# Fast Computation of Auxiliary Quantities in Local Polynomial Regression

B. A. TURLACHand M. P. WAND

We investigate the extension of binning methodology to fast computation of several auxiliary quantities that arise in local polynomial smoothing. Examples include degrees of freedom measures, cross-validation functions, variance estimates, and exact measures of error. It is shown that the computational effort required for such approximations is of the same order of magnitude as that required for a binned local polynomial smooth.

**Key Words:** Binning; Cross-validation; Error degrees of freedom; Kernel estimator; Linear smoother; Mean average squared error; Smoother matrix; Standard error.

## 1. INTRODUCTION

Fast computational methods for local polynomial kernel regression have received considerable attention in the recent literature. Witness the work of Cleveland and Grosse (1991), Härdle and Scott (1992), Fan and Marron (1994), Seifert, Brockmann, Engel, and Gasser (1994), Loader (1994), and Wand (1994).

In each of these articles the main focus has been on fast computation of the curve of interest, usually a regression function estimate or "scatterplot smoother." In nonparametric regression analysis, however, there are often several auxiliary quantities that need to be computed. Examples are:

- the error degrees of freedom measure for diagnosis and comparison of different smoothers (e.g., Hastie and Tibshirani 1990);
- the cross-validation criterion function for automatic choice of the smoothing parameter;
- estimates of the error variance; and
- estimates of standard errors in partially linear models.

If the data set contains $n$ observations, then each of these auxiliary quantities requires $O(n^2)$ operations for exact computation. This can mean an enormous computational cost

B.A. Turlach is Research Associate, Centre for Mathematics and its Applications, Australian National University, Canberra, 2601, Australia; e-mail: berwin@alphasun.anu.edu.au. M. P. Wand is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, Sydney, 2052, Australia; e-mail: wand@agsm.unsw.edu.au.

*Journal of Computational and Graphical Statistics, Volume 5, Number 4, Pages 337–350*

for large data sets and has resulted in—most notably in the smoothing spline literature—the development of approximations to some of these quantities. For example, Hastie and Tibshirani (1990) devoted an appendix of their monograph to the development of an approximation to the error degrees of freedom that can be computed in $O(n)$ operations.

The use of binning to speed up computation of kernel estimators was first developed in the density estimation context by Scott (1981, 1985) and Silverman (1982). The close similarity between their two approaches is not generally recognized. They essentially differ only in the kernel used and the method by which the discrete convolutions are computed, with Silverman (1982) using the fast Fourier transform for this task. Fan and Marron (1994) extended binning ideas to local polynomial kernel estimators. The computational speed of this approach is apparent in the fact that the binned approximation to a function estimate can be computed on a grid of size $M$ using $O(M)$ kernel evaluations. Moreover, it can be shown that the binned approximation can be made arbitrarily accurate by increasing the value of $M$ (e.g., Hall and Wand in press).

The purpose of this article is to show how these ideas can be applied to fast computation of the auxiliary quantities of the type described previously. In each case it is seen that use of the binning principle reduces the computational labor to that of computing a regression estimate over a grid, and therefore also requires $O(M)$ evaluations.

The local polynomial smoother is described in Section 2. In Section 3 we briefly describe the binning principle and, in Section 4, some key results for handling common forms are highlighted. Section 5 illustrates how the binning principle can be used for fast computation of a variety of auxiliary quantities. Section 6 discusses some generalizations and Section 7 contains an assessment of the accuracy of the binned approximations of the preceding sections.

## 2. LOCAL POLYNOMIAL SMOOTHERS

Each of the auxiliary quantities that we consider can be defined for general linear smoothers (e.g., Hastie and Tibshirani 1990) so we will start at this level of generality.

A *smooth* of the regression data set $(X_1, Y_1), \ldots, (X_n, Y_n)$ is defined to be

$$\widehat{m} = [\widehat{m}(X_1), \ldots, \widehat{m}(X_n)]^T,$$

where $\widehat{m}(x)$ denotes the value of a regression estimate, or scatterplot smoother, at the point $x$. Common methods for obtaining $\widehat{m}(x)$ are smoothing splines, regression splines, local polynomials, kernel estimators, and wavelets. If there exists an $n \times n$ matrix $S$ such that

$$\widehat{m} = SY,$$

where $Y$ denotes the vector of the $Y_i$'s, then $\widehat{m}$ is called a *linear smooth* of the data. We usually refer to $S$ as the *smoother matrix.* There are a number of important auxiliary quantities that can be defined for general linear smoothers. For example, Hastie and Tibshirani (1990) defined the *error degrees of freedom* of a smoother as

$$\mathrm{df}^{\mathrm{err}} = n - 2\mathrm{tr}(S) + \mathrm{tr}(SS^T).$$

The class of linear smoothers that we consider in this article are those commonly referred to as local polynomial smoothers. For the $p$th degree local polynomial smoother the $(i,j)$ entry of $S$ is

$$S_{ij} = e_1^T(X_{X_i}^T W_{X_i} X_{X_i})^{-1} X_{X_i}^T W_{X_i} e_j, \qquad (2.1)$$

where $e_i$ is the column vector with 1 in the $i$th position and zeroes elsewhere,

$$X_x = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix} \quad \text{and} \quad W_x = \operatorname*{diag}_{1\le i\le n} K\left(\frac{X_i - x}{h}\right) w(X_i). \qquad (2.2)$$

Typically, $K$ is a smooth bell-shaped function such as the standard normal density, called the *kernel*, and $h > 0$ is a scaling parameter, usually referred to as the *bandwidth*. The function $w$ is equal to the identity for local polynomial smoothing, but in likelihood-based models (discussed in Sec. 6.1) $w(X_i)$ may be something different. For example, in diagnostics for binary response models $w(X_i)$ is an estimate of $P(Y_i = 0|X_i)P(Y_i = 1|X_i)$. The local polynomial smooth at a general point $x$ is

$$\widehat{m}_p(x) = e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x Y. \qquad (2.3)$$

## 3. THE BINNING PRINCIPLE

The need for fast computational methods in local polynomial smoothing is borne out by the fact that the explicit expression for $\widehat{m}_p(x)$ depends on summations of the form

$$\widehat{\theta}(x) = \sum_{i=1}^{n} L_x(X_i) \quad \text{and} \quad \widehat{\psi}(x) = \sum_{i=1}^{n} L_x(X_i)Y_i, \qquad (3.1)$$

where $L_x$ is a generic function depending on $x$. For example, $\widehat{m}_0(x) = \widehat{\theta}(x)/\widehat{\psi}(x)$ with $L_x(u) = K\{(u - x)/h\}$. This means that direct computation of $M$ values of $\widehat{m}_p(x)$ requires $O(n \times M)$ function calls. Auxiliary quantities such as $\mathrm{df}^{\mathrm{err}}$ require $n$ evaluations of $\widehat{m}_p(x)$ resulting in an $O(n^2)$ algorithm. If $K$ has compact support then the number of kernel evaluations can be reduced to $O(n^2h)$. The computational labor for computation of such quantities can be significantly reduced by appealing to the *binning principle*, which we now describe.

Let $g_1 < \cdots < g_M$ be an equally spaced grid over the range of the $X_i$'s and let $\delta = (g_M - g_1)/(M - 1)$ be the gap between successive grid points. The grid count $(c_\ell, d_\ell^Y)$ at grid point $g_\ell$, with respect to linear binning, is given by

$$c_\ell = \sum_{i=1}^{n}(1 - |\delta^{-1}(X_i - g_\ell)|)_+ \quad \text{and} \quad d_\ell^Y = \sum_{i=1}^{n}(1 - |\delta^{-1}(X_i - g_\ell)|)_+ Y_i, \qquad (3.2)$$

where $x_+ = \max(0, x)$. Fan and Marron (1994) described a fast algorithm for obtaining the $(c_\ell, d_\ell^Y)$. The binning principle says that the quantities in (3.1) be replaced by

$$\widetilde{\theta}(x) = \sum_{\ell=1}^{M} L_x(g_\ell)c_\ell \quad \text{and} \quad \widetilde{\psi}(x) = \sum_{\ell=1}^{M} L_x(g_\ell)d_\ell^Y,$$

respectively. The approximation of $\widehat{\theta}(x)$ by $\widetilde{\theta}(x)$ and $\widehat{\psi}(x)$ by $\widetilde{\psi}(x)$ can be made arbitrarily better by making the grid sufficiently fine (see Sec. 7).

For local polynomial estimators, $L_x(u) = \kappa(u-x)$ for some function $\kappa$. This entails that

$$\widetilde{\theta}(g_j) = \sum_{\ell=1}^{M} \kappa\{\delta(\ell - j)\}c_\ell, \quad j = 1, \ldots, M, \tag{3.3}$$

from which it is apparent that no more than $M$ evaluations of $\kappa$ are necessary for computation of $\widetilde{\theta}$ over the entire grid. If the kernel has compact support, then even fewer, $O(Mh)$, kernel evaluations are necessary. Similar comments apply to the $\widetilde{\psi}(g_j)$. An algorithm for efficient computation of a vector of $\widetilde{\theta}(g_j)$ values is given in Scott (1992, p. 118). Alternatively, a fast Fourier transform algorithm could be used (Silverman 1982).

Fan and Marron (1994) applied this principle to obtain fast approximations to the $p$th degree local polynomial smoother by writing $\widehat{m}_p(x)$ in terms of expressions of the form $\widehat{\theta}(x)$ and $\widehat{\psi}(x)$. For our purposes it is more convenient to use binned versions of the weighted least squares notation used to define $\widehat{m}_p(x)$ at (2.3). Let $\widetilde{m}_p(x)$ be the binned approximation to $\widehat{m}_p(x)$ as obtained by Fan and Marron (1994) using the binning principle. Define

$$\widetilde{X}_x = \begin{bmatrix} 1 & g_1 - x & \cdots & (g_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_M - x & \cdots & (g_M - x)^p \end{bmatrix},$$

$$\widetilde{W}_x = \operatorname*{diag}_{1 \le \ell \le M} K\left(\frac{g_\ell - x}{h}\right) w(g_\ell),$$

$$C = \operatorname*{diag}_{1 \le \ell \le M} (c_\ell),$$

and

$$d^Y = (d_1^Y, \ldots, d_M^Y)^T.$$

Then the binned analogue of the smoother matrix is $\widetilde{S}$, where

$$\widetilde{S}_{\ell\ell'} = e_1^T(\widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} C \widetilde{X}_{g_\ell})^{-1} \widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} e_{\ell'}, \tag{3.4}$$

because it can be easily shown that

$$\widetilde{m}_p \equiv [\widetilde{m}_p(g_1), \ldots, \widetilde{m}_p(g_M)]^T = \widetilde{S} d^Y.$$

In other words, $\widetilde{S}$ maps the $Y$ grid counts to the vector of binned smooths at the grid points.

# 4. SOME KEY RESULTS

Several auxiliary quantities of interest can be expressed in terms of the diagonal entries of the matrices $S$ and $SAS^T$, where $A$ is some diagonal matrix. Therefore, it is useful to first study the properties of binned approximations to these two forms.

The $i$th diagonal entry of $S$ is

$$S_{ii} = e_1^T(X_{X_i}^T W_{X_i} X_{X_i})^{-1} X_{X_i}^T W_{X_i} e_i = K(0) e_1^T (X_{X_i}^T W_{X_i} X_{X_i})^{-1} e_1,$$

but its binned version is

$$\widetilde{S}_{\ell\ell} = K(0) e_1^T (\widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} C \widetilde{X}_{g_\ell})^{-1} e_1, \qquad (4.1)$$

corresponding to the grid points $g_1, \ldots, g_M$. For example, the $\widetilde{S}_{\ell\ell}$ can be used to approximate $\mathrm{tr}(S)$ as follows:

$$\mathrm{tr}(S) = \sum_{i=1}^{n} S_{ii} \simeq \sum_{\ell=1}^{M} \widetilde{S}_{\ell\ell} c_\ell.$$

Next, we treat $(SAS^T)_{ii}$. Typically, $A_{ii} = a(X_i)$ for some function $a(\cdot)$. For example, for calculation of the mean average squared error of a linear smoother (described in Sec. 5.5), one needs to take $a(X_i) = \mathrm{var}(Y_i|X_i)$. Let $\widetilde{A}$ be the diagonal $M \times M$ matrix with $\widetilde{A}_{\ell\ell} = a(g_\ell)$. First, observe that

$$(SAS^T)_{ii} = e_1^T (X_{X_i}^T W_{X_i} X_{X_i})^{-1} X_{X_i}^T W_{X_i}^2 A X_{X_i} (X_{X_i}^T W_{X_i} X_{X_i})^{-1} e_1.$$

The binned approximation to this quantity is then

$$(\widetilde{SAS^T})_{\ell\ell} = e_1^T (\widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} C \widetilde{X}_{g_\ell})^{-1} \widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell}^2 \widetilde{A} C \widetilde{X}_{g_\ell} (\widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} C \widetilde{X}_{g_\ell})^{-1} e_1 = (\widetilde{S}\widetilde{A}C\widetilde{S}^T)_{\ell\ell}. \qquad (4.2)$$

From this we see that the computation of the $(\widetilde{SAS^T})_{\ell\ell}$, $1 \le \ell \le M$, requires about the same amount of work as the computation of a binned regression function estimate over $g_1, \ldots, g_M$ and can therefore be carried out relatively quickly.

# 5. ILLUSTRATIONS

## 5.1 Error Degrees of Freedom

The binned approximation to $\mathrm{df}^{\mathrm{err}}$ follows very straightforwardly from the results of the previous section.

$$\mathrm{df}^{\mathrm{err}} = n - \sum_{i=1}^{n} \{2S_{ii} - (SS^T)_{ii}\} \simeq n - \sum_{\ell=1}^{M} \{2\widetilde{S}_{\ell\ell} - (\widetilde{SS^T})_{\ell\ell}\} c_\ell. \qquad (5.1)$$

The vectors of $\widetilde{S}_{\ell\ell}$ and $(\widetilde{SS^T})_{\ell\ell}$ values can be computed using (4.1) and (4.2).

## 5.2 CROSS-VALIDATION

For a general linear smoother with smoother matrix $S$ the cross-validation criterion can be defined to be

$$\mathrm{CV} = \sum_{i=1}^{n} \left\{ \frac{Y_i - \widehat{m}(X_i)}{1 - S_{ii}} \right\}^2 = \sum_{i=1}^{n} \left[ \frac{\{(I-S)Y\}_i}{1 - S_{ii}} \right]^2 .$$

For certain common smoothers, this is equivalent to a leave-one-out sum of squares. Typically, CV is a function of a smoothing parameter on which $S$ depends. The cross-validation choice of the smoothing parameter is the one that minimizes $CV$.

Expanding the numerator we see that, for $\widehat{m} = \widehat{m}_p$,

$$\mathrm{CV} = \sum_{i=1}^{n} \frac{Y_i^2 - 2\widehat{m}_p(X_i)Y_i + \widehat{m}_p(X_i)^2}{(1 - S_{ii})^2},$$

which can be written as a sum of terms of the form (3.1). Using the ideas in Section 3, and applying the binning principle once, we get

$$\mathrm{CV} \simeq \sum_{\ell=1}^{M} \frac{d_\ell^{Y^2} - 2\widehat{m}_p(g_\ell)d_\ell^Y + \widehat{m}_p(g_\ell)^2 c_\ell}{\{1 - e_1^T(X_{g_\ell}^T W_{g_\ell} C X_{g_\ell})^{-1} X_{g_\ell}^T W_{g_\ell} e_\ell\}^2},$$

where the $d_\ell^{Y^2}$ are obtained by binning the $Y_i^2$'s. Applying the binning principle once more, this time to the smooths in CV, we arrive at the binned approximation

$$\widetilde{\mathrm{CV}} = \sum_{\ell=1}^{M} \frac{d_\ell^{Y^2} - 2\widetilde{m}_p(g_\ell)d_\ell^Y + \widetilde{m}_p(g_\ell)^2 c_\ell}{\{1 - K(0)e_1^T(\widetilde{X}_{g_\ell}^T \widetilde{W}_{g_\ell} C \widetilde{X}_{g_\ell})^{-1} e_1\}^2} = \sum_{\ell=1}^{M} \frac{d_\ell^{Y^2} - 2\widetilde{m}_p(g_\ell)d_\ell^Y + \widetilde{m}_p(g_\ell)^2 c_\ell}{(1 - \widetilde{S}_{\ell\ell})^2}.$$

## 5.3 VARIANCE ESTIMATION

For a general linear smoother with matrix $S$, an estimate of the variance, $\sigma^2$, in a homoscedastic nonparametric regression model is

$$\widehat{\sigma}^2 = \frac{Y^T(I-S)^T(I-S)Y}{\mathrm{df}^{\mathrm{err}}}.$$

The numerator of $\widehat{\sigma}^2$ is the residual sum of squares, and the denominator is chosen so that $\widehat{\sigma}^2$ is an unbiased estimate of $\sigma^2$ in those situations where there is no bias in the smooth $SY$. Ruppert, Sheather, and Wand (in press) study variance estimators of this type in the local polynomial context.

Binned computation of the denominator for local polynomials is described in Section 5.1. The numerator equals

$$Y^TY - 2Y^TSY + (SY)^T(SY) = Y^TY - 2\sum_{i=1}^{n} \widehat{m}_p(X_i)Y_i + \sum_{i=1}^{n} \widehat{m}_p(X_i)^2,$$

so repeated application of the binning principle leads to the approximation

$$Y^TY - 2\sum_{\ell=1}^{M} \widetilde{m}_p(g_\ell)d_\ell^Y + \sum_{\ell=1}^{M} \widetilde{m}_p(g_\ell)^2 c_\ell.$$

## 5.4 STANDARD ERRORS

Consider the *partially linear model*

$$E(Y_i|X_i, Z_i) = m(X_i) + \beta^T Z_i, \quad i = 1, \ldots, n \tag{5.2}$$

where $m$ is some unspecified function, and $\beta$ is a vector of parameters with the same dimension as the $Z_i$'s. Also, assume that $\text{var}(Y_i|X_i, Z_i) = \sigma^2$ and conditional independence of the $Y_i$'s. If $S$ is a smoother matrix corresponding to a smooth of the $Y_i$'s on the $X_i$'s then a common estimate of $\beta$ is

$$\hat{\beta} = \{Z^T(I - S)Z\}^{-1} Z^T(I - S)Y$$

(e.g., Hastie and Tibshirani 1990), where $Z = (Z_1, \ldots, Z_n)^T$.

The covariance matrix of $\hat{\beta}$ can be estimated by

$$V = \hat{\sigma}^2 \{Z^T(I - S)Z\}^{-1} Z^T(I - S)(I - S)^T Z\{Z^T(I - S)Z\}^{-1}. \tag{5.3}$$

See, for example, Carroll, Fan, Gijbels, and Wand (1995). Expansion of (5.3) reveals that the difficult-to-compute components of $V$ are

$$Z^T S Z \qquad \text{and} \qquad (S^T Z)^T (S^T Z).$$

The entries of $Z^T S Z$ can be approximated straightforwardly by noting that they can be expressed in terms of a smooth of an appropriate column of $Z$. Let the notation for a $p$th degree polynomial smooth given at (2.3) be extended to $\hat{m}_p(x)^Y$ so that the $Y$ vector to which the smoothing is being applied is specified, and let $\tilde{m}_p(x)^Y$ denote the corresponding binned approximation. Also, let $\widetilde{Z}_i$ denote the $i$th column of $Z$. Then the $(i, j)$ entry of $Z^T S Z$ is

$$(Z^T S Z)_{ij} = \sum_{k=1}^{n} \hat{m}(X_k)^{\widetilde{Z}_j} \widetilde{Z}_{ki} \simeq \sum_{\ell=1}^{M} \tilde{m}(g_\ell)^{\widetilde{Z}_j} d_\ell^{\widetilde{Z}_i}.$$

Binned approximation of $(S^T Z)^T (S^T Z)$ takes a little more work. First observe that

$$\{(S^T Z)^T (S^T Z)\}_{ij} = \sum_{k=1}^{n} \left( \sum_{s=1}^{n} S_{sk} Z_{si} \right) \left( \sum_{s=1}^{n} S_{sk} Z_{sj} \right) \simeq \sum_{\ell=1}^{M} u_\ell^{\tilde{Z}_i} u_\ell^{\tilde{Z}_j} c_\ell, \tag{5.4}$$

where $u_\ell^Y = \sum_{\ell'=1}^{M} \widetilde{S}_{\ell'\ell} d_{\ell'}^Y$. It is a relatively straightforward exercise to show that a vector of $u_\ell^Y$ values, $1 \le \ell \le M$, can be computed with the same computational effort as a binned local polynomial smooth.

## 5.5 MEAN AVERAGE SQUARED ERROR

Exact risk analysis is a very useful technique for understanding the properties of curve estimators in finite samples (e.g., Marron and Wand 1992). The ideas presented in this article can be easily extended to fast computation of the exact mean average squared error (MASE) of a local polynomial kernel estimator. Suppose that the data are generated according to the model

$$Y_i = m(x_i) + v(x_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n,$$

where the $\varepsilon_i$ are uncorrelated random variables with zero mean and unit variance and $m$ and $v$ are known. Let

$$m = [m(x_1), \ldots, m(x_n)]^T \quad \text{and} \quad V = \text{diag}\{v(x_1), \ldots, v(x_n)\},$$

respectively, denote the mean vector and covariance matrix of $Y$. Then the MASE of $\widehat{m}$, can be expressed as

$$\begin{aligned} \text{MASE}(\widehat{m}) &= \frac{1}{n}\sum_{i=1}^{n} E\{\widehat{m}(x_i) - m(x_i)\}^2 \\ &= \frac{1}{n}\left[\text{tr}(SVS^T) + \sum_{i=1}^{n}\{(S-I)m\}_i^2\right] \\ &\simeq \frac{1}{n}\sum_{\ell=1}^{M}\left[(\widetilde{SVS^T})_{\ell\ell} + \{\widetilde{m}(g_\ell)^m - m(g_\ell)\}^2\right] c_\ell, \end{aligned}$$

where $\widetilde{m}(\cdot)^m$ is the binned approximation to the local polynomial smooth of $m$.

One could also use these ideas to compute accurate approximations to other global error criteria, such as mean integrated squared error.

# 6. GENERALIZATIONS

## 6.1 Likelihood-Based Models

The smoothers described in the previous three sections can be motivated by least squares considerations, which is equivalent to maximum likelihood under the assumption of normal errors. In recent years there has been a significant amount of research into the extension of smoothers to more general likelihoods (e.g., Hastie and Tibshirani 1990). In the generalized settings, the auxiliary quantities considered in Section 5 are replaced by weighted versions and the function $w$ in (2.2) is no longer the identity. For example, the error degrees of freedom corresponding to a smooth on binary response variables is given by

$$\text{df}^{\text{err}} = n - 2\text{tr}(S) + \text{tr}(ASA^{-1}S^T),$$

where $A$ is a diagonal matrix with $i$th diagonal entry equal to an estimate of $P(Y_i = 0|X_i)P(Y_i = 1|X_i)$. This situation was considered by Hastie and Tibshirani (1990, p. 306), who stated that, in the spline smoothing case, approximation of $\text{tr}(ASA^{-1}S^T)$ cannot be easily assessed.

In the case of local polynomial smoothing the binning principle is able to handle extensions of this type quite easily because of

$$\text{tr}(ASA^{-1}S^T) = \sum_{i=1}^{n} A_{ii}(SA^{-1}S^T)_{ii} \simeq \sum_{\ell=1}^{M} \widetilde{A}_{\ell\ell}(\widetilde{SA^{-1}S^T})_{\ell\ell}c_\ell$$

and result (4.2). Note that in this example $w(X_i) = A_{ii}$ so there is some cancellation when the expression is written out in full.

## 6.2 SINGLE-INDEX MODELS

Carroll et al. (1995) generalized (5.2) to

$$E(Y_i|X_i, Z_i) = m(\alpha^T X_i) + \beta^T Z_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where the $X_i$'s are $d$-dimensional variables and $\alpha$ is a $d \times 1$ vector of coefficients satisfying $\alpha^T \alpha = 1$. This is the partially linear extension of the *single-index* model Härdle, Hall, and Ichimura (1993). Carroll et al. (1995) derived local polynomial estimates of $m$, $\alpha$, and $\beta$ in model (6.1).

The covariance matrix of the estimates $(\widehat{\alpha}, \widehat{\beta})^T$ can be estimated by

$$V = \hat{\sigma}^2 \{PR^T(I - S)R\}^- PR^T(I - S)(I - S)^T RP\{PR^T(I - S)R\}^-, \tag{6.2}$$

where $A^-$ denotes a generalized inverse of a square matrix $A$,

$$R = \begin{bmatrix} \widehat{m'}(X_1)X_1^T & Z_1^T \\ \vdots & \vdots \\ \widehat{m'}(X_n)X_n^T & Z_n^T \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} I - \widehat{\alpha}\widehat{\alpha}^T & 0 \\ 0 & I \end{bmatrix}. \tag{6.3}$$

As in the univariate $X_i$ setting, discussed in Section 5.4, the hard work is the computation of $R^T SR$ and $(S^T R)^T (S^T R)$, so one can apply exactly the same ideas described there to obtain binned approximations to the estimated covariance matrix in this more general context. The extension to general likelihood-based models is also straightforward.

## 6.3 LOCAL BANDWIDTHS

The simplest example of a binned kernel estimator that arises in local polynomial smoothing contexts is the special case of (3.3) with $\kappa(\cdot) = K(\cdot/h)$:

$$\widetilde{\theta}(g_j) = \sum_{\ell=1}^{M} K\{\delta(\ell - j)/h\} c_\ell = \sum_{\ell=-L}^{L} K(\delta\ell/h) c_{\ell-j}, \tag{6.4}$$

where $L$ is the highest $\ell$ for which $K(\delta\ell/h) > 0$. From this second expression it is apparent that $L$ evaluations of $K$ are required to compute $\widetilde{\theta}(g_j)$ over the entire grid. However, this result is dependent on there being just a single *global* bandwidth $h$ for all $j$. Often it is desirable to have a set of *local* bandwidths $h_\ell$, $1 \leq \ell \leq M$, where $h_\ell$ is used for estimation at $g_\ell$. In this case, (6.4) generalizes to

$$\widetilde{\theta}(g_j) = \sum_{\ell=-L_j}^{L_j} K(\delta\ell/h_j) c_{\ell-j},$$

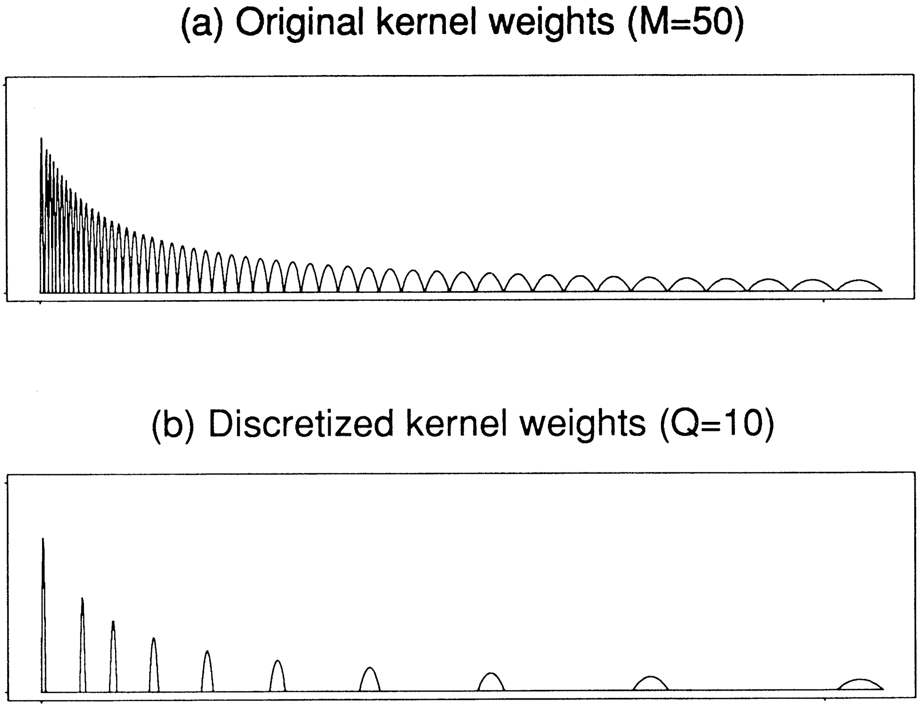

*Figure 1. Illustration of Bandwidth Discretization Idea. The 50 kernel weights in (a) are discretized to the 10 kernel weights in (b). The discretization is done using a logarithmically equally spaced grid of bandwidths.*

where $L_j$ is the smaller of $M$ and the highest $\ell$ for which $K(\delta\ell/h_j) > 0$ (i.e., $L_j = \min(M, \max\{\ell : K(\delta\ell/h_j) > 0\})$), which means that $\sum_{j=1}^{M} L_j$ kernel evaluations are required. This can be relatively expensive if $M$ is large, such as $M = 400$.

An attractive way out of this problem is to discretize the bandwidths onto a grid of $Q$ logarithmically equally spaced bandwidths, where $Q$ is much smaller than $M$, say $Q = 25$. This idea is conveyed by Figure 1, where (a) shows the set of $M = 50$ kernel weights, with bandwidths in increasing order. There is little difference between adjacent kernels, however. A subset of size $Q = 10$ is shown in (b). It is apparent that, if the kernels in (a) are replaced by their respective closest kernels in (b) there will be little change in the binned approximation.

## 6.4 Multivariate Smoothers

The extension of the ideas presented here to multivariate smoothers is reasonably straightforward. Wand (1994) provided an account of multivariate local polynomial smoothers. Following the notation there, let $g_\ell$ be a typical member of a $M_1 \times \cdots \times M_d$ mesh of grid points, where $\ell$ is a $d$-variate vector ranging over the index set

$$\mathcal{I} = \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_d\}.$$

The $d$-variate analogue of the binned smoother matrix is not a matrix for $d > 1$, but rather a $2d$-dimensional array $(\widetilde{S}_{\ell\ell'})_{\ell\in\mathcal{I},\ell'\in\mathcal{I}}$ that satisfies

$$\widetilde{m}(g_\ell) = \sum_{\ell'\in\mathcal{I}} \widetilde{S}_{\ell\ell'} d^Y_{\ell'}.$$

For most of the auxiliary quantities discussed in previous sections, the binned approximation for multivariate smoothers is a natural extension of the univariate counterpart. For example, the multivariate analogue of (5.3) is

$$Y^TY - 2\sum_{\ell\in\mathcal{I}} \widetilde{m}(g_\ell)d^Y_\ell + \sum_{\ell\in\mathcal{I}} \widetilde{m}(g_\ell)^2 c_\ell.$$

## 7. ASSESSMENT OF ACCURACY

### 7.1 ASYMPTOTICS

Consider the generic estimator $\widehat{\psi}(x) = \sum_{i=1}^n L_x(X_i)Y_i$ and its binned approximation $\widetilde{\psi}(x) = \sum_{\ell=1}^M L_x(g_\ell)d^Y_\ell$. If the $d^Y_\ell$ are obtained using the linear binning strategy, then it is a simple exercise in Taylor expansion to show that for sufficiently smooth $L_x$,

$$\widetilde{\psi}(x) - \widehat{\psi}(x) \simeq \tfrac{1}{2}\delta^2 \sum_{i=1}^n R(\delta^{-1}X_i)\{1 - R(\delta^{-1}X_i)\}L''_x(X_i)Y_i, \tag{7.1}$$

where $R(x) = x -$ (greatest integer not exceeding $x$) (Hall and Wand in press). Because $R$ is bounded it is clear from (7.1) that the $\widetilde{\psi}(x)$ converges to $\widehat{\psi}(x)$ as $\delta \to 0$. This, in turn, guarantees the accuracy of the binned approximations to the quantities presented in the previous sections, since they can each be expressed as smooth functions of versions of $\widehat{\psi}(x)$.

### 7.2 SIMULATION RESULTS

To test the accuracy of binned approximations in practical circumstances we conducted a small simulation study. For $\mathrm{df}^{\mathrm{err}}$, CV, $\hat{\sigma}^2$, and MASE, data were generated according to model: $Y_i = \sin(a\pi X_i) + .5\varepsilon_i$ where, for $i = 1, \ldots, 1{,}000$, the $X_i$ are independently and uniformly distributed on the unit interval, and $\varepsilon_i$ are independent standard normal variates. We considered the extended model $Y_i = \sin(a\pi X_i) + \beta Z_i + .5\varepsilon_i$, where $Z_i$ are equi-probable 0–1 random variables to assess the accuracy of binned approximations to std. err.$(\hat{\beta}) = V^{1/2}$, the estimated standard error of $\hat{\beta}$.

The grid size $M$ was fixed at 401, the kernel was the standard normal density truncated to $[-4, 4]$, and the bandwidth was taken to be

$$h_0 = [2{,}000(a\pi)^3\pi^{1/2}\{2a\pi - \sin(2a\pi)\}]^{-1/5},$$

an approximation to the conditional MASE-optimal bandwidth. The simulation involved 500 replications. Table 1 shows the averages and standard deviations of the ratios of the exact quantity to its binned approximation for various values of $a$.

Table 1. Averages (standard deviations) of the Ratios of the Exact Quantity to its Binned Approximation for 500 Replications of Simulated Data With Bandwidth $h_0$. A full description is given in the text.

| | $a = 1$ | $a = 5$ | $a = 10$ |
|---|---|---|---|
| $\mathrm{df}^{\mathrm{err}}$ | 1.00001<br>(.0035e-4 ) | 1.00004<br>(.0385e-4 ) | 1.00009<br>(.1596e-4 ) |
| CV | .99998<br>(.0358e-4 ) | .99934<br>(.5831e-4 ) | .99717<br>(2.2403e-4 ) |
| $\hat{\sigma}^2$ | .99998<br>( .0350e-4 ) | .99936<br>( 0.5705e-4 ) | .99726<br>( 2.1664e-4 ) |
| std.err.($\hat{\beta}$) | .99998<br>( 0.3102e-4 ) | 1.00001<br>(.0438e-4 ) | 1.00007<br>(.1902e-4 ) |
| MASE | .99990<br>( 0.2212e-4 ) | .99890<br>( .8121e-4 ) | .99651<br>(2.7935e-4 ) |

One should expect the accuracy to worsen for larger $a$ because there is finer structure in the underlying regression function and the optimal bandwidth is smaller. The table shows that even in the most extreme case considered here, $m(x) = \sin(10\pi x)$, the binning error using 401 grid points is negligible.

It is well-known (e.g., Fan and Marron 1994) that binning error is greater for smaller bandwidths, since the corresponding function estimates have more curvature. To investigate the effect of a decrease in the bandwidth on the computation of auxiliary quantities, we re-ran the simulation with $h = h_0/5$. We hesitated to use a bandwidth smaller than this because the local polynomial estimates tend to become numerically unstable due to not having enough points in the fitting window (Seifert and Gasser in press). The results are given in Table 2.

We see that there is some loss of accuracy, although in most cases the error is still negligible. The main exception is CV which, for some of the samples, had a binned approximation that was much larger than the exact quantity. The reason for this appears to be the fact that each summand of the CV function has a pole at a certain small bandwidth.

Table 2. Averages (standard deviations) of the Ratios of the Exact Quantity to its Binned Approximation for 500 Replications of Simulated Data With Bandwidth $h_{0/5}$. A full description is given in the text.

| | $a = 1$ | $a = 5$ | $a = 10$ |
|---|---|---|---|
| $\mathrm{df}^{\mathrm{err}}$ | 1.0001<br>( .0861e−4 ) | 1.0004<br>( 2.9790e−4 ) | .9983<br>( 2.1600e−3 ) |
| CV | .9996<br>( 0.6841e−4 ) | .9736<br>( 0.6339e−1 ) | .8469<br>(1.7632e−1 ) |
| $\hat{\sigma}^2$ | .9997<br>(.6436e−4 ) | .9848<br>( 1.6760e−3 ) | .9341<br>( 5.9306e−3 ) |
| std.err.($\hat{\beta}$) | 1.0000<br>(.1036e−4 ) | 1.0019<br>( 3.1504e−4 ) | 1.0103<br>( 1.3591e−3 ) |
| MASE | .9995<br>( .8648e−4 ) | 1.0046<br>( 1.0126e−3 ) | 1.0155<br>( 3.1497e−3 ) |

Table 3. Averages (standard deviations) of the Ratios of the Elapsed Exact Quantity to its Binned Approximation for 500 Replications of Simulated Data With Bandwidth $h_0$. A full description is given in the text.

| | $a = 1$ | | $a = 5$ | | $a = 10$ | |
|---|---|---|---|---|---|---|
| df $^{err}$ | 16.86 | (2.64) | 27.13 | (6.88) | 31.05 | (9.59) |
| CV | 8.74 | (1.17) | 16.37 | (3.76) | 19.71 | (5.28) |
| $\hat{\sigma}^2$ | 4.07 | (0.54) | 6.87 | (1.45) | 8.23 | (2.43) |
| std.err.($\hat{\beta}$) | 16.01 | (2.03) | 27.76 | (6.06) | 32.74 | (8.96) |
| MASE | 14.54 | (1.84) | 25.66 | (5.65) | 30.47 | (8.26) |

It is easiest to explain this problem when $p = 0$. In this case the $i$th summand of CV has a pole at $h = K(0)/\{n\hat{f}(X_i;h)\}$, where $\hat{f}(\cdot\,;h)$ is the kernel density estimator based on $K$. If the bandwidth is such that $h \simeq K(0)/\{n\tilde{\hat{f}}(g_\ell;h)\}$ at a certain grid point $g_\ell$, then the binned approximation tends to inflate because of the pole. This problem occurred for only a small percentage of the simulated data sets. If, for example, 1% of the lowest values are trimmed from the sample of ratios for CV when $a = 10$, then the corresponding table entry becomes 1.0154 (3.1554e−3). This problem is not a major concern because, in practice, the objective is to find the minimizer of CV that will occur at a much larger bandwidth. To verify this, for each sample in our simulation study we calculated the CV curve at 20 logarithmically equally spaced bandwidths ranging from five times the binwidth to sixty times the binwidth. We calculated the exact CV curve and its binned approximation and the minimizer of each curve. In the case of $a = 1$, the ratio of these minimizers had an average of .9982 (3.181e−3) and in one sample the (absolute) minimum of both curves occurred at the smallest bandwidth. For $a = 5$ the average ratio was .9922 (1.037e−2), and for seven samples the minimum of both curves occurred at the smallest bandwidth. In this setting the minimum of the exact CV curve occurred for two further samples at the smallest bandwidth whereas the binned approximation for these two samples had its minimum within the range of bandwidths used. However, for $a = 10$, the minimum of the exact and the binned CV curve was at the smallest bandwidth for all of the samples. In this case the CV curve would need to be calculated over a grid of smaller bandwidths. This would require use of a smaller binwidth since, in our experience, the binwidth should be at most one-fifth of the bandwidth for a normal kernel.

Finally, we investigated the question of how much of a saving binning offers in terms of computation time. The simulations used to produce Table 1 were timed using the elapsed time component of the S-Plus function `unix.time()`. Table 3 contains the results. It can be clearly seen that the use of binning results in substantial time savings.

## ACKNOWLEDGMENTS

We are grateful to three anonymous referees for their helpful comments.

*[Received April 1995. Revised February 1996.]*

# REFERENCES

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1995), "Generalized Partially Linear Single-Index Models," unpublished manuscript.

Cleveland, W. S., and Grosse, E. (1991), "Computational Methods for Local Regression," *Statistics and Computing,* 1, 47–62.

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Hall, P., and Wand, M. P. (in press), "On the Accuracy of Binned Kernel Density Estimators," *Journal of Multivariate Analysis.*

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single Index Models," *The Annals of Statistics*, 21, 157–178.

Härdle, W., and Scott, D. W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics,* 7, 97–128.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, in *Monographs on Statistics and Applied Probability* (vol. 43), London: Chapman and Hall.

Loader, C. (1994), "Computing Nonparametric Function Estimates," in *Computationally Intensive Statistical Methods*, eds. J. Sall and A. Lehman, *Computing Science and Statistics* (vol. 26), Fairfax, VA: Interface Foundation of North America, Inc., pp. 356–361.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Errors," *The Annals of Statistics*, 20, 712–736.

Ruppert, D., Sheather, S. J., and Wand, M. P. (in press) "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association.*

Scott, D. W. (1981), "Using Computer-Binned Data for Density Estimation," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, New York: Springer-Verlag, pp. 282–294.

——— (1985), "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *The Annals of Statistics* 13, 1024–1040.

——— (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

Seifert, B., Brockmann, M., Engel, J., and Gasser, T. (1994), "Fast Algorithms for Nonparametric Curve Estimation," *Journal of Computational and Graphical Statistics,* 3, 192–213.

Seifert, B., and Gasser, T. (in press), "Finite Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association.*

Silverman, B. W. (1982), "Kernel Density Estimation Using the Fast Fourier Transform Statistical Algorithm AS 176," *Applied Statistics,* 31, 93–97.

Wand, M. P. (1994), "Fast Computation of Multivariate Kernel Estimators," *Journal of Computational and Graphical Statistics,* 3, 433–445.