1. The need for computationally efficient smoothing algorithms

Smoothing of data is a method of re-expressing the data points in a form that is easier to understand than the raw point cloud itself. In a *regression smoothing* problem a $(d+1)$-dimensional point cloud is observed consisting of observations $\{Y_i\}_{i=1}^n$ at values $\{X_i\}_{i=1}^n$ of the predictor variable, where $X_i \in \mathbb{R}^d$. It is assumed that with observation errors $\{\varepsilon_i\}_{i=1}^n$,

$$Y_i = m(X_i) + \varepsilon_i. \tag{1.1}$$

The goal of regression smoothing is to re-express the point cloud by approximating the function m. In a *density smoothing* problem a d-dimensional point cloud is observed and one is interested to understand the structure of the data by estimating the unknown density $f(x)$ from observations $\{X_i\}_{i=1}^n$.

Nonparametric techniques for regression and density smoothing are known for their flexibility and their ability to detect structures deviating from a postulated parametric model. For an overview of nonparametric techniques, see the recent monographs of Eubank (1988), Härdle (1990), Müller (1988), Silverman (1986), and Wahba (1990). A drawback, however, of nonparametric smoothing techniques is that they face rapidly increasing computational complexity as the dimension increases beyond one dimension, or when the sample size is bigger than some threshold, say several thousand points in one dimension. This computational burden is especially handicapping when resampling techniques are used to determine the smoothing parameter or to construct error bars.

In this paper nonparametric smoothing of data points by *Weighted Averaging of Rounded Points* will be discussed. This approach is based on discretizing the data which allows efficient computing in low and high dimensions. The approach is entirely natural for raw data automatically collected in rounded form. For example, each datum collected by the LANDSAT remote sensing satellite is stored in just eight bits; see section 4. We consider smoothing in "low and high" dimensions. A dimension will be called "low" if it is less than three or four since the smoothing operation can be interactively visualized on modern computing systems. At present, it seems reasonable to call dimensions beyond 10 "high" since our experience indicates that one rarely has enough observations to explore spaces of that dimension. The dimensions from $4-10$ are the subject of intense study, and it is not clear exactly where "high" begins; see Scott and Wand (1990).

Smoothing in high dimensional spaces results in an inherent lack of sufficient statistical precision. Huber (1985) has discussed the so-called "curse of dimensionality." Practically speaking, observations in higher-dimensional spaces are very sparse. This sparseness limits the application of smoothing methods since they are all basically constructed by local averages over sample points. To illustrate this surprising sparseness, we recall the small example by Friedman and Stuetzle (1981) who considered a uniform distribution on a 10-dimensional unit cube. How sparse are the points in this cube?

If the dimensions of the neighborhood are chosen to cover 10 percent of the range of each coordinate, then it will (on the average) contain only $(.1)^{10}$ *of the sample, and thus will nearly always be empty. If, on the other hand, one adjusts the neighborhood to contain 10 percent of the sample, it will cover (on the average)* $(.1)^{1/10} \simeq 80$ *percent of the range of each coordinate. This problem of sparsity basically limits the success of direct d-dimensional local averaging.*

Should we give up now that we know that smoothing in high dimensions is almost impossible unless we have billions of data points that we can't analyze effectively? No, we could still try to pursue the goal to extract the most interesting low dimensional feature. One very attractive class of such models that greatly improve statistical efficiency are the *additive* ones (Stone, 1985). A simple additive model is, for example, one where the regression function m is decomposed into a sum of simpler one-dimensional functions, i.e.,

$$m(\underset{\sim}{x}) = \sum_{j=1}^{d} g_j(x_j),$$

see Hastie and Tibshirani (1987). A more general class of additive models are those that are based on projecting the $\underset{\sim}{x} \in \mathbb{R}^d$ onto the real line, i.e.,

$$m(\underset{\sim}{x}) = \sum_{j=1}^{N} g_j(\beta^T \underset{\sim}{x}),$$

see Friedman and Stuetzle (1981). A similar model has been recently studied by Duan and Li (1990). They use Sliced Inverse Regression (SIR), a method of discreting the response variable in order to find interesting projections.

The assumption of additivity significantly improves statistical convergence properties but does not eliminate the complexity of computation. The additive model algorithms proposed in the literature are heavily based on efficient smoothing algorithms since most of them are iterative or optimize the smoothing parameter. Friedman's (1984) *supersmoother*, for example, used in the alternating conditional expectation (ACE) algorithm (Breiman and Friedman, 1985) needs three pilot symmetrized k–NN "smooths" (the tweeter, woofer and midrange).

Current algorithmic techniques in typical smoothing scenarios in low or high dimensions involve stepping through some of the following operations.

SMOOTHING. A nonparametric "smooth" is first computed for a particular smoothing parameter. This smooth may be a one–dimensional building block for a higher dimensional additive model or the smooth may be a simple possibly multivariate exploratory computation.

OPTIMIZATION. Often a functional of "interestingness" is optimized over a set of parameters. Huber (1985, Chapter III) presents several projection indices for projection pursuit. In the specific projection pursuit technique the optimization is performed simultaneously over linear combinations of the (transformed) data and a nonparametric regression smooth.

ITERATION. The process of extracting interesting features often involves changing the interestingness functional, or running an optimization technique from residuals or transformed statistics in order to find good approximations to additive models. For instance, the "local scoring" algorithm of Tibshirani and Hastie (1987, Section 2) to fit generalized additive models (GAM) by "backfitting" is highly iterative. It loops over the elements of the predictor variable and uses the Aitken-weighted least squares technique in each element.

CALIBRATION. A necessary operation in nonparametric smoothing is to calibrate the smooth by finding a good smoothing parameter. There are several concepts for defining an optimal smoothing parameter but all have the goal of construction of narrow confidence bands for comparison with alternative (possibly parametric) models. Cross-validation or related resampling methods have been used both for regression and density smoothing, see Härdle, Hall and Marron (1988), Silverman (1986), or Scott and Terrell (1987). A nonasymptotic way to set up confidence bands is to use some sort of resampling scheme. Härdle and Bowman (1988) and Härdle and Marron (1989) use the bootstrap from estimated residuals to calibrate the smoothing parameter and construct variability bands.

Most of these operations involve computations that are typically a function of the squared sample size, $O(n^2)$, if one uses straightforward implementations of the formulae. Improved computation may be obtained either by reducing the number of arithmetic operations or by using novel approaches that eliminate one or more of the above four basic operations. For example, equation (2.11) below illustrates a method that avoids an optimization step. Our plan is to demonstrate how the Weighted Averaging of Rounded Points (WARPing) method achieves the goal of improved computational efficiency. We have intentionally chosen this abbreviation because of the connotation of speed, although it might lead the reader to the conclusion that we distort the data. Later, we will show that the asymptotics of this method are not unreasonable.

WARPing consists of first rounding the data points, i.e., to reduce the complexity in a step that is linear in the sample size, $O(n)$, and then performing a weighted average of these rounded points. A significant benefit of this prebinning step is that the resulting array of estimates, which falls on a uniform grid, is precisely in the form required for surface visualization algorithms such as marching cubes (Lorensen and Cline, 1987). Thus the decision to prebin is directly related to the final visualization desired (Scott and Hall, 1989) and enhances the speed of interpretation and analysis of data.

Other authors have used prebinning for one-dimensional smoothing problems. Silverman (1982) has given an algorithm based upon the Fast Fourier Transform (FFT) for computing a Gaussian kernel density estimate by rounding data to a mesh with 2^M points. At least three examples of prebinning from regression exist. Cleveland's (1979) LOWESS algorithm bins the x data in order to take advantage of a clever updating scheme for computing a sliding series of local linear regressions. O'Sullivan's unpublished BART smoothing spline implementation uses the same trick. Härdle (1987) prebins the x and y data to apply the FFT for one step M-smoothing.

Our purpose in this paper is to illustrate the effectiveness of prebinning. The usefulness of multivariate prebinning is not widely acknowledged. An alternative approach is to use any of the efficient algorithms that exist for computing k–th nearest neighbor (k–NN) estimate, which is a particular example of an adaptive kernel estimate. We hope to expand the use of the simple but significant trick of prebinning, which Marron has suggested "could make nonparametric methods accessible to a PC."

2. Weighted Averaging of Rounded Points (WARPing)

2.1 Density and regression smoothing

It is easiest to demonstrate our approach in the univariate density smoothing context. The simplest density estimator for data $\{X_i\}_{i=1}^n$ is the histogram with bin width h,

$$HG_h(x;x_0) = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in B(x;x_0,h)) \tag{2.1}$$

where $B(x;x_0,h)$ denotes the unique bin containing x of the form $[x_0+kh, x_0+(k+1)h)$ for some integer k. Note that the histogram is a function of the origin x_0. It is well known that different choices of x_0 may result in quite different shapes of density estimates. Figure 1 shows several histograms of the Buffalo snowfall data (Parzen, 1979). Depending on the origin, the histogram $HG_h(x;x_0)$ has one, two, or three modes and looks skewed to either the right or left. A quite drastic difference can be observed between the estimates with bin origin 2.5 (secondary mode at left) and bin origin 10.0 (secondary mode at right).

Figure 1. Six histograms with bin width h=13.5 and bin origins x_0=0,2.5,5,7.5,10,12.5 for the Buffalo snowfall data.

A natural way of eliminating this nuisance parameter is to construct an average of histograms, each with bin width h, over a collection of "origin" choices (Scott, 1985a). In particular, consider the set $\{x_{0,\ell} = \ell h/M, \ell = 0, ..., M-1\}$. Then our estimate is simply

$$\hat{f}_M(x) = \frac{1}{M}\sum_{\ell=0}^{M-1} HG_h(x, x_{0,\ell}). \tag{2.2}$$

It is straightforward to see that formula (2.2) may be written as

$$\begin{aligned}\hat{f}_M(x) &= \frac{1}{M}\sum_{\ell=0}^{M-1} \frac{1}{nh}\#\{i : X_i \in B(x;x_{0,\ell},\ h)\} \\ &= \frac{1}{nh}\sum_{\ell=1-M}^{M-1} \frac{(M-|\ell|)}{M}\#\{i : X_i \in B(x;x_{0,\ell},\ h/M)\}\end{aligned} \tag{2.3}$$

Another way to view formula (2.3) is to say we are weighting points aggregated or rounded into the smaller bins of width $\delta = h/M$. In formula (2.3), observe that the data enter only in the form of frequencies in the bin $B(\cdot; x_{0,k}, \delta)$.

The averaged shifted histogram (ASH) in (2.3) is a special case of WARPing. As $M \to \infty$, formula (2.3) can be written in the familiar form of a kernel estimate

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.4}$$

where $K(u) = (1 - |u|)I(|u| \leq 1)$.

The triangular weights $(1 - |\ell|/M) \equiv w_M(\ell)$ in formula (2.3) can be generalized in an obvious way to other kernel weights. A simple way to generate weight sequences is to discretize a continuous weight function that is defined on the interval $[-1, 1]$. For example, $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$, the so-called biweight or quartic kernel, corresponds to the weights $w_M(\ell) = \frac{15}{16}(1 - \ell^2/M^2)^2$ for $|\ell| < M$. In practice we would normalize this weight sequence by a constant $C_{M,K}$ so that $M^{-1} \sum_\ell w_M(\ell) = 1$. This technique has been applied to the Buffalo snowfall data (Parzen, 1979). The sequence of WARP density estimates as the number of averages increases is shown in Figure 2.

Figure 2. Construction of the WARP density estimate of Buffalo snowfall data with M=1,2,4,6,8,16 using the biweight kernel. Bandwidth h fixed.

Using this generalization we can rewrite formula (2.3) in the general form

$$\hat{g}(x) = \frac{1}{M} \sum_{|\ell| < M} w_M(\ell) RP_{\iota(x)+\ell}\ , \tag{$WARP$}$$

where $\iota(x)$ is the bin in which x falls and where, in the above case of density smoothing, RP_j is the frequency of rounded points ($\equiv RP$) in the j-th bin. We will show that the above notion is not restricted to estimating the density itself, but also related functionals such as a regression smooth. By allowing the weight sequence to take on negative values, we can estimate other quantities such as the derivative of the density. For the biweight kernel, $K'(u) = \frac{15}{4} u(u^2 - 1)I(|u| \leq 1)$ so that the effective WARP weight $w_M(\ell)$ for derivatives is $\frac{15}{4}\ell/M(\ell^2/M^2 - 1)$. Further generalizations to multivariate data are accomplished by rounding along each coordinate axis. For example, the multivariate version of the ASH in equation (2.3) converges to a multivariate product triangle kernel estimate as $M \to \infty$.

Let us give an overview and summary of this technique for this introductory example of density smoothing.

- One has a small bin width δ defining a sequence of bins

$$B_j = [(j - 1/2)\delta, (j + 1/2)\delta], \quad j \in \mathbb{Z}$$

  and the *index function* $\iota(x) = j \;\Leftrightarrow\; x \in B_j$.

- The approximation parameter $M = h/\delta$ determines how many adjacent bins enter into the averaging process.

- The WARPed density estimate is

$$\hat{f}_M(x) = \frac{C_{M,K}}{nh} \sum_{i=1}^{n} K\left(\frac{\iota(x) - \iota(X_i)}{h}\right),$$

  where the factor $C_{M,K}$ is introduced to guarantee that $\hat{f}_M$ integrates to one. The constant $C_{M,K}$ is given by

$$C_{M,K} = M / \sum_{\ell=1-M}^{M-1} K\left(\frac{\ell}{M}\right).$$

  The discretized quartic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \le 1),$$

  for example, has the effective weight

$$C_{M,K}\ K\left(\frac{\ell}{M}\right) = \frac{15M^4}{16M^4 - 1}\left(1 - \left(\frac{\ell}{M}\right)^2\right)^2.$$

For regression smoothing, we wish to approximate $m(x)$ by a weighted average of the responses for which the predictor values are in a neighborhood of x (Stone 1977). Using the WARP approach this amounts first to rounding the predictor values to the closest bin center and computing the average response in each bin; then second, appropriately weighting these bin averages in a neighborhood of x to estimate $m(x)$. For the regression case we consider two different estimators, a step function approximation

$$\hat{m}_M^{(S)}(x) = \frac{\sum_{i=1}^{n} K\big(\frac{\iota(x)-\iota(X_i)}{M}\big) Y_i}{\sum_{i=1}^{n} K\big(\frac{\iota(x)-\iota(X_i)}{M}\big)} \tag{2.5}$$

and a polygon through the midpoints of the steps of $\hat{m}_M^{(S)}(x)$. For the latter, define $d_x = \iota(x) - x/\delta$ the distance between x and the bias center $\iota(x)\delta$. The polygon approximation is defined through

$$\begin{aligned} \hat{m}_M^{(P)}(x) &= (1 - |d_x|)\ \hat{m}_M^{(S)}(\iota(x)\delta) \\ &\quad + |d_x|\big(I(x \in [(\iota(x) - 1/2)\delta), \iota(x)\delta))\ \hat{m}_M^{(S)}((\iota(x) - 1)\delta) \\ &\quad + I(x \in [\iota(x)\delta, (\iota(x) + 1/2)\delta))\ \hat{m}_M^{(S)}((\iota(x) + 1)\delta)\big). \end{aligned} \tag{2.6}$$

To give some insight note that by construction

$$K\big(\frac{\iota(x) - \iota(X_i)}{h}\big) = \sum_{\ell \in \mathbb{Z}} K(\ell/M) \times I(X_i \in B_{\iota(x)+\ell})$$

and therefore

$$\hat{m}_M^{(S)}(x) = \frac{\sum_{\ell=1-M}^{M-1} K(\ell/M) Y_{\bullet,\iota(x)+\ell}}{\sum_{\ell=1-M}^{M-1} K(\ell/M) n_{\iota(x)+\ell}}, \tag{2.7}$$

where

$$Y_{\bullet,j} = \sum_{i=1}^{n} Y_i\ I(X_i \in B_j),$$

$$n_j = \sum_{i=1}^{n} I(X_i \in B_j).$$

The rounded points RP_j are here the sum of the response variables and the frequency of the predictor variable in bin B_j. In principle, the Y_i values can be binned, but there is no computational advantage in doing so.

Figure 3 shows $\hat{m}_M^{(S)}(x)$, $\hat{m}_M^{(P)}(x)$, and the Nadaraya–Watson kernel estimator

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

Figure 3. WARPing approximations (solid lines) and the Nadaraya–Watson kernel estimator (dashed line) for the Geyser data-set. Bandwidth h=0.8, M=15.

Qualitatively, the general WARP formula says that we first discretize the data and then smooth; compare equations (2.5) and (2.7), for example. Thus the computational advantage of the WARP method results from the partial decoupling of the sample size and the smoothing computation. The first step of binning is a computational burden of order $O(n)$ that produces a data compression into $1/\delta$ bin counts on a mesh that can be easily maintained. The smoothing is then applied to this smaller bin object taking advantage of the regular spacing of bins which in particular pays off if we use a kernel with compact support. Such a kernel allows fast computation since we know from the regularly spaced bin points where the weighting scheme gives weight zero outside the support of the kernel. Alternative degrees of smoothing can be tried without having to rebin the data.

2.2 Comparison of computational costs

To have some insight into the computational advantage of the WARP method, consider a comparison of three possible procedures for estimating a density. Assume that the density is to be computed at δ^{-1} equidistant points. The brute force kernel method amounts to averaging the rescaled kernel functions $K((x - X_i)/h)$ over the whole sample at δ^{-1} points resulting in $\delta^{-1}n$ operations. If the kernel has compact support and $d = 1$, a presorting can reduce this to $2nh/\delta$ but, in general, the cost is n/δ. The method of Silverman (1982) and of Härdle (1987) consists of first discretizing the data into δ^{-1} bins (assumed to be a power of 2) and then using the FFT. After the FFT of the discretized data, smoothing is performed in the frequency domain (δ^{-1} multiplications for a low-pass filter). Then the back FFT is applied. The discretization cost is n and that of the two FFT's is $2\delta^{-1}\, log_2\, \delta^{-1}$, which results in the number given in Table 1.

As explained above, the WARP method is performed in two steps: first n steps to discretize the data and, then, in a window of $2M-1$ the rounded points $RP_{\iota(x)+\ell}$ are averaged, resulting in $\delta^{-1}(2M-1)+n$ total cost. We have excluded comparison with other smoothing methods such as discrete maximum penalized likelihood estimators (DMPLE or DiMPLE) (Scott, Tapia, and Thompson, 1980) and orthogonal polynomials (Cencov, 1962), since they require iterative solution or are equivalent to a particular kernel method, respectively. The k–NN estimator is also relatively efficient, $O(n \, log \, n + \delta^{-1} \, log \, n)$, but is deficient for many purposes since it is often too bumpy and visually rough (Silverman, 1986). Observe that the $O(n \, log \, n)$ work for sorting is replaced by prebinning, which is essentially a sorting operation.

Method	Operation Count
Kernel	$\delta^{-1}n$
WARP	$\delta^{-1}(2M-1)+n$
FFT	$\delta^{-1}(1+2 \, log_2 \, \delta^{-1})+n$

Table 1. Operation counts for some univariate nonparametric density estimators.

Of course, the costs of the WARP and FFT are quite similar since they are both based on discretization ideas. For example, when we discretized into 1000 bins the factor of δ^{-1} is 21 for the FFT method and 19 for $M = 10$. Such an M is already rather big, we work mostly with $M = 5$ in practice. The differences become much more drastic if we consider high dimensional smoothing problems or data with many empty bins. While the FFT technique of Silverman (1982) extends to bivariate (x, y) regression smoothing (see Härdle, 1987), it does not extend so easily to higher dimensions. By keeping the pointers to nonempty bins the discretization can be performed only for the nonempty bins so that the computation count for WARPing is in fact smaller, namely, dependent upon the number of nonempty bins, $NB(\delta, n)$.

Method	Operation Count
Kernel	$2(\delta^{-1})^d n$
WARP	$2NB(\delta, n)^d (2M-1)^d + nd$

Table 2. Comparison of operation counts between brute force kernel smoothing in d dimensions and the WARP method.

Of course, the number of nonempty bins $NB(\delta, n)$ is in $\{1, \ldots, n\}$, but in most cases the number is much smaller than n. To see this in the case of $d = 1$, denote by $R = X_{(n)} - X_{(1)}$ the span of the data and $[R/\delta]$ the next integer larger than R/δ. Of course, $NB(\delta, n) \leq min(n, [R/\delta])$, but let us look more closely at the random variable $NB(\delta, n)$ for the case of uniform X-variables, i.e., $P(X_i \in B_j) \approx 1/[R/\delta]$. This assumption is rather conservative in the sense that other X-distributions will produce a smaller $NB(\delta, n)$.

Proposition 2.1 *The number of nonempty bins $NB(\delta, n)$ can be represented as a homogeneous Markov chain with transition matrix*

$$P = \frac{1}{[R/\delta]} \begin{pmatrix} 1 & [R/\delta]-1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 2 & [R/\delta]-2 & 0 & \ldots & \ldots & 0 \\ & & & \ldots & & & \\ 0 & \ldots & k & [R/\delta]-k & 0 & \ldots & 0 \\ & & & \ldots & & & \\ 0 & \ldots & \ldots & \ldots & 0 & [R/\delta]-1 & 1 \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & [R/\delta]. \end{pmatrix}$$

The starting distribution is given by

$$p_1^T = (1 \; 0 \; \cdots \; 0).$$

From this proposition we can compute the expectation

$$E(NB(\delta, n)) = p_1^T P^{n-1} (1 \; 2 \; 3 \; \cdots \; [R/\delta])^T.$$

Figure 4 shows the increase of computational cost for the kernel smoother as a function of n. Figure 5 shows the computational cost for WARPing. One clearly sees that the cost stays relatively constant as n increases.

Figure 4. Computational cost of the kernel smoother as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

Figure 5. Computational cost for WARPing as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

The operation counts for density or regression estimation for one-dimensional x's will be slightly high, since empty bins require no operations. For bivariate x's, usually a substantial number of bivariate bins are empty, so that an operation count estimate of δ^{-1} may in fact be only $0.3\delta^{-1}$. For three-dimensional x's, it is possible to proceed exactly as in the previous two cases, accepting a large number of empty bins. However, if n is not too large, then it may be more attractive to keep an unarrayed list of bin counts, specifically, three integers pointing to a trivariate bin plus the integer bin count (with an additional real variable containing the sum of the bin responses in the regression case).

When the dimension $d \geq 4$, it is almost always sensible to store the data in such an intermediate compressed form, since only three-dimensional slices will be computed and displayed. It is possible to maintain $30^4 < 10^6$ bins in memory in the quadravariate case, which may be necessary when using animation techniques (Scott, 1986). Only rarely would it be sensible to estimate at one instant the density on a full mesh in R^d for $d > 5$. In such cases graphical display is not the goal but rather estimates at the sample points may be desired.

It is interesting to mention the symmetrized k–NN smoother that is used in *supersmoother*, Friedman's (1984) variable span smoother. This smoother is only defined for one-dimensional scatterplot smoothing and requires pre-sorting which results in $n(1 + log\, n)$ operations altogether. By contrast, the WARP kernel smoother requires $n + \delta^{-1}M$ operations to produce a smooth curve. In fact, Friedman has tried substituting a one-dimensional WARP estimator for a k–NN smoother not for any computational reason but because the smoothness of the WARPing approach greatly improved the iteration in one version of the ACE algorithm.

As we will see in Section 3 below, the discretization step does not affect, in an asymptotic sense, the statistical accuracy of the WARP smoother. The decoupling of the sample size by introducing the rounded points makes the optimization of the nonparametric smoothers much more efficient. This becomes apparent when regression and density smoothers of any kind have to be calibrated by a smoothing parameter. In the WARP–estimator above this smoothing parameter is the number of bins M over which the average is performed. The difference in computational cost becomes drastic if cross-validation or related scores are to be computed for optimization of smoothing parameters. Scott and Schmidt (1988) in an analysis of British income data compared the least-squares and biased cross-validation techniques to optimize the smoothing parameter. Using the direct formulae used up as much as 8 CPU hours on an IBM 3081 mainframe for $n = 7123$. Using the WARPing approach brought the computation of these scores down to 30 seconds on an IBM AT (Processor 80286, 10MHz).

For regression smoothing, ordinary cross-validation requires multiple computation of the smoother at all the sample points of the cross-validation score at several bandwidths, although efficient algorithms exist for certain linear smoothers associated with spline smoothers (Eubank, 1988). The WARP estimator has the advantage that once the discretization step is performed, all further smoothing operations are linear in δ^{-1}.

2.3 Smoothing in high dimensions by means of additive models

In this section we illustrate the use of WARPing in high dimensions by an additive model example. The process of modeling high dimensional point clouds by additive models typically involves iterative approximations in order to minimize some score function. Adopting the additive model (Breiman and Friedman, 1985), construct nonparametric functions ψ^*, $\{\varphi_j^*\}_{j=1}^d$ such that

$$e^2(\psi, \varphi_1, ..., \varphi_d) = \frac{E\{[\psi(Y) - \sum_{j=1}^d \varphi_j(X_j)]^2\}}{E\psi^2(Y)} \tag{2.8}$$

is minimized. Alternating conditional expectations have to be computed based on an efficient one-dimensional scatterplot smoother. In this context, the WARP kernel smoother has to perform $2\delta^{-1}M + n$ operations for this task. Depending on δ and M it can be made highly efficient compared to a kernel or symmetrized k–nearest neighbor smoother (Carroll and Härdle, 1988).

The projection pursuit regression (PPR) algorithm of Friedman and Stuetzle (1981) requires an efficient one-dimensional scatterplot smoother as well. The PPR algorithm searches first for the best pair (β, g) such that with a suitably normalized direction β the residual sum of squares

$$\sum_{i=1}^{n}(Y_i - g(X_i^T\beta))^2 \tag{2.9}$$

is minimized. This task is done by iterating over several smoothing operations. In a further step residuals are fitted and the same procedure is run for the set of estimated residuals. Then another set of residuals is computed and this iteration is continued until a convergence criterion is met.

The direction β of a one-step PPR can also be estimated without optimization or iteration by proving that β is proportional to a certain expectation:

$$\eta \equiv E_X[\nabla m(X)], \tag{2.10}$$

where $\nabla m(x)$ denotes the gradient vector of partial derivatives. The so-called average derivative, η, can be estimated by

$$\hat{\eta} = \frac{-1}{n}\sum_{i=1}^{n} Y_i \frac{\hat{f}_h'(X_i)}{\hat{f}_h(X_i)}, \tag{2.11}$$

where $\hat{f}_h'(\cdot)$ denotes the vector of partial derivatives of the kernel estimate with bandwidth h. This can be done effectively with WARPing since averaging has to be performed only over nonempty bins. Once the estimate $\hat{\eta}$ is obtained, the one-dimensional function g is estimated by smoothing the bivariate scatterplot $\{(\hat{\eta}^T X_i, Y_i)\}_{i=1}^n$. The estimator $\hat{\eta}$ from (2.11) has desirable root–n convergence properties, and the additivity of the models enables a rate of convergence for m that is typical for one-dimensional problems; for details, see Härdle and Stoker (1989). The average derivative is also helpful in estimating parameters in the partial linear model

$$Y = \beta^T x + m(z) + \varepsilon, \qquad x \in \mathbb{R}^d, \quad z \in \mathbb{R}^k \ ;$$

see Spiegelman (1976), Rice (1986), and Heckman (1986). The first d components of the average derivative are equal to β.

Generalized additive modeling involves, similarly to ACE, finding functions $\{g_j\}_{j=1}^d$ such that

$$m(\underset{\sim}{x}) = G\left(\sum_{j=1}^{d} g_j(x_j)\right)$$

with a nonlinear (inverse) link function $G(\cdot)$. The functions g_j are determined by a combination of the *backfitting* and the *local scoring algorithm*. For a definition of this algorithm, see Hastie and Tibshirani (1987). We applied this algorithm to a data-set of simulated side-impacts with Opel Kadetts. The response variable was $Y = 0$ or 1 for survival or nonsurvival, respectively, as determined by examining the occupants after impact. The predictor variables for survival were X_1 = AGE, X_2 = VELocity, X_3 = ROSYM, a measure of acceleration at the chest. A draftsman's graphic of this four-dimensional data-set is presented in Figure 6. We will return to these data in section 4.

Figure 6. A draftsman's plot of the four side-impact variables. Made with XploRe (1990). X_1= AGE, X_2= VEL, X_3= ROSYM, Y= response.

3. Asymptotics of WARPing

3.1 How precise is WARPing?

Since WARP algorithms are identical to kernel smoothing when the fine bin width $\delta = h/M \to 0$, it is not surprising that the asymptotic bias and variance expressions of WARP and kernel estimators are quite similar. For example, with a univariate WARP density estimator with asymptotic triangular kernel $K(u) = (1 - |u|)I(|u| \leq 1)$, the mean integrated squared error is given by

$$(3.1)\quad MISE \simeq \frac{2}{3nh}(1+\frac{1}{2M^2}) + \frac{1}{12}\delta^2 \int (f')^2 + \frac{1}{144}(1-\frac{2}{M^2}+\frac{3}{5M^4}) \int (f'')^2.$$

The second and third terms are bias terms, with the second being a histogram-like bias term. Clearly, the fine bin width δ should be chosen sufficiently small (or, equivalently, M sufficient large since $h = M\delta$ with h conceptually fixed near its optimal value) so that the second term is negligible compared to the third term. In practice, the data are often presented with finite precision so that δ (and hence h) cannot be made arbitrarily small. When δ can be chosen arbitrarily small and assuming a reasonable choice of h is known, then a (δ, h) pair should be specified so that $M \geq 5$ but not so large to eliminate the benefit of binning. Scott (1985a) also showed that the histogram-like bias term could be eliminated by the additional work of constructing a piecewise linear interpolant of the WARP estimates at the bin centers, for which

$$MISE \simeq \frac{2}{3nh} + \frac{1}{144}h^4(1 + \frac{1}{M^2} + \frac{9}{20M^4}) \int (f'')^2.$$

This expression is much easier to understand and for relatively small $M \geq 3$, the WARP estimate is both theoretically and graphically virtually identical to the triangle kernel estimate. This particular issue is discussed in greater detail in Scott and Sheather (1985) and Jones (1989). Hjort (1986) has developed expressions for multivariate averaged shifted histograms using a linear blend interpolation procedure.

The mean squared error expansion is more complicated for the WARPed regression smoothers. Let $\sigma^2(x) = \mathrm{var}(\varepsilon \mid x)$. Breuer (1990) proved the following propositions.

Proposition 3.1 *As $h \sim n^{-1/5}$, $M \sim n^{\beta}$, $\beta \geq 1/5$,*

$$\begin{aligned} MSE[\hat{m}_M^{(S)}(x)] & \\ &\simeq (nh)^{-1}\frac{\sigma^2(x)}{f(x)} \parallel K_M \parallel_2^2 + \frac{h^2}{M^2} d_x^2 (m'(x))^2 \\ &+ \frac{h^3}{M} d_x \left(\mu_2(K_M) + \frac{12d_x^2+1}{12M^2} \right) m'(x) \left\{ m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \right\} \\ &+ \frac{h^4}{4} \left(\mu_2(K_M) + \frac{12d_x^2+1}{12M^2} \right) \left\{ m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \right\}^2 \end{aligned}$$

where

$$\|K_M\|_2^2 = \frac{C_{M,K}^2}{M} \sum_{\ell} K^2\left(\frac{\ell}{M}\right)$$

and

$$\mu_2(K_M) = \frac{C_{M,K}^2}{M^3} \sum_{\ell} \ell^2 K\left(\frac{\ell}{M}\right).$$

Proposition 3.2 *Let $h \sim n^{-1/5}$, $M \sim n^{\beta}$, $\beta > 0$,*

$$MSE[\hat{m}_M^{(P)}(x)]$$

$$\simeq (nh)^{-1}\frac{\sigma^2(x)}{f(x)}\Big(\| K_M \|_2^2 - 2|d_x|(1-|d_x|)(\| K_M \|_2^2 - \gamma(K_M))\Big)$$
$$+ \frac{h^4}{4}\Big[m''(x)\Big(\mu_2(K_M) + \frac{12|d_x|(1-|d_x|)+1}{12M^2}\Big)$$
$$+ 2m'(x)\frac{f'(x)}{f(x)}\Big(\mu_2(K_M) + \frac{1}{12M^2}\Big)\Big]^2$$

where

$$\gamma(K_M) = \frac{C^2_{M,K}}{M} \sum_{\ell=1-M}^{M-1} K\left(\frac{\ell}{M}\right) K\left(\frac{\ell+1}{M}\right).$$

These two propositions are remarkable. They show the basic difference between the step function $\hat{m}_M^{(S)}$ and the continuous polygon approximation $\hat{m}_M^{(P)}$: One sees that if binning is done so that $M \sim n^{1/5}$ the terms depending on M have the same speed as those depending solely on h. Thus if $M \sim n^{1/5}$ the MSE of WARPing is different from that of the kernel smoother $\hat{m}_h$. By contrast the estimator $\hat{m}_M^{(P)}$ will reflect the correct MSE properties of $\hat{m}_h$ provided $M \sim n^{\beta}$, $\beta > 0$.

3.2 The price of discretization

An example of the MSE of $\hat{m}_M^{(S)}$ is plotted as a function of x in Figure 7. One sees especially in the right half of the interval the increased bias of $\hat{m}_M^{(S)}$ for the fact of not being directly in a bin center. The effect is, of course, here made very drastic since we are averaging over a very large bandwidth and use only $M = 5$ bins for this h.

Figure 7. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(S)}$ (dashed line) for a simulated example $m(x)=x\,sin(2\pi x)$, $x \in U(0,1)$, $\varepsilon \sim N(0,\sigma^2)$, $\sigma^2=0.25$, $h=0.25$, $M=5$, $n=100$, $K=$ Quartic.

Figure 8 shows the MSE as a function of x for the estimator $\hat{m}_M^{(P)}$. Note that the scale is different here. We can see how much better $\hat{m}_M^{(P)}$ is compared to $\hat{m}_M^{(S)}$. For more details, in particular how to bound the difference between the MSE curves, see Breuer (1990).

Figure 8. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(P)}$ (short dashed line) together with bounds on the difference between the two MSE curves. $m(x){=}x\,sin(2\pi x)$, $x{\in}U(0,1)$, $\varepsilon{\sim}N(0,\sigma^2)$, $\sigma^2{=}0.25$, $h{=}0.25$, $M{=}5$, $n{=}100$, $K{=}$ Quartic.

4. WARPing in practice with visualization

The WARP technique was applied to an analysis of laboratory-simulated car side-impacts, see the data in Section 2. Figure 9 shows the WARP density estimate of three biomechanical variables for 29 dummies. The visualization technique has been discussed in Scott and Thompson (1983). The "flying duck" shape of the distribution of the dummy's variables is substantially different from the "frozen duck" shape in Figure 10 that was computed for a different group of experiments. Interestingly there are two clusters at the "drumsticks" of the frozen duck. Both density contours are plotted on the same scale and show very different shapes due to the fact that the frozen duck has higher z-values and lower x-values than the flying duck. Note that the contours would be spherical if the data were independent and normal.

Figure 9. WARP dummy estimates (biweight kernel) for the variables ($ROSYM$, $RUSYM$, $T12RM$)$=(x,y,z)$ giving maximal acceleration (in g) at different regions of the thorax of the dummy's bodies. The contour is at a level 5% of the modal level.

Figure 10. WARP estimates (as in Figure 9) for a different group.

To investigate these data further the GAM/WARP approach has been used. Note that this model generalizes related concepts like logistic regression models where the link function G is known. Figure 11 shows the fifth iteration of the local scoring algorithm (done with XploRe (1990)) using the logistic link function $G(u) = exp(u)/(1 + exp(u))$. It is interesting that the lower left curve shows a typical nonlinear structure. Figure 12 finally shows the functions $\hat{g}_j(x_j)$, $j = 1, 2, 3$. These functions are not linear and not all monotone as might be assumed. However with such small samples, such conclusions can only be tentative.

Figure 11. Fifth step of the local scoring algorithm applied to the side-impact data. The info box shows the decrease in deviance as the iteration proceeds. Behind the info box one sees the current estimate $\sum_{j=1}^{d} \hat{g}_j(x_j)$ plotted versus Y with the link function G.

Figure 12. The WARP estimated function, $\hat{g}_j(x_j)$, j=1,2,3 for the side-impact example.

The WARP technique also applies directly to four–dimensional data, but one has to look at a sequence of contour slices with one variable fixed (Scott 1986). We re–examined the classical Iris data with variables (x_1, x_2, x_3, x_4) given by sepal length, petal width, petal length, and sepal width, respectively. The contours of $\hat{f}(x_1, x_2, x_3, x_4 = 3.14\,cm)$ (not shown) clearly reveal three separate spherical contours representing the three clusters of Iris flower varieties present. Other analyses show that the data for two of these varieties are not well-separated. This example is provocative since the sample size is so small.

The final example illustrates the application of WARPing to large datasets. The simpler visualization techniques used previously can be replaced by more sophisticated surface rendering methods, since large-sample estimates warrant closer inspection (Scott and Hall, 1989). These data were derived from LANDSAT IV measurements taken in summer 1977 on segment 1663, a 5 by 6 nautical mile region in North Dakota in primarily agricultural use (Scott, 1983; Scott and Jee, 1984; and Scott, 1985b). Each segment contains 117 scan lines each with 196 picture elements (pixels), for a total of 22,932 1.1 acre pixels. Five acquisitions were obtained by the 4-channel satellite, which measure light reflectance intensities in fairly narrow bands. These 20 dimensions of data (ignoring any spatial information) were projected into 3 dimensions by the nonlinear model of Badhwar (1980). For agricultural pixels, the three transformed variables have the following interpretations; x, the time of peak "greenness"; y, the ripening period; and z, the peak greenness level.

Figure 13 shows raw 256–bin histograms of these variables. (Image data are usually digitized into 8 bits, numbered 0 to 255, for obvious reasons.) Two features are apparent from these histograms: first, the x-variable is strongly bimodal; second, much of the "bandwidth" in the 8 bits is wasted, particularly in the first two channels.

Figure 13. Histograms of the 3 channels of transformed LANDSAT data.

Figure 14 shows the pairwise scatter diagrams of these data. To improve the visualization, the data have been "blurred" by adding uniform noise to the integer values. More multivariate structure is apparent in this diagram than in Figure 13, but it is well to remember that the features away from the central cluster represent less that 10% of the data.

Figure 14. Pairwise scatter diagrams of the 3 channels of transformed LANDSAT data.

A WARP density estimate was computed over the region $[65, 130] \times [0, 75] \times [0, 113]$, which excluded 474 points ($< 2.1\%$). Each axis was divided into 40 bins and $M_i = 3$ was chosen subjectively. Figure 15 displays the $\alpha = 20\%$ density contour, which is clearly bimodal. The left ellipse, which represents primarily sunflower crops, actually leans towards the viewer, while the right ellipse, which represents small grain crops, leans away and is further from the viewer. Observe that the WARP bin structure and triangulation of the marching cube algorithm are both readily visible. Some further surface shading algorithms may be contemplated, but smoothing away the bin edges and local noise information is not to be recommended generally.

Figure 15. Density shell $S_{20\%}$ for the WARP estimate of the LANDSAT data.

Figures 16–18 display the $S_{10\%}$, $S_{5\%}$, and $S_{2\%}$ contour shells. (Conceptually, these shells are nested, and a transparency visualization algorithm could be employed.) The bump towards the lower right in Figure 17 actually protrudes more than the bump towards the bottom left, as is clear from a stereo or animated representation. Figure 18 displays not only the beautiful structure arising from the complex mixing of different pure crop data, but also begins to show some small bumps in the tails of the data. Given the very large sample, the detailed structure is most likely real and not an artifact.

We believe these surface renderings will become commonplace in nonparametric density and regression applications and that WARPing is essential to implement interactive versions of such applications.

Figure 16. Density shell $S_{10\%}$ for the WARP estimate of the LANDSAT data.

Figure 17. Density shell $S_{5\%}$ for the WARP estimate of the LANDSAT data.

Figure 18. Density shell $S_{2\%}$ for the WARP estimate of the LANDSAT data.

5. Discussion

The WARPing algorithm was originally devised to handle the very large trivariate LANDSAT data. Exploring even the bivariate data in Figure 14 proved infeasible with ordinary kernel estimates. A bivariate histogram was a partial solution, but the MISE of the corresponding kernel estimate is an order of magnitude smaller for a sample of this size. The MISE advantage for the trivariate data is similar. The LANDSAT example illustrates the substantial difference in exploring large raw trivariate scatter diagrams and smooth functionals of those data.

WARPing has proven appropriate even with small multivariate data-sets such as the Iris example. The coupling of the estimation algorithm and the visualization scheme simplifies the understanding of the roles and interactions of statistical errors and (numerical) approximation errors. The WARP estimates carry along the interpolation machinery so no further consideration of numerical approximation errors of visualization is required.

Computational efficiency of the algorithms is acutely felt when hundreds of repetitions are required for bootstrapping, animations, or cross-validation. In the latter case, numerous convolutions are often required. WARPing simply substitutes discrete convolutions on rounded data for the exact calculations. Such an approach has been commonplace in time series analysis for many years. [Many spline-based estimators have exact representations for solutions with exactly n basis functions. Using rounded points could reduce the solution sizes as well; see Wahba (1990).]

For data in more than four dimensions, a two-stage analysis is often most appropriate. As in the LANDSAT example, the high dimensional data are projected to four or fewer dimensions. The projected data are then analyzed using a WARP estimator. In many cases, WARPing can be used in the projection stage.

Our goal is to encourage the use of the growing array of nonparametric ideas by stimulating the development of computationally feasible algorithms such as those provided by the WARPing framework.

References.

Badhwar, G.G. (1980). Crop Emergence Data Determination From Spectral Data. *Photogram. Eng. Remote Sens., 46, 369–377.*

Becker, R.A. and Chambers, J.M. (1984). S: An Interactive Environment for Data Analysis and Graphics. *Wadsworth, Belmont, California.*

Breiman, L. and Friedman, J.H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association, 80, 580–619.*

Breuer, K. (1990). Approximation von Kernglättern durch die WARPing-Methode. *Diploma Thesis.*

Broich, Th., Härdle, W. and Krause, A. (1990). XploRe — A computing environment for eXploRatory Regression. *Springer-Verlag, Berlin.*

Carroll, J. and Härdle, W. (1989). Symmetrized nearest neighbor regression estimates. *Statistics and Probability Letters , 7, 315–318.*

Cencov, N.N (1962). Evaluation of an Unknown Distribution Density from Observations. *Soviet Mathematics, 3, 1559–1562.*

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74, 829–836.*

Duan, N. and Li, K.-C. (1990). Slicing Regression: A Link-Free Regression Method *Annals of Statistics, in press.*

Eubank, R. (1988). Spline Smoothing and Nonparametric Regression. *Marcel Dekker, New York.*

Friedman, J.H. (1984). A Variable Span Smoother. *Tech Rep LCS5, Dept. of Statistics, Stanford.*

Friedman, J.H. and Stuetzle, W. (1981). Project Pursuit Regression. *Journal of the American Statistical Association, 76, 817–823.*

Härdle, W. (1987). Resistant Smoothing using the Fast Fourier Transform. *Appl. Statistics, AS 222, 36, 104–111.*

Härdle, W. (1990). Applied Nonparametric Regression. *Econometric Society*

Monograph Series, Cambridge University Press, in press

Härdle, W. and Bowman, A. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association, 83, 102–110.*

Härdle, W., Hall, P., and Marron, J.S. (1988). How Far are Automatically Chosen Regression Smoothing Parameters from Their Optimum? (with discussion). *Journal of the American Statistical Association, 83, 86–101.*

Härdle, W. and Marron, J.S. (1990). Bootstrap Simultaneous Error Bars for Nonparametric Regression. *CORE D.P. 8923, Universite Catholque de Louvain, Belgium.*

Härdle, W. and Stoker, T. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association, 84, 986–995.*

Hastie, T. and Tibshirani, R. (1987). Generalized Additive Models: Some Applications *Journal of the American Statistical Association, 82, 371-386.*

Heckman, N. (1986). Spline Smoothing in a Partly Linear Model *Journal of the Royal Statistical Society, Series B, 48, 244–248.*

Hjort, N. (1986). On Frequency Polygons and Average Shifted Histograms in Higher Dimensions. *TR 22, Dept. of Statistics, Stanford.*

Huber, P.J. (1985). Projection Pursuit. *Annals of Statistics, 13, 435–475.*

Jones, M.C. (1989). Discretised and Interpolated Kernel Density Estimates. *Journal of the American Statistical Association, 84, 733-741.*

Kallieris, D., Mattern, R., and Härdle W. (1986). Belastbarkeitsgrenze und Verletzungsmechanik des angegurteten PKW-Insassen beim Seitenaufprall. Phase II: Ansätze zur Verletzungsprädiktion. *FAT Schriftenreihe 60, Forschungsvereinigung Automobiltechnik e.V. (FAT).*

Lorensen, W.E. and Cline, H.E. (1987). Marching Cubes: A High Resolution 3D Surface Reconstruction Algorithm. *ACM Computer Graphics, 21, 163–169.*

Müller, H.-G. (1988). Nonparametric Regression Analysis of Longitudinal Data *Springer–Verlag, New York.*

Parzen, E. (1979). Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association, 74, 105–131.*

Priestley, M.B. and Chao, M.T. (1972). Non-parametric Function Fitting. *Journal of the Royal Statistical Society, Series B, 34, 385–392.*

Rice, J.A. (1986). Convergence Rates for Partially Splined Models. *Statistics and Probability Letters, 4, 203–208.*

Scott, D.W. (1983). Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions. *Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing, pp. 387–397.*

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

Scott, D.W. (1985b). Classification Using Multivariate Nonparametric Density Estimation. *Proceedings of the Sixth Annual National Computer Graphics Association Conference, Volume III, 715–718.*

Scott, D.W. (1986). Data Analysis in 3 and 4 Dimensions With Nonparametric Density Estimation *in Statistical Image Processing, E.J. Wegman and D. DePriest, Eds., Marcel Dekker, New York, pp. 291–305.*

Scott, D.W. (1988). Software for Cross-Validation of Density Estimates. *Rice Technical Report 88-8-311.*

Scott, D.W. and Hall, M.R. (1989). Interactive Multivariate Density Estimation in the S Language. *Proceedings of the 20th Interface of Computer Science and Statistics, American Statistical Association, Alexandria, Virginia, pp. 241–245.*

Scott, D.W. and Jee, R. (1984). Nonparametric Analysis of Minnesota Spruce and Aspen Tree Data and Landsat Data *Proceedings of the Second Symposium on Mathematical Pattern Recognition and Image Analysis, Dept. of Math, Texas A&M University, pp. 27–49*

Scott, D.W. and Schmidt, H.-P. (1988). Calibrating Histograms with Applications to Economic Data. *Empirical Economics, 13, 155–168.*

Scott, D.W. and Sheather, S.J. (1985). Kernel Density Estimation with Binned

Data. *Communications in Statistics, 14, 1353–1359.*

Scott, D.W., Tapia, R.A., and Thompson, J.R. (1980). Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria. *Annals of Statistics, 8, 820–832.*

Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased Cross-Validation in Density Estimation. *Journal of the American Statistical Association, 82, 1131–1146.*

Scott, D.W. and Thompson, J.R. (1983). Probability Density Estimation in Higher Dimensions. *Proceedings of the 15th Symposium on the Interface of Computer Science and Statistics, J.E. Gentle, Ed., North-Holland, Amsterdam, pp. 173–179.*

Scott, D.W. and Wand, M.P. (1990). Feasibility of Multivariate Density Estimates. *To appear, Biometrika.*

Silverman, B.W. (1982). Kernel Density Estimation Using the Fast Fourier Transformation. *Applied Statistics, 31, 93–97.*

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

Spiegelman, C.H. (1976). Two Techniques for Estimating Treatment Effects in the Presence of Hidden Variables: Adaptive Regression and a Solution of Reiersols's Problem *Unpublished Ph.D. Thesis, Northwestern University, Dept. Mathematics.*

Stone, C.J. (1977). Consistent Non-parametric Regression. *Annals of Statistics, 5, 595–645.*

Stone, C.J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Annals of Statistics, 10, 1040–1053.*

Stone, C.J. (1985). Additive Regression and other Nonparametric Models. *Annals of Statistics, 13, 689–705.*

Wahba, G. (1990). Spline Models for Observational Data *SIAM, Philadephia*

XploRe (1990). See Broich, Härdle and Krause, 1990.

Smoothing by Weighted Averaging of Rounded Points

Wolfgang K. Härdle [1,2,4]
David W. Scott [1,3]

C.O.R.E.
Université Catholique de Louvain
34 Voie du Roman Pays
B-1348 Louvain-la-Neuve

and

Department of Statistics
Rice University
Houston, Texas 77251-1892

[1,2,3,4]Research supported by (1) Sonderforschungsbereich 303, Deutsche Forschungsgemeinschaft; (2) The Center for the Study of Institutions and Values, School of Social Sciences, Rice University; (3) the Office of Naval Research and Army Research Office under contracts N00014-90-J-1176 and DAAL03-88-K-0131, respectively; and (4) CORE, Université Catholique de Louvain.

Abstract

Nonparametric smoothing techniques are generating much interest not only among theoretical statisticians but among applied workers in biostatistics, economics, and engineering. The benefits of this more flexible method come at the cost of greater computation. In higher dimensions, the computational burden can also be enormous when resampling methods for confidence intervals are used. One idea for reduction of computational cost is to do a data compression. In the case of multivariate density estimation, for example, the averaged shifted histogram is such an algorithm with significantly reduced computational effort. The ideas of the averaged shifted histogram algorithm can be extended to other nonparametric estimation problems such as regression and also to algorithms for additive modeling of high dimensional surfaces. In this paper the common framework for the so-called *Weighted Averaging of Rounded Points (WARPing)* is presented in these situations and examples are given with real data from LANDSAT observations and from a study of binomial response variables. The reduction of computational cost is discussed versus the loss in statistical efficiency.

Keywords

Smoothing, Nonparametric Density Estimation, Nonparametric Regression, Binning, WARPing, Generalized Additive models.

Contents

Figure Captions.

Figure 1. Six histograms with bin width h=13.5 and bin origins x_0=0,2.5,5,7.5,10,12.5 for the Buffalo snowfall data.

Figure 2. Construction of the WARP density estimate of Buffalo snowfall data with M=1,2,4,6,8,16 using the biweight kernel. Bandwidth h fixed.

Figure 3. WARPing approximations (solid lines) and the Nadaraya–Watson kernel estimator (dashed line) for the Geyser data-set. Bandwidth h=0.8, M=15.

Table 1. Operation counts for some univariate nonparametric density estimators.

Table 2. Comparison of operation counts between brute force kernel smoothing in d dimensions and the WARP method.

Figure 4. Computational cost of the kernel smoother as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

Figure 5. Computational cost for WARPing as a function of n for R=1, h=0.27, M=1 (long dashed line), M=3 (short dashed line), M=9 (short-long dashed line), M=27 (solid line). M is here the number of grid points per h.

Figure 6. A draftsman's plot of the four side-impact variables. Made with XploRe (1990). X_1= AGE, X_2= VEL, X_3= ROSYM, Y= response.

Figure 7. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(S)}$ (dashed line) for a simulated example $m(x)=x\,sin(2\pi x)$, $x \in U(0,1)$, $\varepsilon \sim N(0,\sigma^2)$, σ^2=0.25, h=0.25, M=5, n= 100, K= Quartic.

Figure 8. $MSE(x)$ for $\hat{m}_h(x)$ (solid line) and $\hat{m}_M^{(P)}$ (short dashed line) together with bounds on the difference between the two MSE curves. $m(x){=}x\,sin(2\pi x)$, $x{\in}U(0{,}1)$, $\varepsilon{\sim}N(0{,}\sigma^2)$, $\sigma^2{=}0.25$, $h{=}0.25$, $M{=}5$, $n{=}100$, $K{=}$ Quartic.

Figure 9. WARP dummy estimates (biweight kernel) for the variables ($ROSYM$, $RUSYM$, $T12RM$)=(x,y,z) giving maximal acceleration (in g) at different regions of the thorax of the dummy's bodies. The contour is at a level 5% of the modal level.

Figure 10. WARP estimates (as in Figure 9) for a different group.

Figure 11. Fifth step of the local scoring algorithm applied to the side-impact data. The info box shows the decrease in deviance as the iteration proceeds. Behind the info box one sees the current estimate $\sum_{j=1}^{d} \hat{g}_j(x_j)$ plotted versus Y with the link function G.

Figure 12. The WARP estimated function, $\hat{g}_j(x_j)$, $j{=}1{,}2{,}3$ for the side-impact example.

Figure 13. Histograms of the 3 channels of transformed LANDSAT data.

Figure 14. Pairwise scatter diagrams of the 3 channels of transformed LANDSAT data.

Figure 15. Density shell $S_{20\%}$ for the WARP estimate of the LANDSAT data.

Figure 16. Density shell $S_{10\%}$ for the WARP estimate of the LANDSAT data.

Figure 17. Density shell $S_{5\%}$ for the WARP estimate of the LANDSAT data.

Figure 18. Density shell $S_{2\%}$ for the WARP estimate of the LANDSAT data.