Journal of the Korean
Statistical Society
Vol. 27, No. 1, 1998

Estimation of Density via Local Polynomial Regression †

B.U. Park, W.C. Kim, J. Huh and J.W. Jeon[1]

ABSTRACT

A method of estimating probability density using regression tools is presented here. It is based on equal-length binning and locally weighted approximate likelihood for bin counts. The method is particularly useful for densities with bounded supports, where it automatically corrects edge effects without using boundary kernels.

Key Words : Density estimation, boundary effects, likelihood-based local linear regression, binning.

1. INTRODUCTION

Nonparametric density and regression function estimation are two main subjects of kernel smoothing. One of the major challenges of these problems has been developing methods that correct edge effects when density or regression function has bounded support. Use of boundary kernels (e.g. Rice

†This research was supported in part by the Basic Science Research Institute Program, Ministry of Education, 1996 Project No. 1418

[1]Department of Statistics, Seoul National University, Seoul 151–742, Korea.

1984, Müller 1991) is perhaps the most well-known resolution. One of the unappealing features of this approach is that one needs to use different kernel function for each point of evaluation near boundaries. In regression setting, there is a popular technique of local linear smoothing (e.g. Fan 1993, Fan and Gijbels 1992) which is only minimally influenced by edge effects without using boundary kernels. An interesting question is: can one adapt the idea of local linear regression to density estimation to automatically accommodate edge effects?

In this note we present a methodology which allows regression techniques to be directly applicable to estimating density. It is particularly useful to resolve the aforementioned difficulty. This is illustrated in Figure 1 which is based on the suicide data of Copas and Fryer (1980) given in Table 2.1 of Silverman (1986). The dashed curve is the ordinary kernel density estimate, and the solid one the result of applying the method discussed below, with a histogram represented by dots, and circles on the bottom representing the original data. We can imagine that the solid curve is a considerably improved estimate of the underlying density at and near the origin.

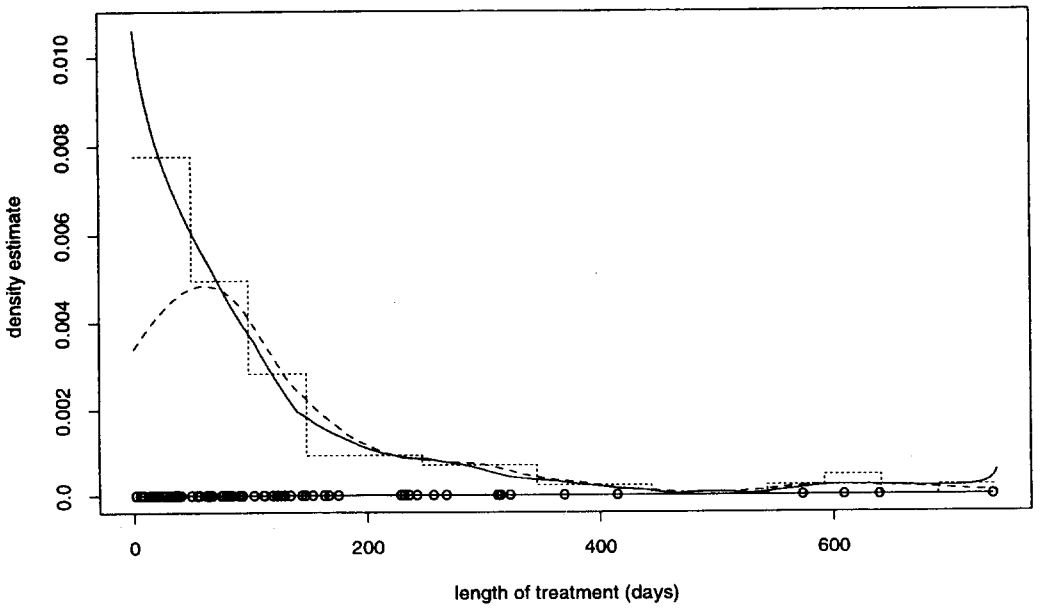


Figure 1. Ordinary kernel density estimate (dashed curve) and Poisson likelihood-based kernel density estimate (defined in (2.1); solid curve) with histogram (represented by dots) for the suicide data ($n = 86$, $h = 100$, Epanechnikov kernel).

The method starts with equal-length binning on the data space and recording the bin frequencies. The main idea is to note that the relative bin frequencies in density scale resembles a regression data in the sense that (bin frequency in density scale) = (density) + (error), and that the distribution of each bin count is approximately Poisson if the bin length reduces to zero at a proper speed. We point out that the operation of performing the likelihood-based local linear regression (Fan, Heckman and Wand 1995) with Poisson likelihood and the canonical link (log function in this case) successfully estimates the target density. In particular, the procedure always produces nonnegative estimates despite of using regression techniques, which may be tailored further to integrate to one by proper scaling. Furthermore, without using boundary kernels it accommodates edge effects particularly well as illustrated in Figure 1, and allows increased computational speed through equal-length binning. Details of the methodology will be given in Section 2, and illustrated numerically in Section 3 where the method is shown to outperform the optimal boundary kernel estimator (Müller 1991). Theory will be outlined in Section 4.

A different way of adapting local linear regression to density estimation has been introduced by Jones (1993) where the empirical distribution function or the Dirac delta function is used in place of response variable. Moreover, the above binning idea to get "regression-like" data has been suggested by Hall, Park and Turlach (1998). We note, however, that direct application of local linear regression to the Dirac delta function or to the binned data as suggested there would yield negative estimates.

2. METHODOLOGY

We observe independent and identically distributed X_j, for $1 \leq j \leq n$, and wish to estimate their common probability density function f. Although the same idea can apply to general supports, we consider in the sequel f with bounded support since the method is particularly useful in this case. Furthermore, we suppose for the sake of definiteness that the density is supported on $[0, 1]$. The method we propose performs equal-length binning, and then, treating each bin center and count as a data pair in regression smoothing, it carries out local linear regression based on an approximate likelihood of the bin counts.

Specifically, for a given $c > 1$ let $m = [n/c]$ be the number of bins of length $1/m$ where $[a]$ denotes the greatest integer not exceeding a, let N_i denote the number of X_j's falling in the i-th bin, and write x_i for the i-th bin center. We have now the data (x_i, N_i) for $1 \le i \le m$. Let $\eta(x) = \log\{nf(x)/m\}$. Treating N_i's as if they were independent and had Poisson distribution with mean $e^{\eta(x_i)}$, we have the approximate likelihood $\sum_{i=1}^m Q(e^{\eta(x_i)}; N_i)$ where $Q(v; y) = y \log v - v$. Local linear kernel regression based on this approximate likelihood suggests

$$\hat{f}_1(x) = me^{\hat{\beta}_0(x)}/n \tag{2.1}$$

where $\hat{\beta}_0(x)$ together with $\hat{\beta}_1(x)$ maximizes the locally kernel-weighted approximate likelihood

$$\sum_{i=1}^{m} Q(\exp\{\beta_0 + \beta_1(x_i - x)\}; N_i)K\{(x_i - x)/h\} \tag{2.2}$$

with respect to (β_0, β_1), K is a second-order kernel, and h a bandwidth. If all the bins with centers belonging to the smoothing interval $[x-h, x+h] \cap [0,1]$ are empty, then there exists no maximizer of (2.2). In this case we define $\hat{f}_1(x) = 0$. Note that $\hat{f}_1$ is nonnegative, while it may not integrate to one. A bona fide density may be obtained by proper scaling, i.e., dividing $\hat{f}_1$ by $\int \hat{f}_1(x)dx$. We call it $\hat{f}_2$.

Suppose that the kernel K is a probability density with support $[-1, 1]$. Let $\mathcal{I} \equiv \mathcal{I}(x,h) = [(x-1)/h, x/h] \cap [-1,1]$, $\nu_j \equiv \nu_j(\mathcal{I}) = \int_{\mathcal{I}} z^j K(z)dz$, and $K_1(z) = (\nu_0\nu_2 - \nu_1^2)^{-1}(\nu_2 - z\nu_1)K(z)$. It turns out that the function K_1 produces "actual" weights applied to N_i's. As h goes to zero, the interval $\mathcal{I}(x,h)$ reduces to $[-1,1]$ for an interior point $x \in [h, 1-h]$, and to $[-1, c_1]$ or $[-c_2, 1]$ for a boundary point of the form $c_1 h$ or $1 - c_2 h$ $(0 \le c_1, c_2 < 1)$. Write $\kappa_1 = \int_{\mathcal{I}} z^2 K_1(z)dz$, $\kappa_2 = \int_{\mathcal{I}} \{K_1(z)\}^2 dz$. Here we suppress dependence of κ_1 and κ_2 on x. In fact, κ_1 and κ_2 reduce to $\int_{-1}^1 z^2 K(z)dz$ and $\int_{-1}^1 K^2$ respectively when x is an interior point. Now, the estimator $\hat{f}_1$ has pointwise asymptotic bias and variance,

$$\begin{aligned} \text{asymp. bias}\{\hat{f}_1(x)\} &= \tfrac{1}{2}h^2\kappa_1 f(x)\{\log f(x)\}'' \\ \text{asymp. var}\{\hat{f}_1(x)\} &= (nh)^{-1}\kappa_2 f(x)\,. \end{aligned} \tag{2.3}$$

If we let $\kappa_0 = \{\int_{-1}^1 z^2 K(z)dz\}\{\int_0^1 (\log f(z))'' f(z)dz\}$, then the pointwise asymptotic bias of the scaled version $\hat{f}_2$ is given by

$$\text{asymp. bias}\{\hat{f}_2(x)\} = \tfrac{1}{2}h^2 f(x)\{\kappa_1(\log f(x))'' - \kappa_0\}\,. \tag{2.4}$$

The asymptotic variance of $\hat{f}_2$ turns out to be the same as that of $\hat{f}_1$. Details will be given in Section 4.

Some conclusions may be drawn from (2.3) and (2.4). Note particularly that for both estimators $O(h^2)$ bias is retained right up to the ends of $[0, 1]$ without boundary kernels being used. Furthermore, the kernel function need not be symmetric for (2.3) and (2.4). This is mainly because K_1 has zero first moment even though K is not symmetric. These features are not shared by other methods for density estimation. For proper comparison with the ordinary kernel estimator for an interior point, let us assume that K is symmetric (otherwise the ordinary kernel estimator would have $O(h)$ bias even for an interior point). Then, in comparison with the ordinary kernel estimator, we note that variance is unchanged, while bias is slightly changed with $f(\log f)''$ for $\hat{f}_1$ and $f\{(\log f)'' - \int_0^1 (\log f(z))'' f(z)dz\}$ for $\hat{f}_2$ now playing the role of f''. This means that the bias formula of $\hat{f}_1$ or $\hat{f}_2$ has additional term $-(f'/f)^2 f$ or $-(f'/f)^2 f - f\int_0^1 (\log f(z))'' f(z)dz$ respectively. In case of a truncated normal density f (see p_1 in Section 3 for example), the bias coefficient of $\hat{f}_2$ is zero since $(\log f)''$ is constant.

Note also that the variance components of $\hat{f}_1$ and $\hat{f}_2$ have not been inflated (to first order) by binning, i.e., using less number $m = [n/c] < n$ of data in the smoothing step. Moreover, the choice of c does not affect first-order properties of the estimator. It will influence only second-order properties.

Our approach may be applicable to estimating derivatives of the density function. In fact, $\hat{\beta}_1(x)$ is a consistent estimator of $\eta'(x)$, and so $f'(x)$ may be estimated by $m\beta_1(x)e^{\hat{\beta}_0(x)}/n$. Similarly, estimators of higher order derivatives could be obtained by fitting locally higher order polynomial (instead of linear) regression.

3. NUMERICAL EXPERIMENTS

We consider 500 pseudo samples of size 100. Those are generated from two population densities: (i) a truncated normal with density $p_1(x) = c_1\phi\{6(x - \frac{1}{2})\}I_{[0,1]}(x)$, (ii) a folded normal with density $p_2(x) = c_2[\phi\{4(x - 0.3)\} + \phi\{4(x + 0.3)\}]I_{[0,1]}(x)$ where ϕ is the standard normal density function and c_1, c_2 are the constants to make p_1, p_2 integrate to one. Note that p_1 has low density at both boundaries, while p_2 has high density at the left boundary.

We choose Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)I(|x| \leq 1)$. Our preliminary investigation shows that $\hat{f}_2$ dominates $\hat{f}_1$, which leads us to focus on the scaled version here. The number of bins is $m = 25$. We have tried other values of $m(20 \sim 30)$, but found that the performance of $\hat{f}_2$ is not sensitive to the choice of m in the range. We compare $\hat{f}_2$ with the scaled version of the so called optimal boundary kernel estimator (BKE) as proposed in Müller (1991).

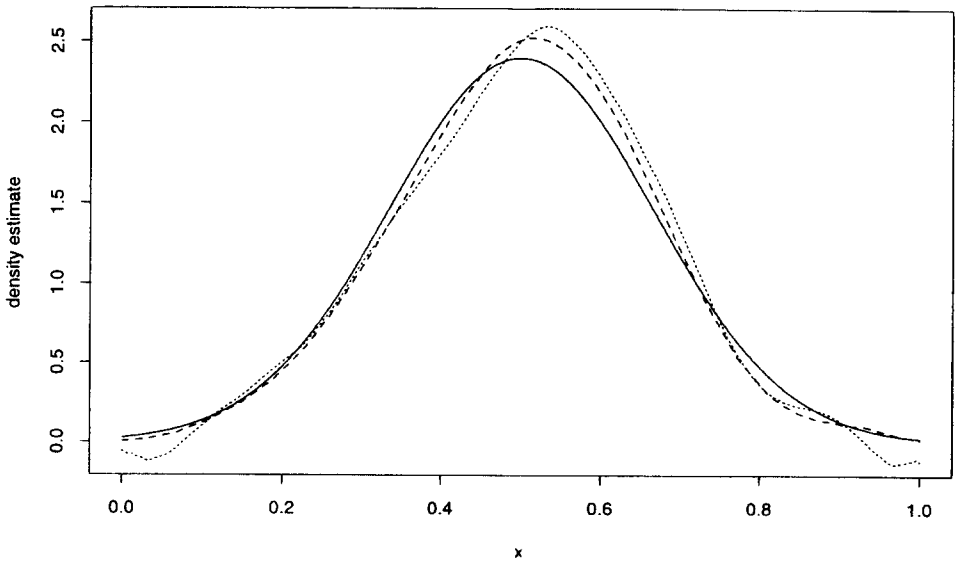
Figure 2 depicts, for a simulated dataset from (a) the truncated normal p_1 and (b) the folded normal p_2, the true density denoted by solid curve, BKE and $\hat{f}_2$ represented by dots and dashes respectively. The bandwidths used for BKE and $\hat{f}_2$ are those minimising mean integrated squared error (MISE). Figure 3 shows Monte Carlo estimates MISE as a function of bandwidth. Note that the minimum MISE of $\hat{f}_2$ is roughly 65% for p_1 and 90% for p_2 of that of BKE. In particular, the MISEs of $\hat{f}_2$ is less than those of BKE in the whole range of bandwidths for p_2, and in the range $h > .15$ for p_1. A first thought on the reason for $\hat{f}_2$ being inferior in the range of smaller bandwidths may be that it is due to presence of low density areas of p_1. However, we found that it is not true. In fact, reversely $\hat{f}_2$ is superior at the low density areas near $x = 0$ and 1, mostly due to smaller variance. This can be explained as follows: in a smoothing interval there are $2mh$ binned data for $\hat{f}_2$ regardless of how small the density is, while BKE has approximately $2nhf(x)$ data. This means that BKE is more sensitive to low density. The inferiority of $\hat{f}_2$ for p_1 in the range $h < .15$ is rather due to larger variance at high density areas. Comparing the variances at $x = 0$ and $x = \frac{1}{2}$, we found that the minimal variances of $\hat{f}_2$ and BKE in the range $h < .15$ are .00334 and .03042 respectively at $x = 0$, and .06232 and .04400 at $x = \frac{1}{2}$. The biases of both estimators are negligible in that range of bandwidths.

4. TECHNICAL ARGUMENTS

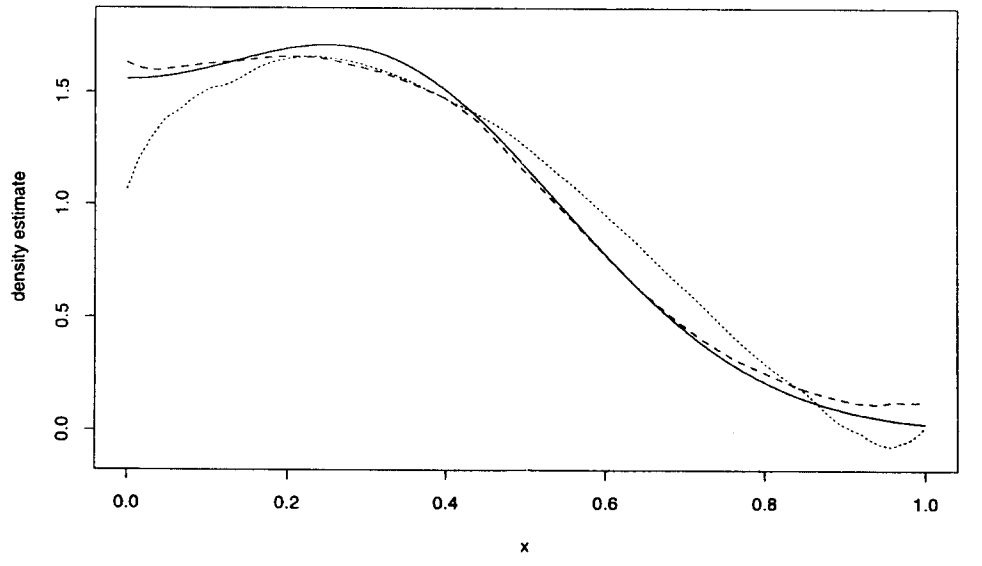
Write $b_i(x, h)$ for the asymptotic biases of $\hat{f}_i$ for $i = 1$ and 2 given in (2.3) and (2.4). We state a more explicit version of (2.3) and (2.4):

$$\sqrt{nh}(\kappa_2 f(x))^{-1/2}\{\hat{f}_i(x) - f(x) - b_i(x,h) + o(h^2)\} \to_d N(0,1)\,. \qquad (4.1)$$

The following regularity conditions are sufficient for (4.1):f has two continuous derivatives and is nonzero on its support $[0, 1]$; K is a probability density

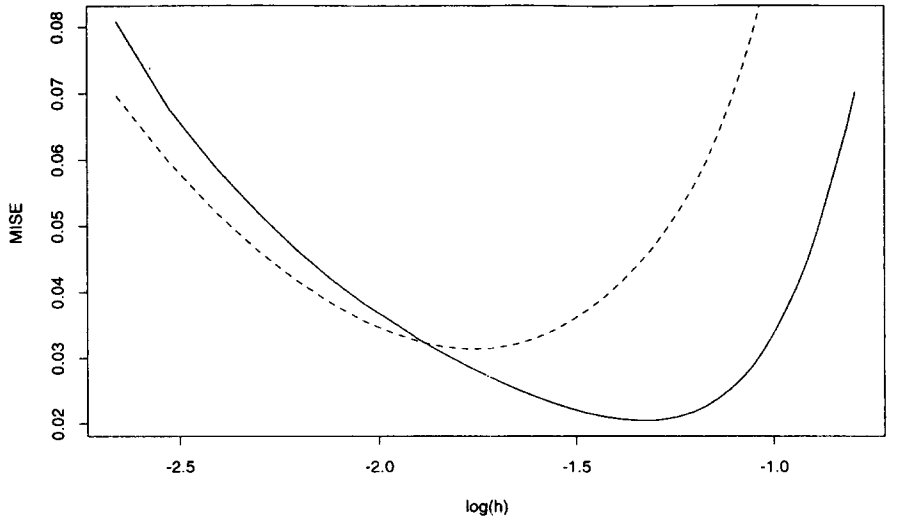


(a)

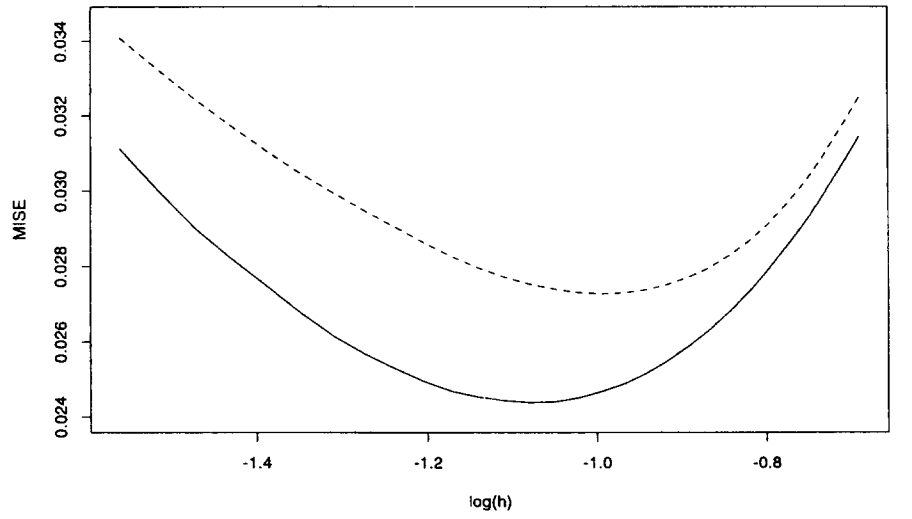


(b)

Figure 2. Typical estimates representing BKE (dotted curve) and $\hat{f}_2$ (dashed curve) with true density (solid curve) for (a) the truncated normal p_1 and (b) the folded normal p_2.



(a)



(b)

Figure 3. Mean integrated squared errors of BKE (dashed) and $\hat{f}_2$ (solid) as a function of bandwidth based on 500 Monte Carlo samples of size $n = 100$ generated from (a) the truncated normal p_1 and (b) the folded normal p_2.

supported on $[-1,1]$; n/m converges to some positive constant; and $h = h_n \to 0$, $nh^3 \to \infty$ as $n \to \infty$. We will outline the proof of (4.1) for $\hat{f}_1$ first. Put $\bar{\eta}(x,u) = \eta(x)+\eta'(x)(u-x)$, and $K_2(z) = (\nu_0\nu_2 - \nu_1^2)^{-1}(\nu_0 z - \nu_1)K(z)$. Write $\tilde{K}(z) = (K_1(z), K_2(z))^T$, $\alpha_n(x) = (m/\{nf(x)\}, -hm\kappa_1 f'(x)/\{nf(x)^2\})^T$ and let

$$W_n(x) = (mh)^{-1/2} \sum_{i=1}^{m} \{N_i - e^{\bar{\eta}(x,x_i)}\}\tilde{K}\{(x_i - x)/h\}\,.$$

In the sequel we will suppress x in all the notations. Derivation of (4.1) is based on the following stochastic representation:

$$\sqrt{mh}(\hat{\beta}_0 - \eta) = \alpha_n^T W_n + o_p(h)\,. \tag{4.2}$$

A proof of (4.2) follows lines similar to those in Fan, Heckman and Wand (1995). Note that its derivation is rather simplified since $(\partial^k/\partial x^k)Q(e^x; y)$ does not depend on the second argument y for $k \geq 2$. The mean and variance of $\alpha_n^T W_n$ are now given by

$$\mathrm{E}\,(\alpha_n^T W_n)\tfrac{1}{2}m^{1/2}h^{5/2}\kappa_1(\log f)'' + o(m^{1/2}h^{5/2})\,,$$

$$\mathrm{var}\,(\alpha_n^T W_n) = m\kappa_2/(nf) + o(1)\,. \tag{4.3}$$

The results (4.3) can be obtained by using the facts $\mathrm{E}\,(N_i) = nf(x_i)/m + O(1/n)$, $\mathrm{cov}\,(N_i, N_j) = nf(x_i)/m + O(1/m)$ for $i = j$, and $-nf(x_i)f(x_j)/m^2 + O(1/m^2)$ for $i \neq j$. Note also that in (4.3) there are integral approximation errors of order $O(m^{-1/2}h^{-1/2})$, which have been included in the remainders since $nh^3 \to \infty$. Finally, $\mathrm{cov}\,(W_n)^{-1/2}\{W_n - \mathrm{E}\,(W_n)\}$ converges in law to the bivariate standard normal distribution. This can be verified easily by using the moment generating function of multinomial distribution. This together with (4.2) and (4.3) implies (4.1).

Now for $\hat{f}_2$ it follows that

$$\hat{f}_2 - f = (\hat{f}_1 - f) - f\int_0^1 (\hat{f}_1 - f) + o_p(h^2 + n^{-1/2}h^{-1/2})\,. \tag{4.4}$$

The integral term of (4.4) has negligible variance. This follows from the fact that $\mathrm{var}\,(\int_0^1 \alpha_n^T W_n f) = O(h)$. This concludes the proof.

REFERENCES

(1) COPAS, J.B. AND FRYER, M.J. (1980). Density estimation and suicide risks in psychiatric treatment. *Journal of the Royal Statistical Society*, Series A **143**, 167-176.

(2) FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics* **21**, 196-216.

(3) FAN, J. AND GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20**, 2008-2036.

(4) FAN, J., HECKMAN, N.E. AND WAND, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141–150.

(5) HALL, P., PARK, B.U. AND TURLACH, B. (1998). A note on design transformation and binning in nonparametric curve estimation. *Biometrika*, to appear.

(6) JONES, M.C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing* **3**, 135-146.

(7) MÜLLER, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521-530.

(8) RICE, J.A. (1984). Boundary modification for kernel regression. *Communications in Statistics - Theory and Methods* **13**, 893-900.

(9) SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.