothing by weighted averaging of rounded points. *Computational Statistics*, **7**, 97–128

Nadaraya, E.A. (1964): On estimating regression. *Theory Probab. Appl.* **9**, 141–142

Seifert, B.; M. Brockmann; J. Engel; T. Gasser (1994): Fast algorithms for nonparametric curve estimation, *J Computational and Graphical Statistics* **3**, 192–213

Seifert, B.; T. Gasser (1994): Infinite variance for local polynomials: analysis and solutions. *under consideration*

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

Wu, J.S.; C.K. Chu (1992).:Double smoothing for kernel estimators in nonparametric regression. *J. Nonparametric Statist.* **1**, 375–386

CAPTIONS

Figure 1. True regression function m_1 (dotted line line), and 3 (shifted) nonparametric estimates based on one realization of $n = 100$ design points from a uniform distribution and $\sigma^2 = .1$: G–M estimator (small dashes), local linear estimator (solid line) and N–W estimator (large dashes)

Figure 2. Regression Functions. From left above to right below the functions $m_1, \ldots, m_4$ are plotted.

Figure 3. Plot of log(average ISE) against log(n) for regression function m_1, variances .1,1.,2.5 and design density beta(1,1), beta(2,2) and b(3,3) for BCE (solid), local linear estimator (large dashes), G–M method (small dashes) and N–W method (dotted line)

Figure 4. Plot of log(average ISE) against log(n) for regression function m_2, variances .1,1.,2.5 and design density beta(1,1), beta(2,2) and b(3,3) for BCE (solid), local linear estimator (large dashes), G–M method (small dashes) and N–W method (dotted line)

Figure 5. Plot of log(average ISE) against log(n) for regression function m_3, variances .1,1.,2.5 and design density beta(1,1), beta(2,2) and b(3,3) for BCE (solid), local linear estimator (large dashes), G–M method (small dashes) and N–W method (dotted line)

Figure 6. Boxplots of ISE for regression function m_4, data with design density Beta(2,2) and error variance $\sigma^2 = 1$. Sample size from $n = 50$ to $n = 2500$.