

Fast Implementation of Density-Weighted Average Derivative Estimation

Berwin A. Turlach

C.O.R.E & Institut de Statistique
34, Voie du Roman Pays
1348 Louvain-la-Neuve, Belgium

Abstract

Given random variables $X \in \mathbb{R}^d$ and Y such that $E[Y|X = x] = m(x)$, the average derivative δ_0 is defined as $\delta_0 = E[\nabla m(X)]$, i.e., as the expected value of the gradient of the regression function. Average derivative estimation has several applications in econometric theory (Stoker, 1992) and thus it is crucial to have a fast implementation of this estimator for practical purposes.

We present such an implementation for a variation known as *density-weighted average derivative estimation*. This algorithm is based on the ideas of binning or **W**eighted **A**veraging of **R**ounded **P**oints (WARPiNg). The basic idea of this method is to discretize the original data into a d -variate histogram and to replace in the non-parametric smoothing steps the actual observations by the appropriate bincenters. The non-parametric smoothing steps become thus a (multi-dimensional) convolution between the (discretized) data and the (discretized) smoothing kernel.

A Monte-Carlo study demonstrates that with this binned implementation substantial reduction in computing time can be achieved. But it will also become clear that in higher dimension the choice of **how to bin** is crucial.

1 Introduction

Average derivative estimation tries to estimate the mean slope of the conditional mean of the response variable, i.e., given a response variable Y , whose expectation is assumed to depend on a d -dimensional variable X via a smooth function m , the aim of average derivative estimation is to estimate the average slope of this function. In other words, if

$$E[Y|X = x] = m(x)$$

and ∇ denotes the gradient of partial derivatives with

respect to the coordinates of X , the aim is to estimate

$$\delta_0 = E[\nabla m(X)] \quad (1)$$

respectively a weighted version

$$\delta_w = E[\nabla m(X)w(X)] \quad (2)$$

where $w(\bullet)$ is a non-negative weight function. If we choose as weight function $w(x) \equiv f(x)$, the marginal density of X , our estimand becomes:

$$\begin{aligned} \delta &= E[\nabla m(X)f(X)] \\ &= -2E[Y\nabla f(X)] \end{aligned} \quad (3)$$

Where (3) follows by partial integration. The problem of estimating the *density-weighted average derivative*, as given by (3), was studied by Powell, Stock and Stoker (1989).

Average derivative estimation can be used in many econometric models (Stoker, 1992; Härdle, Hildenbrand and Jerison, 1991). As one example, we want to mention *single-index* models (also called *one-term projection pursuit* models). In these models the regression function m has the form

$$m(x) = g(x^T \beta), \quad (4)$$

where g is an unknown univariate function and β is a d -dimensional (projection) vector. Stoker (1986) gives an extensive discussion and motivation for models of the form (4). The semiparametric model (4) covers a broad range of important parametric models such as probit and logit models, censored regression, Tobit models etc.

It is easy to see, that in this case we have

$$\nabla m(x) = g'(x^T \beta)\beta$$

and thus

$$\delta_0 = E[g'(X^T \beta)]\beta \quad \text{and} \quad \delta_w = E[g'(X^T \beta)w(X)]\beta.$$

This means that (weighted) average derivative estimation allows us to estimate the unknown projection β up to a scale constant. This is in fact the best we can do in the semiparametric single-index model given by (4). If the pair (g, β) fulfills model (4) then for any $c \in \mathbb{R}$, $c \neq 0$, the pair $(\tilde{g}, \tilde{\beta})$ with

$$\tilde{g}(\bullet) = g(\bullet/c) \quad \text{and} \quad \tilde{\beta} = c\beta$$

does so too.

The rest of this article is structured as follows, Section 2 will describe the density-weighted average derivative estimator as proposed by Powell et al. (1989). In Section 3 we will propose how to implement this estimator using binning ideas and to achieve thus considerable run-time gains. Finally in Section 4 we will discuss some further points related to the binning method.

2 Direct implementation

2.1 Estimator for δ

To estimate the density-weighted average derivative δ , Powell et al. (1989) propose to estimate the gradient of the marginal density of the X variables nonparametrically at each observation point by, say, $\widehat{\nabla}f(x_i)$. Their estimator for δ is

$$\hat{\delta} = -\frac{2}{n} \sum_{i=1}^n y_i \widehat{\nabla}f(x_i) \quad (5)$$

which can be motivated as a method of moment estimator in which the unknown function ∇f is replaced by a nonparametric estimate of it.

To estimate the gradient of f nonparametrically, Powell et al. (1989) use the gradient of a multivariate kernel density estimator (Silverman, 1986; Scott 1992). Given a d -variate kernel \mathcal{K} (think of \mathcal{K} as a d -variate density function) and a $d \times d$ positive definite matrix H of smoothing parameters a nonparametric estimate of the marginal density f at a point $x \in \mathbb{R}^d$ would be

$$\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\det(H)} \mathcal{K}(H^{-1}(x - x_i)). \quad (6)$$

For numerical ease, a common choice is to take \mathcal{K} as a product of d univariate kernels K , and to reduce H to a diagonal matrix, so that we have only a d -dimensional vector h of smoothing parameters. Wand and Jones (1993) discuss for the two-dimensional case the implications of this simplification. With this choices (6) simplifies to

$$\hat{f}_h(x) = \frac{1}{nh_1 \dots h_d} \sum_{i=1}^n \prod_{k=1}^d K\left(\frac{x_k - x_{ik}}{h_k}\right). \quad (7)$$

where $x = (x_1, \dots, x_d)^T$ and $x_j = (x_{j1}, \dots, x_{jd})^T$.

Powell et al. (1989) do not use the nonparametric density estimator given in (7) directly, but a *leave-one-out* version of it. (For this reason the estimator $\hat{\delta}$ has a U -statistic structure and can be easily analyzed.) Thus to estimate the marginal density f at the observation x_i , they drop x_i from the sample and calculate $\hat{f}_h(x_i)$ from the remaining sample (of size $n-1$). As a further simplification they use only *one* bandwidth for all dimensions. So the estimator $\widehat{\nabla}f(x_i)$ which they use in (5) is:

$$\begin{aligned} \widehat{\nabla}f(x_i) &= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{1}{h^{d+1}} \mathcal{K}'\left(\frac{x_{ik} - x_{jk}}{h}\right) \\ &= \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{pmatrix} \left(\frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \prod_{k=1}^d K_h(x_{ik} - x_{jk}) \right). \end{aligned} \quad (8)$$

with $K_h(u) = K(u/h)/h$.

2.2 Asymptotic properties

Powell et al. (1989) showed that under certain regularity conditions and a suitable choice for K and the rate with which h tends to zero, the estimator $\hat{\delta}$ given in (5) is consistent and has an asymptotic normal distribution. More specifically they proved that

$$\sqrt{n} (\hat{\delta} - \delta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

where

$$\begin{aligned} \Sigma &= 4\mathbb{E}[r(X, Y)r(X, Y)^T] - 4\delta\delta^T, \\ r(x, y) &= f(x)\nabla m(x) - \{y - m(x)\}\nabla f(x). \end{aligned}$$

2.3 Estimator for the variance

To estimate the asymptotic variance Σ of $\hat{\delta}$ Powell et al. (1989) propose to estimate $r(x_i, y_i)$ by:

$$\hat{r}(x_i, y_i) = -\frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{1}{h^{d+1}} \mathcal{K}'\left(\frac{x_i - x_j}{h}\right) (y_i - y_j) \quad (9)$$

and thus Σ by:

$$\hat{\Sigma} = 4 \frac{\sum_{i=1}^n \hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T}{n} - 4\hat{\delta}\hat{\delta}^T. \quad (10)$$

In the next section we will discuss how fast implementations for $\hat{\delta}$ and $\hat{\Sigma}$ can be obtained by using binning techniques.

3 Binned implementation

3.1 Basic idea

The basic idea of binning methods is to replace each observation of x_i by the nearest point b_z from a regular spaced grid. To fix ideas consider kernel density estimation in the one-dimensional case,

$$\hat{f}_h(x_i) = \frac{1}{n} \sum_{j=1}^n K_h(x_i - x_j), \quad (11)$$

and take the regular grid $\{b_z : b_z = z\Delta, z \in \mathbb{Z}\}$ where Δ is a fixed constant, the *binwidth*. Replacing now each x_i in (11) by the nearest b_z , we see that we have to evaluate the kernel K only at integer multiple of Δ/h :

$$w_l = \frac{1}{h} K\left(\frac{\Delta}{h}l\right), \quad l = -L, \dots, L \quad (12)$$

Here L is chosen such that $\Delta L/h \approx 1$ if K has compact support on $[-1, 1]$ (if K is the Gaussian kernel, i.e., the kernel has no compact support, Wand (1993) recommends $\Delta L/h \approx 4$). If we denote further by n_z the number of observations x_i which have b_z as their nearest point in the grid, we see that we can approximate (11) by (let b_z be the point nearest to x_i):

$$\begin{aligned} \hat{f}_h(x_i) &= \frac{1}{n} \sum_{j=1}^n K_h(x_i - x_j) \\ &\approx \frac{1}{n} \sum_{j=1}^n w_{z-l_j}, \quad b_{l_j} \text{ is nearest to } x_j \\ &= \frac{1}{n} \sum_{l=-L}^L w_{z-l} n_l. \end{aligned}$$

The last formula is a discrete convolution between the vector of weights (the discretized kernel) and the vector of *bincounts* n_z (the discretized data).

Silverman (1982) uses a fast fourier transformation to calculate this discrete convolution. Another algorithm which does not use the fast fourier transform is given in Scott (1985) (see also Härdle and Scott, 1992; Härdle, 1991). Fan and Marron (1994) describe how to use these ideas for other nonparametric curve smoothers.

Fan and Marron (1994) also quantify the run-time gains achievable using these ideas. These run-time gains are mainly due to two facts. First we have much less kernel evaluations, in fact we have to evaluate the kernel only once on a finite grid of points. Secondly, once the data is discretized the nonparametric curve smoother is estimated at the grid points b_z and not at the original observations x_i . Usually the number of grid points

at which the smoother is evaluated is (much) smaller than n . The estimate at an original observation x_i is either taken as the estimate at the nearest b_z or obtained by linear interpolation between the estimates of the two nearest grid points (Jones, 1989).

3.2 Application to $\hat{\delta}$

The ideas presented in Section 3.1 above are readily extendable to the multivariate case (Wand, 1993) and to the estimator $\hat{\delta}$.

Again we define a (multivariate) grid of equidistant points $b_z \in \mathbb{R}^d$ and replace $x_i \in \mathbb{R}^d$ by the nearest b_z . To fix ideas let $\Delta = (\Delta_1, \dots, \Delta_d)^T$ be a fixed d -dimensional vector and define b_z by

$$b_z = z\Delta = (z_1\Delta_1, \dots, z_d\Delta_d)^T$$

for each multi-index $z = (z_1, \dots, z_d)^T \in \mathbb{Z}^d$. Note the pointwise multiplication of the vectors z and Δ above. In the rest of this article, if not indicated differently, we mean this kind of pointwise vector multiplication rather than the standard matrix multiplication when we multiply two vectors.

For each $z \in \mathbb{Z}^d$, let again n_z denote the number of observed x_i for which b_z is the nearest grid point. For a binned implementation of the estimator $\widehat{\nabla}f$ we also need to discretize the derivative of the kernel K :

$$\tilde{w}_{lj} = \frac{1}{h^2} K'\left(\frac{\Delta_j}{h}l\right), \quad l = -L_j, \dots, L_j, \quad j = 1, \dots, d \quad (13)$$

and define w_{lj} analogous to (12) by replacing Δ by Δ_j . If we define now for each multi-index $l = (l_1, \dots, l_d)^T \in \mathbb{Z}^d$ the corresponding weight $w'_l \in \mathbb{R}^d$ by:

$$w'_l = \begin{pmatrix} \tilde{w}_{l_1 1} w_{l_2 2} \cdots w_{l_d d} \\ w_{l_1 1} \tilde{w}_{l_2 2} \cdots w_{l_d d} \\ \vdots \\ w_{l_1 1} w_{l_2 2} \cdots \tilde{w}_{l_d d} \end{pmatrix}$$

we see that analogous to the example in Section 3.1 a binned version of the estimator $\widehat{\nabla}f$ is:

$$\widehat{\nabla}f(b_z) = \frac{1}{n-1} \sum_{l=-L}^L w'_{z-l} n_l. \quad (14)$$

Note that the sum in (14) is actually a sum over d indices l_1, \dots, l_d , each l_j taking values from $-L_j$ to L_j , $j = 1, \dots, d$. Also, the multi-index $z-l$ in (14) is $z-l = (z_1 - l_1, \dots, z_d - l_d)^T$.

Thus a binned version of the density-weighted average derivative δ is:

$$\hat{\delta} = -\frac{2}{n} \sum_{z \in \mathbb{Z}^d} n_z \bar{y}_z \widehat{\nabla}f(b_z) \quad (15)$$

where \bar{y}_z is the average over all observation y_i such that b_z is the nearest grid point to the corresponding x_i . Note that the summation in (15) is actually only over all $z \in \mathbb{Z}^d$ such that $n_z \neq 0$ and is not an infinite sum. Furthermore, if we compare (5) with (15) we see that the only approximation error we do is due to replacing $\widehat{\nabla}f(x_i)$ by $\widehat{\nabla}f(b_z)$. With respect to the y we “keep the full resolution”.

3.3 Application to $\hat{\Sigma}$

In this section we will discuss the implementation of a binned estimator for the asymptotic variance Σ given in Section 2.2. A naive way of implementing such an estimator would be to plug into (10) a binned estimate, say, $\hat{r}(b_z)$ for $\hat{r}(x_i, y_i)$, given in (9), to obtain:

$$\hat{\Sigma} = 4 \frac{\sum_{z \in \mathbb{Z}^d} \hat{r}(b_z) \hat{r}(b_z)^T}{n} - 4 \hat{\delta} \hat{\delta}^T \quad (16)$$

with $\hat{\delta}$ from (15). The binned estimate $\hat{r}(b_z)$ is easily derived in the same way as demonstrated in Section 3.1. Let b_z be the grid point nearest to x_i , then we have:

$$\begin{aligned} \hat{r}(x_i, y_i) &= \\ &= -\frac{1}{n-1} \sum_{j=1}^n \frac{1}{h^{d+1}} \mathcal{K}'\left(\frac{x_i - x_j}{h}\right) (y_i - y_j) \\ &\approx -\frac{1}{n-1} \sum_{j=1}^n w'_{z-l_j} (y_i - y_j), b_{l_j} \text{ is nearest to } x_j \\ &= -\frac{1}{n-1} \sum_{l=-L}^L w'_{z-l} n_l (y_i - \bar{y}_l) = \hat{r}(b_z, y_i) \\ &\approx -\frac{1}{n-1} \sum_{l=-L}^L w'_{z-l} n_l (\bar{y}_z - \bar{y}_l) = \hat{r}(b_z) \end{aligned}$$

Note that the only approximation error in $\hat{r}(b_z, y_i)$ is due to replacing the x_i by the grid point b_z . Thus for $\hat{r}(b_z, y_i)$ we have still the full resolution in the y -direction. Only if we go to $\hat{r}(b_z)$ we make an approximation error in that direction too. The motivation for this approximation is, that if several x_i exist which have b_z as nearest grid point then we should average over the corresponding $\hat{r}(b_z, y_i)$ to get a unique estimate $\hat{r}(b_z)$ at b_z .

However, the binned implementation which we get if we insert $\hat{r}(b_z)$ in (16) does not work. The reason for this is explained and graphically illustrated in Proença and Turlach (1994). On one side we make an approximation error in the y -direction by going from $\hat{r}(b_z, y_i)$ to $\hat{r}(b_z)$. On the other side we want to approximate $\hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T$ which involves a squared term

in y . Thus we have to take into account what Proença and Turlach (1994) call the *within-bin-variability* of y . This means that we can not find a binned estimator for $\hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T$ by finding one just for $\hat{r}(x_i, y_i)$, but that we really have to consider this product directly. Hence a “correct” binned estimator can be found by observing that:

$$\begin{aligned} \hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T &\approx \\ &\approx \hat{r}(b_z, y_i) \hat{r}(b_z, y_i)^T \\ &= \left(\frac{1}{n-1}\right)^2 \sum_{l=-L}^L \sum_{l'=-L}^L w'_{z-l} w'_{z-l'}^T \times \\ &\quad n_l (y_i - \bar{y}_l) n_{l'} (y_i - \bar{y}_{l'}) \\ &= \left(\frac{1}{n-1}\right)^2 \sum_{l, l'=-L}^L \left\{ w'_{z-l} w'_{z-l'}^T \times \right. \\ &\quad \left. n_l (y_i - \bar{y}_z + \bar{y}_z - \bar{y}_l) n_{l'} (y_i - \bar{y}_z + \bar{y}_z - \bar{y}_{l'}) \right\} \\ &= \hat{r}(b_z) \hat{r}(b_z)^T + \\ &\quad + \left(\frac{1}{n-1}\right)^2 \sum_{l, l'=-L}^L \left\{ w'_{z-l} w'_{z-l'}^T \times \right. \\ &\quad \left. n_l n_{l'} (y_i - \bar{y}_z) (2\bar{y}_z - y_l - y_{l'}) \right\} \\ &\quad + \left(\frac{1}{n-1}\right)^2 \sum_{l, l'=-L}^L w'_{z-l} w'_{z-l'}^T n_l n_{l'} (y_i - \bar{y}_z)^2 \end{aligned}$$

And thus the sum $\sum_{i=1}^n \hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T$ can be approximated as:

$$\begin{aligned} \sum_{i=1}^n \hat{r}(x_i, y_i) \hat{r}(x_i, y_i)^T &\approx \\ &\approx \sum_{z \in \mathbb{Z}^d} \sum_{i=1}^n \hat{r}(b_z, y_i) \hat{r}(b_z, y_i)^T \\ &= \sum_{z \in \mathbb{Z}^d} \left\{ \hat{r}(b_z) \hat{r}(b_z)^T + \right. \\ &\quad \left. \left(\frac{1}{n-1}\right)^2 \sum_{l, l'=-L}^L w_{z-l} w_{z-l'}^T n_l n_{l'} n_z (\bar{y}_z^2 - \bar{y}_z^2) \right\} \\ &= \sum_{z \in \mathbb{Z}^d} \left\{ \hat{r}(b_z) \hat{r}(b_z)^T + n_z (\bar{y}_z^2 - \bar{y}_z^2) \widehat{\nabla}f(b_z) \widehat{\nabla}f(b_z)^T \right\} \\ &= \widehat{rr}^T \end{aligned}$$

Note that because of the summation over i the term which includes $(y_i - \bar{y}_z)(2\bar{y}_z - \bar{y}_l - \bar{y}_{l'})$ drops out, i.e., the sum is zero. Also, \bar{y}_z^2 denotes the square of \bar{y}_z and \bar{y}_z^2 denotes the mean of all y_i^2 such that x_i has b_z as nearest grid point. This term, namely $n_z (\bar{y}_z^2 - \bar{y}_z^2)$, measures the

variability of Y around the grid point b_z . This term is obtained by expanding $(y_i - \bar{y}_z)^2$ and summing over i . Note that if we choose Δ so small, that each grid point b_z has at most one observation x_i for which it is the nearest point then all of these within-bin-variability terms vanish and the binned estimator given in (16) would be correct.

However, in general we have to take these terms into account. Thus a “correct” binned estimator for the variance matrix is given by

$$\hat{\Sigma} = 4 \frac{\widehat{rr^T}}{n} - 4\hat{\delta}\hat{\delta}^T$$

with $\hat{\delta}$ from (15).

4 Closing remarks

In the previous section we demonstrated how the simple and intuitive basic binning idea can be applied to the density-weighted average derivative estimator $\hat{\delta}$ and the estimator of the asymptotic covariance matrix $\hat{\Sigma}$. Some questions still remain which we would like to address here.

From (14) we see that $\widehat{\nabla}f(b_z)$ is a discrete convolution, the same is true for $\hat{r}(b_z)$ and $\widehat{rr^T}$. How should we calculate this discrete convolution? As mentioned above Silverman (1982) and Wand (1993) use a fast fourier transformation. However, this method is inappropriate in our situation since we are only interested to calculate these estimates at the points b_z which have some observation close enough to them, i.e., for which $n_z \neq 0$. But a fast fourier transformation method would calculate these estimates at *all* grid points b_z . Just imagine the case where we have a two-dimensional X -variable and we choose our grid such that we have 100 different grid points in each dimension. The complete grid will have 10.000 points b_z . In this case a fast fourier transform method would calculate $\widehat{\nabla}f(b_z), \dots$ at all these grid points. Clearly this involves many unnecessary calculations if the sample size is not too big.

The fast fourier transform approach is feasible if we need estimates at all grid points for example if we want to make a plot. But it is also not clear if the fast fourier transform is the fastest method in such a case. Fan and Marron (1994) find that this approach is not the fastest for the one-dimensional case whereas Wand (1993) favors the fast fourier transform in the two-dimensional case. Scott (1992) describes alternative algorithms which do not use a fast fourier transform. These algorithm step through all grid points b_z with $n_z \neq 0$ and just do the necessary calculations at these points *and* in the neighborhood of b_z (as defined by the L_j), i.e., also these algorithms calculate the estimates on the whole grid. For

the discrete convolution necessary here we recommend to use specialized versions of the algorithms of Scott (1994) which step through all grid points b_z with $n_z \neq 0$ and do the necessary calculations *only* at these points.

Closely related with the question “How to perform the discrete convolution?” is the question “How shall one discretize the data?”. Until now we always used a kind of “histogram” binning in which n_z was integer and each observation was shifted to (replaced by) the nearest grid point b_z . For the one-dimensional density estimation Jones and Lotwick (1984) proposed an alternative called “linear” binning. In this variation the n_z are no longer integer and each observation is distributed onto the *two* nearest grid points. Hall and Wand (1993) propose further variations for the binning procedure and quantify the error which is introduced by using binning techniques (see also González-Manteiga, Sánchez-Sellero and Wand, 1994).

But the use of such techniques in a higher-dimensional setting is problematic. A binning technique like “linear” binning which distributes each observation in one-dimension on two grid points, will distribute each observation in d -dimension onto 2^d grid points. This could have the effect that we have more grid points b_z with $n_z \neq 0$ than observations! Take for example a two-dimensional standard normal variable and use linear binning with a grid where $\Delta = (0.03, 0.03)^T$. If the sample size is $n = 250$ we have on the average 950 grid points b_z at which $n_z \neq 0$. The result of this is that, even if we use the algorithms described above for the discrete convolution, the binned implementation using “linear” binning is *slower* than the direct implementation.

This was verified in a Monte-Carlo study with a bivariate X -variable (and Y generated according to a linear model and a probit model). Using the adapted algorithms from Scott (1992) for the discrete convolution and “linear” binning hardly no run-time gains were observed and for a grid with small Δ the direct implementation was even faster. If “histogram” binning was used, however, we observed run-time gains of a factor 10 over the direct implementation.

Thus we recommend to use “histogram” binning and the (adapted) algorithms of Scott (1993) for functional estimation in higher dimensions.

Acknowledgments

I would like to thank Steve Marron and Matt Wand for the discussions we had about binning techniques. Likewise I would like to thank Tom Stoker and Jim Powell for discussions about (weighted) average derivative estimation. This research was supported by the EC grant B/SPES 915011.

References

- J. Fan and J. S. Marron. Fast implementations of non-parametric curve estimators. *Journal of Computational and Graphical Statistics*, 3(1):35–56, 1994.
- W. González-Manteiga, C. Sánchez-Sellero, and M. P. Wand. Accuracy of binned kernel functional estimators. unpublished manuscript, 1994.
- P. Hall and M. P. Wand. On the accuracy of binned kernel density estimators. Working Paper 93–003, The University of New South Wales, Australian Graduate School of Management, PO Box 1, Kensington NSW 2033, Australia, 1993.
- W. Härdle. *Smoothing Techniques, With Implementations in S*. Springer, New York, 1991.
- W. Härdle, W. Hildenbrand, and M. Jerison. Empirical evidence on the law of demand. *Econometrica*, 59(6):1525–1549, 1991.
- W. Härdle and D. W. Scott. Smoothing by weighted averaging of rounded points. *Computational Statistics*, 7:97–128, 1992.
- M. C. Jones. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84(407):733–741, 1989.
- M. C. Jones and H. W. Lotwick. A remark on algorithm AS176: Kernel density estimation using the fast fourier transform (Remark ASR50). *Applied Statistics*, 33:120–122, 1984.
- J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.
- I. M. Proença and B. A. Turlach. Fast implementation of bandwidth selectors. unpublished manuscript, 1994.
- D. W. Scott. Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *Annals of Statistics*, 13(3):1024–1040, 1985.
- D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, Chichester, 1992.
- B. W. Silverman. Kernel density estimation using the fast fourier transform. Statistical algorithm AS 176. *Applied Statistics*, 31:93–97, 1982.
- B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1986.
- T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481, 1986.
- T. M. Stoker. *Lectures on Semiparametric Econometrics*. Center for Operation Research and Econometrics, Université Catholique de Louvain, Voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium, 1992.
- M. P. Wand. Fast computation of multivariate kernel estimators. Working Paper 93–007, The University of New South Wales, Australian Graduate School of Management, PO Box 1, Kensington NSW 2033, Australia, 1993.
- M. P. Wand and M. C. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88:520–529, 1993.