Smoothing by Local Regression: Principles and Methods

William S. Cleveland and Clive Loader

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.

Summary

Local regression is an old method for smoothing data, having origins in the graduation of mortality data and the smoothing of time series in the late 19th century and the early 20th century. Still, new work in local regression continues at a rapid pace. We review the history of local regression. We discuss four of its basic components that must be chosen in using local regression in practice — the weight function, the parametric family that is fitted locally, the bandwidth, and the assumptions about the distribution of the response. A major theme of the paper is that these choices represent a modeling of the data; different data sets deserve different choices. We describe *polynomial mixing*, a method for enlarging polynomial parametric families. We introduce an approach to adaptive fitting, *assessment of parametric localization*. We describe the use of this approach to design two adaptive procedures: one automatically chooses the mixing degree of mixing polynomials at each x using cross-validation, and the other chooses the bandwidth at each x using C_p. Finally, we comment on the efficacy of using asymptotics to provide guidance for methods of local regression.

Keywords: Nonparametric regression; Loess; Bandwidth; Polynomial order; Polynomial mixing; Ridge regression; Statistical graphics; Diagnostics; Modeling

1 Introduction

1.1 Modeling

Local regression is an approach to fitting curves and surfaces to data by smoothing: the fit at x is the value of a parametric function fitted only to those observations in a neighborhood of x (Woolhouse, 1870; Spencer, 1904a; Henderson, 1916; Macaulay, 1931; Kendall, 1973; Kendall and Stuart, 1976; Stone, 1977; Cleveland, 1979; Katkovnik, 1979; Friedman and Stuetzle, 1982; Hastie and Tibshirani, 1990; Härdle, 1990; Hastie and Loader, 1993; Fan, 1993).

The underlying model for local regression is

$$E(y_i) = f(x_i),\ i = 1, \ldots, n,$$

where the y_i are observations of a response and the d-tuples x_i are observations of d independent variables that form the *design space* of the model. The distribution of the y_i, including the means, $f(x_i)$, are unknown. In practice we must first *model* the data, which means making certain assumptions about f and other aspects of the distribution of the y_i. For example, one common distributional assumption is that the y_i have a constant variance. For f, it is supposed that the function can be well approximated *locally* by a member of a parametric class, frequently taken to be polynomials of a certain degree. We refer to this as *parametric localization.*

Thus, in carrying out local regression we use a parametric family just as in global parametric fitting, but we ask only that the family fit locally and not globally. Parametric localization is the fundamental aspect that distinguishes local regression from other smoothing methods such as smoothing splines (Wahba, 1990), regression splines with knot selection (Friedman, 1991), and wavelets (Donoho and Johnstone, 1994) although the notion is implicit in these methods in a variety of ways.

1.2 Estimation of f

The estimation of f that arises from the above modeling is simple. For each fitting point x, define a neighborhood based on some metric in the d-dimensional design space of the independent variables. Within this neighborhood, assume f is approximated by some member of the chosen parametric family. For example the family might be quadratic polynomials: $g(u) = a_0 + a_1(u - x) + a_2(u - x)^2$. Then, estimate the parameters from observations in the neighborhood; the local fit at x is the fitted function evaluated at x. Almost always, we will want to incorporate a weight function, $w(u)$, that gives greater weight to the x_i in the neighborhood that are close to x and lesser weight to those that are further.

The criterion of estimation depends on the assumptions made about the distribution of the y_i. For example, if we suppose that the y_i are approximately Gaussian with constant variance then it makes sense to base estimation on least-squares. If the parametric family consists of quadratic polynomials, we minimize

$$\sum_{i=1}^{n} w\left(\frac{x_i - x}{h}\right)(y_i - a_0 - a_1(x_i - x) - a_2(x_i - x)^2)^2.$$

In this case, $\hat{f}(x) = \hat{a}_0$. This is a pleasant case. The parameter estimates have a simple closed form expression, and numerical solution of the fitting equations is straightforward. Moreover, provided w and the neighborhood size do not depend on the response, the resulting estimate is a linear function of the y_i; this leads to a simple distribution theory that mimics very closely distributions for parametric fitting, so that t intervals and F-tests can be invoked (Cleveland and Devlin, 1988).

There are many possibilities for the specification of the *bandwidth*, h, for each point x in the design space. One choice is simply to make h a constant. This is *fixed bandwidth* selection; there is one parameter in this case, the halfwidth, h, of the neighborhood. Note that if we are fitting, say, quadratic polynomials locally, then as h gets large, the local regression fit approaches the global quadratic fit. Another is to have h change as a function of x based on the values of x_i. One example is *nearest-neighbor* bandwidth selection; in this case there is one parameter, α. If $\alpha \leq 1$, then we multiply n by α, round to an integer, k, and then take $h(x)$, the neighborhood halfwidth, to be the distance from x to the kth closest x_i. For $\alpha > 1$, $h(x)$ is the distance from x to the furthest x_i multiplied by $\alpha^{1/d}$ where d is the number of numeric independent variables involved in the fit. Thus, as with fixed bandwidth selection, as the bandwidth parameter gets large, the local fit approaches a global parametric fit.

There are many possibilities for the parametric family that is fitted locally. The most common is polynomials of degree p. Later in the paper we will describe polynomial mixing, which enlarges the families of polynomials, replacing the usual integer degree by a continuous parameter, the mixing degree.

Often, we can successfully fit curves and surfaces by selecting a single parametric family for all x; for example we might take p to be 2 so we are fitting quadratics locally. And, we can use a single value of a bandwidth parameter such as α, the nearest-neighbor parameter. Or, we can use an *adaptive* procedure that selects the degree locally as a function of x, or that selects the bandwidth locally, or that selects both locally.

1.3 Modeling the Data

To use local regression in practice, we must choose the weight function, the bandwidth, the parametric family, and the fitting criterion. The first three choices depend on assumptions we make about the behavior of f. The fourth choice depends on the assumptions we make about other aspects of the distribution of the y_i. In other words, as with parametric fitting, we are modeling the data.

But we do not need to rely fully on prior knowledge to guide the choices. We can use the data. We can use graphical diagnostic tools such as coplots, residual-dependence plots, spread-location plots, and residual quantile plots (Cleveland, 1993). To model f we can use more formal model selection criteria such as C_p (Mallows, 1973) or cross-validation (Stone, 1974). For example, Cleveland and Devlin (1988) use C_p to select the bandwidth parameter of nearest-neighbor fitting.

Modeling f usually comes down to a trade-off between variance and bias. In some applications, there is a strong inclination toward small variance, and in other applications, there is a strong inclination toward small bias. The advantage of model selection by a criterion is that it is automated. The disadvantage is that it can easily go wrong and give a poor answer in any particular application. The advantage of graphical diagnostics is great power; we can often see where in the space of the independent variables that bias is occurring and where the variability is greatest. That allows us to decide on the relative importance of each. For example, underestimating a peak in a surface or curve is often quite undesirable and so we are less likely to accept lower variance if the result is peak distortion. But the disadvantage of graphical diagnostics it that they are labor intensive, so while they are excellent for picking a small number of model parameters they are not practical for adaptive fitting, which as a practical matter requires an automated selection criterion. For example, Friedman and Stuetzle (1982) use cross-validation to choose the bandwidth locally for each x. And later in this paper we describe two adaptive procedures, one based on C_p and the other based on cross-validation. Still, when we have a final adaptive fit in hand, it is critical to subject it to graphical diagnostics to study its performance.

The important implication of these statements is that the above choices must be tailored to each data set in practice; that is, the choices represent a modeling of the data. It is widely accepted that in global parametric regression there are a variety of choices that must be made — for example, the parametric family to be fitted and the form of the distribution of the response — and that we must rely on our knowledge of the mechanism generating the data, on model selection diagnostics, and on graphical diagnostic methods to make the choices. The same is true for smoothing.

Cleveland (1993) presents many examples of this modeling process. For example, in one application, oxides of nitrogen from an automobile engine are fitted to the equivalence ratio, E, of the fuel and the compression ratio, C, of the engine. Coplots show that it is reasonable to use quadratics as the local parametric family but with the added assumption that given E the fitted f is linear in C. (It is quite easy to achieve such a *conditionally parametric* fit; we simply ignore C is defining the weights $w((x_i - x)/h)$. In addition, residual diagnostic plots show that the distribution of the errors is strongly leptokurtic, which means that we must abandon least-squares as a fitting criterion and use methods of robust estimation.

Later in the paper we will present other examples to further illustrate and explain the modeling process.

1.4 Why Local Regression?

Local regression has many strengths, some discussed in detail by Hastie and Loader (1993):

1) Adapts well to bias problems at boundaries and in regions of high curvature.

2) Easy to understand and interpret.

3) Methods have been developed that provide fast computation for one or more independent variables.

4) Because of its simplicity, can be tailored to work for many different distributional assumptions.

5) Having a local model (rather than just a point estimate $\hat{f}(x)$) enables derivation of response adaptive methods for bandwidth and polynomial order selection in a straightforward manner.

6) Does not require smoothness and regularity conditions required by other methods such as boundary kernels.

7) The estimate is linear in the response provided the fitting criterion is least squares and model selection does not depend on the response.

Singly, none of these provides a strong reason to favor local regression over other smoothing methods such as smoothing splines, regression splines with knot selection, wavelets, and various modified kernel methods. Rather, it is the combination of these issues that combine to make local regression attractive.

1.5 The Contents of the Paper

The history of local regression is reviewed in Section 2. Then, polynomial mixing is described in Section 3. In the next sections, the four choices required for carrying out local regression are discussed. Section 4 discusses an approach to making judgments about the efficacy of methods of local regression. Section 5 discusses the design of the weight function; the choice here is straightforward since there are just two main desiderata that apply in almost all applications in which the underlying dependence is described by a continuous f, and a single weight function can serve for most of these applications.

Section 6 discusses fitting criteria; there is little to say here because the considerations in making distributional assumptions are the same as those for global parametric regression. Section 7 discusses bandwidth selection and the choice of the local parametric family; the issues here are complex in that there are many potential paths that can be followed in any particular application and we must inevitably rely on the data to help us choose a path. To help convey the salient points made in Section 7, we present three examples in Section 8. In Section 9 we present two adaptive methods; one chooses the bandwidth locally as a function of x and the other chooses the mixing polynomial degree locally as a function of x. In Section 10 we discuss asymptotic theory, in particular, what cannot be determined from asymptotic theory in its current state. Finally, Section 11 summarizes the conclusions drawn in the paper.

2 An Historical Review

2.1 Early Work

Local regression is a natural extension of parametric fitting, so natural that local regression arose independently at different points in time and in different countries in the 19th century. The setting for this early work was univariate, equally spaced x_i; this setting is simple enough that good-performing smoothers could be developed that were computationally feasible by hand calculation. (In our discussion in this subsection we will set $x_i = i$ for $i = 1$ to n.) Also, most of the 19th century work arose in actuarial studies; mortality and sickness rates were smoothed as a function of age.

Hoem (1983) reports that smoothing was used by Danish actuaries as early as 1829, but their methods were not published for about 50 years. Finally, Gram (1883) published work from his 1879 doctoral dissertation on local polynomial fitting with a uniform weight function and with a weight function that tapers to zero. In much of his work he focused on local cubic fitting and used binomial coefficients for weights.

Stigler (1978) reports that in the United States, De Forest (1873, 1874) used local polynomial fitting for smoothing data. De Forest also investigated an optimization problem similar to one studied later by Henderson (1916), which we will describe shortly. Much of De Forest's work focused on local cubic fitting.

In Britain, work on smoothing had begun by 1870 when Woolhouse (1870) published a method based on local quadratic fitting. The method received much discussion but was eventually eclipsed by a method of Spencer (1904a) that became popular because it was computationally efficient and had good performance.

Spencer developed smoothers with several different bandwidths. His 21-point rule, which yields smoothed values for $i = 11, \ldots, n-10$ has the representation

$$\frac{1}{350}[5][5][7][-1, 0, 1, 2].$$

This notation means the following: first, take a symmetric weighted moving average of length 7 with weights $-1, 0, 1, 2, 1, 0, -1$; then take three unweighted moving sums of lengths 7, 5, and 5; and then divide by 350. The resulting fit at i is

$$\sum_{k=-10}^{k=10} c_k y_{i+k}$$

where the c_k are symmetric about $k = 0$ and the values of $350c_k$ for $k = -10$ to 0 are

$$-1, -3, -5, -5, -2, 6, 18, 33, 47, 57, 60.$$

We note three crucial properties. First, the smoother exactly reproduces cubic polynomials. Second, the smoothing coefficients are a smooth function of k and decay smoothly to zero at the ends. Third, the smoothing can be carried out by applying a sequence of smoothers each of which is simple; this was done to facilitate hand computation. Achieving all three of these properties is remarkable; Spencer and others spent some considerable effort in deriving such *summation formulas* of various lengths to satisfy the properties.

One might ask why we have put a summation formula such as Spencer's 21-point rule in the category of local fitting. The answer was first provided by an interesting paper of Henderson (1916) on weighted local cubic fitting. Let w_k be the weight function for $k = -m, \ldots, m$. Henderson showed that the local cubic fit at i can be written as

$$\sum_{k=-m}^{m} \phi(k) w_k y_{i+k}$$

where ϕ is a cubic polynomial whose coefficients have the property that the smoother reproduces the data if they are a cubic. (If w_k is symmetric then ϕ is quadratic.) Henderson also showed the converse: if the coefficients of a cubic-reproducing summation formula $\{c_k\}$ have no more than three sign changes, then the formula can be represented as local cubic smoothing with weights $w_k > 0$ and a cubic polynomial $\phi(k)$ such that $\phi(k)w_k = c_k$. For example, for Spencer's 21 term summation formula we can take

$$\phi(j) = (30 - j^2)/175$$

and the weight function, from $k = -10, \ldots, 0$ to be

$$\frac{1}{140}, \frac{1}{34}, \frac{5}{68}, \frac{5}{38}, \frac{1}{6}, \frac{3}{5}, \frac{9}{14}, \frac{11}{14}, \frac{47}{52}, \frac{57}{58}, 1.$$

1000 times these values are

$$7, 29, 74, 132, 167, 600, 643, 785, 904, 923, 1000.$$

Henderson also considered the problem of obtaining the smoothest possible fit subject to reproduction of cubics. Smoothness is measured by the sum of squares of the third differences of the smoother weights, or equivalently, the sum of squares of the third differences of the fit. The closed form solution for the smoother coefficients $c_k, k = -(m-2), \ldots, (m-2)$ is

$$a_m((m-1)^2 - x^2)(m^2 - x^2)((m+1)^2 - x^2)((3m^2 - 16) - 11x^2)$$

where a_m is a term that makes the weights add to 1. From the result in the previous paragraph, this summation formula is equivalent to local cubic fitting with the neighborhood weight function

$$w_k = ((m-1)^2 - k^2)(m^2 - k^2)((m+1)^2 - k^2)$$

for $|k| \leq m-2$. For large m, this amounts to the triweight weight function $(1-x^2)^3$. Thus, asymptotic optimality problems of the type considered in Müller (1984) were, for equally spaced x_i, solved exactly by Henderson for finite samples.

The work that became well known was that of Henderson (from the U.S.) and the British smoothing research community — which included accomplished applied mathematicians such as Whittaker (1923), who, along with Henderson (1924), also invented smoothing splines. The influence resulted in the movement of local fitting methods into the time series literature. For example, the book *The Smoothing of Time Series* (Macaulay 1931), shows how local fitting methods can be applied to good purpose to economic series. Macaulay not only reported on the earlier local fitting work, but also developed methods for smoothing seasonal time series.

Macaulay's book in turn had a substantial influence on what would become a major milestone in local fitting methods. It began at the U.S. Bureau of the Census, beginning in 1954. A series of computer programs were developed for smoothing and seasonal adjustment of time series, culminating with the X-11 method (Shishkin, Young, and Musgrave, 1967). This represented one of the earliest uses of computer-intensive statistical methods, beginning life on an early Univac. X-11 did not widely penetrate the standard statistics literature at the time because its methods were empirically based rather than emanating from a fully specified statistical model. However, X-11 became the standard for seasonal and trading-day adjustment of economic time series and is still widely used today. In X-11, a time series is fitted by three additive components: a trading day component described by a parametric function of day-of-week variables and fitted by parametric regression; a trend component fitted by smoothing, and a seasonal component fitted by smoothing. The developers of X-11 used what would become known two decades later as the backfitting algorithm (Friedman and Stuetzle, 1981) to iteratively estimate the components. These iterations are nested inside iterations that provide robust fitting by identifying outliers and modifying the data corresponding to the outliers. Thus X-11 at its inception employed semi-parametric additive models, backfitting, and robust estimation two to three decades before these methods became commonplace in statistics.

While the early literature on smoothing by local fitting focused on one independent variable with equally-spaced values, much intuition that was built up about smoothing methods remain valid for smoothing as a function of scattered multivariate measurements. We will invoke this intuition later in the paper.

2.2 Modern Work

The modern view of smoothing by local regression has origins in the 1950's and 1960's, with kernel methods introduced in the density estimation setting (Rosenblatt, 1956; Parzen, 1962) and the regression setting (Nadaraya, 1964; Watson, 1964). This new view extended smoothing as a function of a single independent variable with equally spaced measurements to smoothing as a function of scattered measurements of one or more independent variables. Kernel methods are a special case of local regression; a kernel method amounts to choosing the parametric family to consist of constant functions.

Recognizing the weaknesses of a local constant approximation, the more general local regression enjoyed a reincarnation beginning in the late 1970's (Stone, 1977; Cleveland, 1979; Katkovnik, 1979). The method can also be found in other branches of scientific literature; for example numerical analysis (Lancaster and Salkauskas, 1981 and 1986). Furthermore, while the early smoothing work was based on an assumption of a near-Gaussian distribution, the modern view extended smoothing to other distributions. Brillinger (1977) formulated a general approach and Cleveland (1979) and Katkovnik (1979) developed robust smoothers. Later, Tibshirani and Hastie (1987) substantially extended the domain of smoothing to many distributional settings such as logistic regression and developed general fitting algorithms. The extension to new settings continues today (Fan and Gijbels, 1994a; Loader, 1995)).

Work on local regression has continued throughout the 1980's and 1990's. A major thrust has been the application of smoothing in multidimensional cases. Here, numerous approaches can be taken: Cleveland and Devlin (1988) apply local linear and quadratic fitting directly to multivariate data. Friedman and Stuetzle (1981) use local linear regression as a basis for constructing projection pursuit estimates. Hastie and Tibshirani (1990) use local regression in additive models. These methods have substantial differences in data requirements and the types of surface that can be successfully modeled; the use of graphical diagnostics to help make decisions becomes crucial in these cases.

Accompanying the modern current of work in smoothing was a new pursuit of asymptotic results. It began in the earliest papers (e.g., Rosenblatt, 1956; Stone, 1977) and then grew greatly in intensity in the 1980s (e.g., Müller, 1987; Härdle, 1990; Fan, 1993; Ruppert and Wand 1994). In Section 10 we comment on the role of asymptotics in local regression.

3 Polynomial Mixing

The most common choice for the local parametric family is polynomials of degree p. However, the change from degree 1, say, to degree 2 in many applications often represents a substantial change in the results. For example, we might find that degree 1 fitting distorts a peak in the interior of the configuration of observations of the independent variables, and degree 2 removes the distortion but results in undue flopping about at the boundary. In such cases we can find ourselves wishing for a compromise. Polynomial mixing can provide such a compromise. The idea has been used in global parametric fitting (Mallows, 1974). The description in this paper of its use for local regression is from Cleveland, Hastie, and Loader (1995).

The mixing degree, p, is a nonnegative number. If p is an integer then the mixed fit is simply the local polynomial fit of degree p. Suppose p is not an integer and $p = m + c$ where m is an integer and $0 < c < 1$. Then the mixed fit is simply a weighted average of the local polynomial fits of degrees m and $m + 1$ with weight $1 - c$ for the former and weight c for the latter. It is easy to see that this amounts to a local ridge regression of a polynomial of degree $m+1$ with zero ridge parameters for monomial terms except the term of degree $m + 1$. We can choose a single mixing degree for all x or we can build an adaptive method by letting p vary with x. Both approaches will be used in later sections of the paper.

4 To What Do We Look for Guidance in Making the Choices?

In the next sections we turn to making the choices necessary to carry out local regression — the weight function, the bandwidth, the fitting criterion, and the local parametric family. To what do we look for guidance to make the choices? The answer is, as we have emphasized, to treat choices of degree and bandwidth as modeling the data and use formal model selection criteria and graphical diagnostics to provide guidance. What happens through such a process is that we begin to build up a knowledge base about what tends to provide good models and what does not. In the end, it becomes in some sense a statement about the behavior of data rather than the performance of selection methods. We can never have anything quite as definitive as a theorem, but we can from this process get good guidance about what is likely to work as we approach a new set of data.

The development of methods of parametric regression has had a long history of using model selection criteria and diagnostic methods for the common parametric models fitted to regression data (Daniel and Wood, 1971). It is much to the credit of researchers in parametric regression that these methods have become part of the mainstream of statistical practice. Much reporting of what works in practice from this process has greatly strengthened such parametric fitting.

Attention to the needs of data that arise in practice, and an assessment of methods based on what works in practice was also an important part of the early smoothing literature discussed in Section 2. In the coming sections we will draw on this early literature.

5 Selecting the Weight Function

To begin, suppose the data to be analyzed have an f that is continuous, and suppose that if higher derivatives are not continuous then it is not necessary to produce estimates that reproduce the discontinuity.

In such a case we will almost always want to consider weight functions $W(u)$ that are peaked at $u = 0$, and that decay smoothly to 0 as u increases. One alternative is a rectangular weight function, or boxcar. With a boxcar, all observations within a distance h receive weight 1, and those further away receive weight 0. This results in a noisy estimate; as x changes, observations abruptly switch in and out of the smoothing window. A smooth weight function results in a smoother estimate. This was widely appreciated in the early smoothing literature. For example, Macaulay (1931) writes:

> A smooth weight diagram leads to smoothness in the resulting graduation because smoothing by means of any weighted or unweighted moving average amounts to distributing each observation over a region as long as the weight diagram and of the same shape as the weight diagram.

Second, we will almost always want to use a weight function that is nonzero only on a bounded interval rather than, for example, approaching zero as u gets large. The reason is computational speed; we can simply ignore observations with zero weight.

Given these two constraints, the choice is not too critical; for our examples — which are all cases with a smooth f — we use the tricube weight function, which is used in the loess fitting procedure (Cleveland, 1979; Cleveland and Devlin, 1988),

$$w(u) = \begin{cases} (1 - |u|^3)^3 & |u| < 1 \\ 0 & |u| > 1 \end{cases} .$$

For cases where f cannot be locally approximated by polynomials, it can sometimes be helpful to consider quite different weight functions. One example is the case of discontinuous f, where one-sided kernels can be employed (McDonald and Owen, 1986; Loader, 1993).

6 Selecting The Fitting Criterion

It turns out that virtually any global fitting procedure can be localized. Thus local regression can proceed with the same rich collection of distributional assumptions as have been used in global parametric fitting. And methods for making the choice can proceed as they have for global parametric fitting.

The simplest case, as we have discussed, is Gaussian y_i. An objection to least squares is lack of robustness; the estimates can be quite sensitive to heavy tailed residual distributions. When the data suggest such distributions we can use robust fitting procedures, for example, the iterative downweighting in loess (Cleveland, 1979; Cleveland and Devlin, 1988) implements local M-estimation. Robust fitting is also discussed in some detail in Tsybakov (1986). Another approach, taken by Hjort (1994), is local Bayesian regression.

Other error distributions lead to other fitting criteria (Brillinger, 1977; Hastie and Tibshirani, 1990; Staniswalis, 1988). For example, a double exponential distribution leads to local L_1 regression. Perhaps more interesting is the extension to generalized linear models; for example, binary data. Suppose

$$P(y_i = 1) = 1 - P(y_i = 0) = p(x_i).$$

Then the locally weighted likelihood is

$$\sum_{i=1}^{n} w(\frac{x_i - x}{h})(y_i\theta_i - \log(1 + e^{\theta_i})).$$

with $\theta_i = \log(p(x_i)/(1 - p(x_i)))$. It is sensible in cases like this to model the natural parameters — in this case, the θ_i — by local polynomials and then transform to the quantities of interest — $p(x_i)$.

In density estimation we observe $x_1, \ldots, x_n$ from a density $f(x)$. Loader (1995) uses the local likelihood

$$\sum_{i=1}^{n} w(\frac{x_i - x}{h}) \log f(x_i) - n \int_{\mathcal{X}} w(\frac{u - x}{h}) f(u) du$$

to model the density; it is natural to use polynomials for $\log f(x)$. See also Hjort and Jones (1994) for discussion and further generalizations.

7 Selecting the Bandwidth and Local Parametric Family

In this section we discuss the bandwidth and the local parametric family together because the choices interact strongly. A change from one parametric family to another can often have a dramatic effect on the sensible choice of bandwidth.

The goal in choosing the bandwidth and the local parametric family is to produce an estimate that is as smooth as possible without distorting the underlying pattern of dependence of the response on the independent variables. In other words, we want $\hat{f}$ to have as little bias as possible and as small a variance as possible. Usually, we need to strike a balance between the two, but with the right model selection tools and graphical diagnostics we can in many applications find a parametric family and select the bandwidth to satisfy the needs of the analysis.

Two methods of bandwidth specification are considered in this section — nearest-neighbor and fixed. These are both simple and easily implemented. (In Section 9 we consider adaptive bandwidth selection).

Finally, we will use either ordinary polynomials or mixed polynomials as the parametric family with degrees ranging from 0 to 3, and use a single degree for all x. (In Section 9 we consider adaptive selection of the mixing degree).

For many applications, particularly those with smooth f, they provide sufficient flexibility to get good fits.

7.1 Bandwidth Selection

Nearest neighbor bandwidths are widely used for local regression (e.g., Stone, 1977; Cleveland, 1979; Fan and Gijbels, 1994a). The reason is simple. A fixed-bandwidth estimate often has dramatic swings in variance due to large changes in the density of the data in the design space leading to unacceptably noisy fits; in extreme cases, empty neighborhoods lead to an undefined estimate.

Boundary regions play a major role in bandwidth choice. Suppose the independent variable x is distributed uniformly on $[0, 1]$. When estimating at 0, a bandwidth of h will cover only half as much data as the same bandwidth when estimating at 0.5; moreover, the data are all on one side of the fitting point in the boundary case. Using the same bandwidth at the boundary as the interior point clearly results in high variability. In practice, the situation is even worse, since the data density may be sparse at boundary regions, for example, when the x_i have a Gaussian distribution. And the situation can become much worse in two or more dimensions where most of the data in the design space can lie on the boundary. The boundary-region problem was well known to early researchers in smoothing. For example, Kendall (1973) writes the following about local polynomials:

> This is as we might expect: the nearer the tails, the less reliable is the trend point, as measured by the error-reducing power at that point. The fitted curve it has been said, tends to wag its tail.

While fixed bandwidth selection can provide good fits in many cases, it does not do so in many others, and nearest-neighbors appear to perform better overall in applications because of this variance issue. Of course, nearest-neighbor bandwidth selection can fail to model data in specific examples. But usually, if nearest-neighbor selection fails it is not fixed bandwidth selection that is needed to remedy the problem but rather adaptive methods. We will take this up in Section 9.

As stated earlier, we should approach any set of data with the notion that bandwidths can range from very small to very large. Even though most asymptotic work is based on an assumption of the bandwidth going to zero, quite large bandwidths are important for practice. For example, for some applications, using nearest neighbor bandwidth selection, we get the best fits for small values of α, say $1/10$ or smaller; this is often the case for a time series when the goal is to track a high frequency component. In other applications, the best fits are provided by large values of α, say, $3/4$ or higher.

One reaction might be that if a large bandwidth can be used then there is probably some global parametric function that fits the data. But this is not necessarily so. Large bandwidths can provide a large number of degrees of freedom and substantial flexibility in our fits. For example, for loess fitting — that is, local fitting with the tricube weight function, nearest-neighbor bandwidth selection with bandwidth parameter α, polynomial fitting of degree p, and d independent variables — the degrees of freedom of the fit is roughly $1.2\tau/\alpha$ where τ is the number the degrees of freedom of a global parametric fit of a polynomial of degree p (Cleveland, Grosse and Shyu, 1990). Thus the global parametric degrees of freedom is multiplied by $1.2/\alpha$ for a loess fit which means, for example, that if $\alpha = 3/4$, the global parametric degrees of freedom are multiplied by 1.6.

7.2 Polynomial Degree

The choice of polynomial degree — mixed polynomial degree or ordinary polynomial degree — is, like bandwidth, a bias-variance trade off. A higher degree will generally produce a less biased, but more variable estimate than a lower degree one.

Some have asserted that local polynomials of odd degree beat those of even degree; degree 1 beats degree 0, degree 3 beats degree 2, and so forth. But in applications, it is sensible to think of local regression models as ranging from those with small neighborhoods, which provide very local fits, to those with large neighborhoods, which provide more nearly global fits, to, finally, those with infinite neighborhoods, which result in globally parametric fits. In other words, globally parametric fitting is the limiting case of local regression. Thus it is obvious that we should not rule out fitting with even degrees; this is no more sensible than ruling out even degrees for global polynomial fitting.

There is, however, one degree that very infrequently proves to be the best choice in practice — degree zero, or locally constant fitting. This was widely appreciated in the early smoothing literature. The term often used was "moving average" or "moving average with positive weights". In fact, the standard was methods that preserved either quadratic or cubic polynomials; for our context this would mean degrees of two or higher. Macaulay (1931) writes the following about moving averages away from the boundary:

> For example, a simple moving average, if applied to data whose

> underlying trend is of a second-degree parabolic type, falls always *within* instead of *on* the parabola. ... A little thought or experimentation will quickly convince the reader that, so long as we restrict ourselves to *positive* weights, no moving average, weighted or unweighted, will exactly fit any mathematical curve except a straight line.

Exactly the same point made by Macaulay holds for local regression generally. For most applications, if a single degree for polynomial fitting is to be chosen for all x, careful modeling of the data seldom leads to locally constant fitting. The general problem is that since locally constant fitting cannot even reproduce a line except in special cases, for example, equally-spaced data away from the boundaries. Reducing the lack of fit to a tolerable level requires quite small bandwidths, producing fits that are very rough. By using a polynomial degree greater than zero we can typically increase the bandwidth by a large amount without introducing intolerable bias; despite the increased number of monomial terms in the fitting, the end result is far smoother because the neighborhood size is far larger.

In fact, local linear fitting often fails to provide a sufficiently good local approximation — when there is a rapid change in the slope, for example a local minimum or maximum — and local quadratic fitting does better. It is for this reason that in the early smoothing literature, the methods that were devised preserved at least quadratic functions because peaks and valleys in f are common in practice.

8 Examples

In this section, we will use three examples to demonstrate modeling the data through the choices of the weight function, the fitting criterion, the bandwidth, and the parametric family. We will use cross-validation, C_p, and graphical diagnostics to guide the modeling.

8.1 Ozone and Wind Speed Data

The first example is measurements of two variables: ozone concentrations and wind speed at ground level for 111 days in New York City from May 1 to September 30 of one year. We will take the cube roots of the ozone concentrations — this symmetrizes the distribution of the residuals — and model cube root ozone as a function of wind speed. We will fit with the following choices: the tricube weight function, nearest-neighbor bandwidth selection, least-squares, and polynomial mixing.

Figure 1 shows the cross-validation sum of absolute deviations against the mixing degrees for four values of α from 0.25 to 0.75 in equal percentage steps. Because of the rough approximation referred to above, equal percentage steps in α tend to make the degrees of freedom of fits for a fixed mixing degree change in equal percentage steps. Figure 2 shows fits. Each column contains fits for one value of α. The top row is the mixing fit with the minimum cross-validation score and the remaining rows show the fits for degrees 3 to 0. Figure 3 shows the residuals for each of the 30 fits in Figure 2. Superposed on each plot is a loess smooth with local linear fitting and $\alpha = 1/3$. Such residual plots provide an exceedingly powerful diagnostic that nicely complements a selection criterion such as cross-validation. The diagnostic plots can show lack of fit locally, and we have the opportunity to judge the lack of fit based on our knowledge of both the mechanism generating the data and our knowledge of the performance of the smoothers used in the fitting.

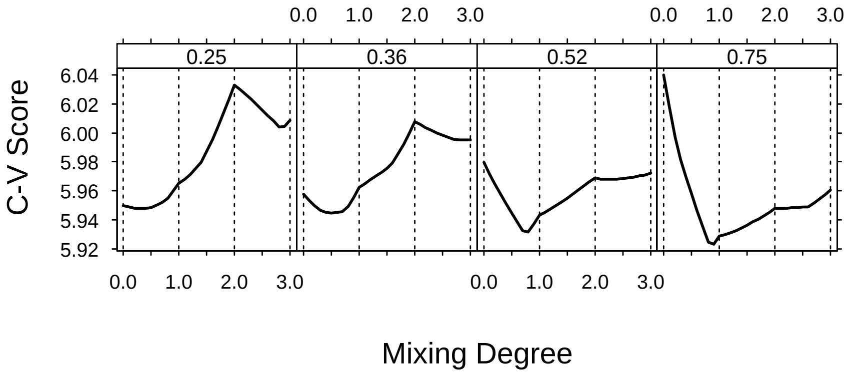


Figure 1: Cross-validation scores for the mixed fits as a function of p.

For the ozone and wind speed data there is a pronounced dependence of the response on the independent variable but overall there are not radical changes in curvature, in particular there are no peaks or valleys. We might expect that locally, a linear family provides a reasonable approximation. The diagnostics and cross-validation show this is the case. For local constant fitting, that is $p = 0$, a small α is needed to capture the dependence of ozone on wind speed without introducing undue distortion. Even for $\alpha = 0.25$, the plot of residuals suggests that there is lack of fit at the smallest values of wind speed, that is, at the left boundary, where there is a large slope. Local constant fitting cannot capture a linear effect at a boundary. The same distortion does not occur at the right boundary because the local slope is zero, and local constant fitting can cope with a zero slope. Despite the distortion, for $\alpha = 0.25$ the cross-validation criterion suggests that overall, local constant fitting is the best fit because for larger values of p the variability of the fit increases rapidly. But the local constant fit is itself quite noisy, far noisier than is reasonable. In other words, we do not have a satisfactory fit for $\alpha = 0.25$.

As we increase α to get a smoother fit, the local constant fit introduces a

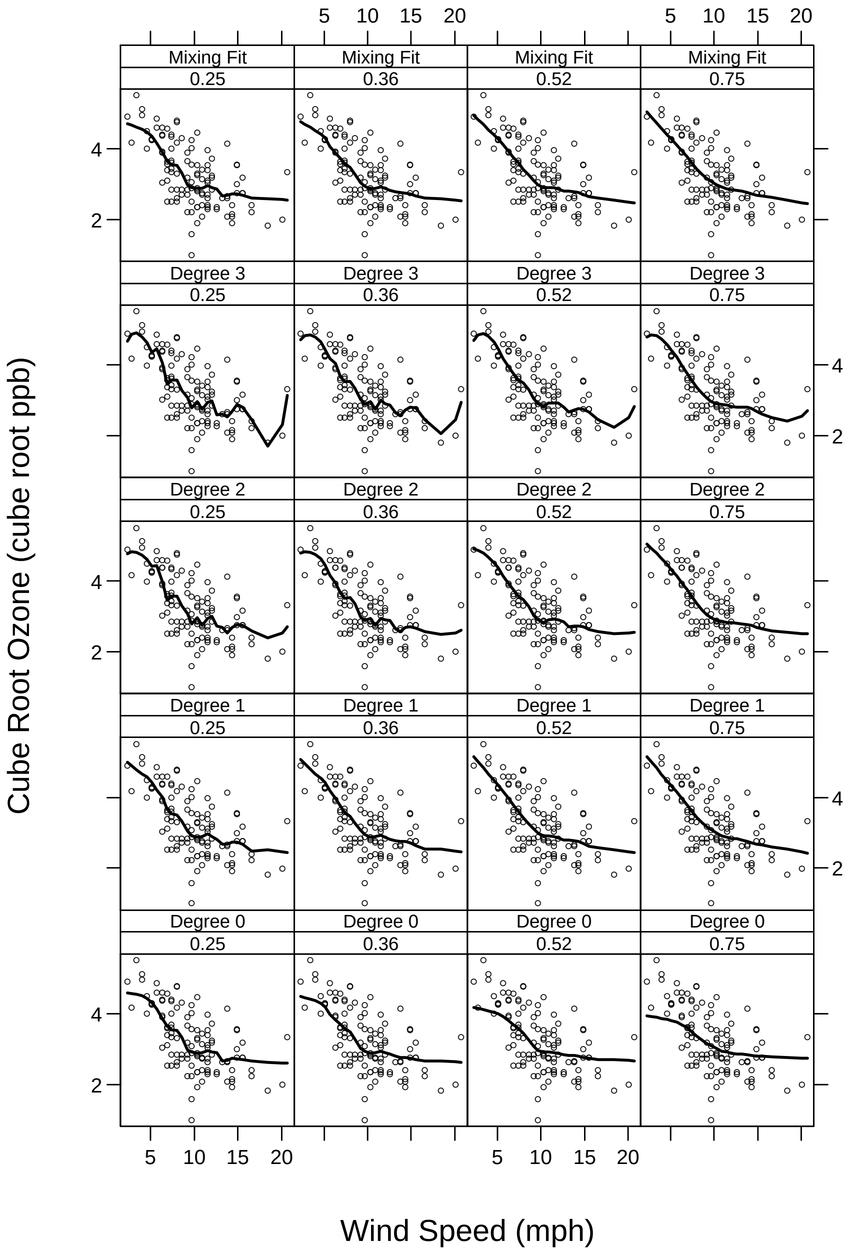


Figure 2: Fits for four nearest-neighbor bandwidths ($\alpha = 0.25$ to 0.75 in equal percentage steps) and five local fitting methods. The top row shows the fits for the mixing method with cross-validation and the remaining rows show local polynomial fits for degrees 3 to 0.

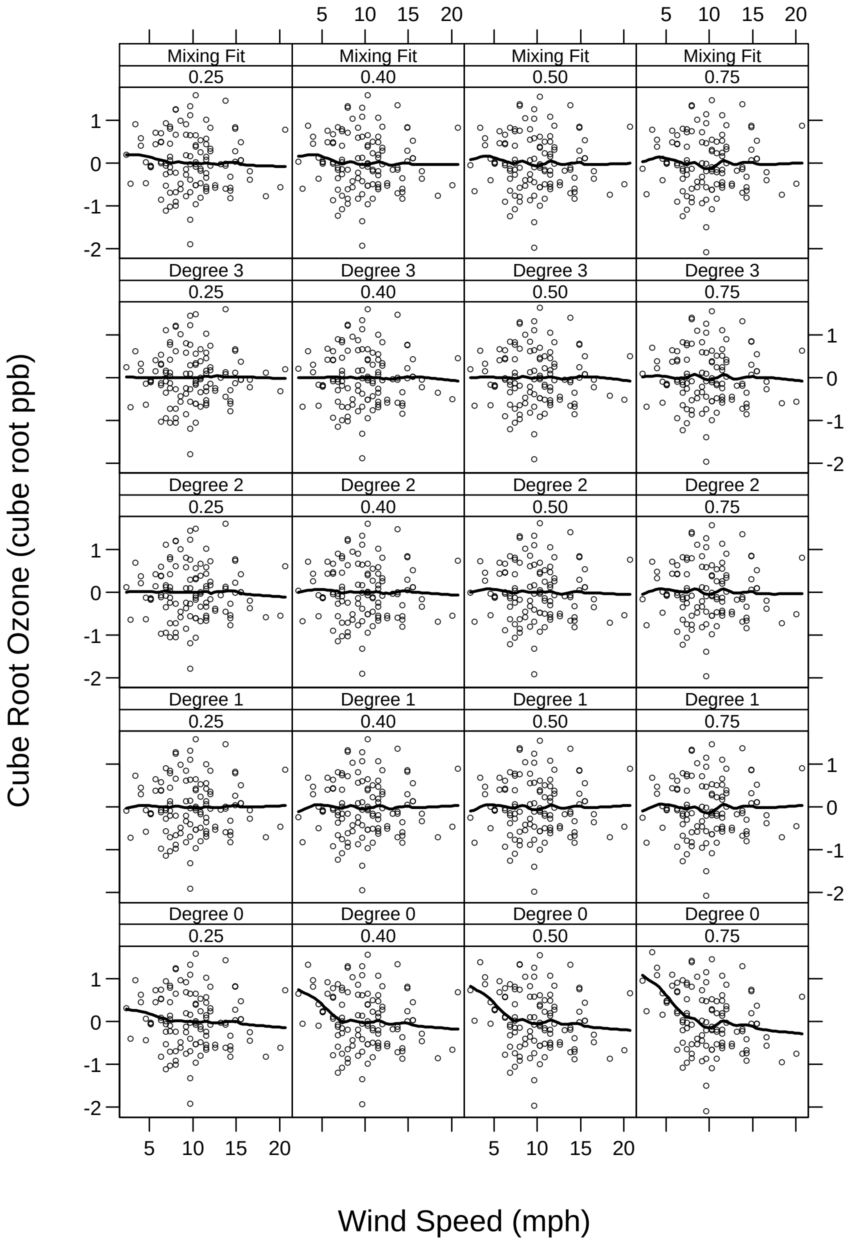


Figure 3: Residual plots for the fits in the previous figure.

major distortion, one that is so bad that cross-validation judges it to be the worst case, and the minimum cross-validation score occurs for p close to 1. For $\alpha = 0.75$, the minimum cross-validation score fit is quite smooth and the residual plot suggests there is no lack of fit.

8.2 An Example with Made-Up Data

We will now turn to data that we generated. To study smoothers, it is useful to use, in addition to real data, made-up data where the true model is known.

Take $n = 100$, $x_i \sim N(0,1)$, $f(x) = 2(x - \sin(1.5x))$ and $\epsilon_i \sim N(0,1)$. We consider smoothing for both fixed and nearest-neighbor selection using local polynomial fitting with degrees 0 to 3.

How should we compare smooths? Were this a real set of data we could use cross-validation or C_p. But in this case, since we know the model, we can compute the true mean-square error summed over the x_i. Thus the criterion is

$$\begin{aligned} R(\hat{f}, f) &= \sum_{i=1}^{n} E((\hat{f}(x_i) - f(x_i))^2) && (1) \\ &= \|(I-L)f\|^2 + \sigma^2 \mathrm{tr}(L^T L), && (2) \end{aligned}$$

where f is the vector of values of $f(x_i)$, $\hat{f}$ is the vector of values of $\hat{f}(x_i)$, and L is the hat matrix of the smoother. The final expression in the above equation decomposes the mean-square error into bias and variance.

To compare different degrees of smoothing — for example, local constant versus local quadratic — the smoothers must be placed on an equal footing. We cannot get meaningful results using the same bandwidth for each; a higher order fit is more variable but less biased. To gain useful insight, one must use equivalent amounts of smoothing for each of the methods under consideration. Thus, we consider the equivalent degrees of freedom of the smooth, which we define here as $\nu = \mathrm{tr}(L^T L)$. This is particularly convenient for $R(\hat{f}, f)$ since ν is, up to the factor σ^2, the variance component of (2). Hence, a plot of $R(\hat{f}, f)$ against ν displays the bias-variance tradeoff without being confounded with meaningless bandwidth effects.

Figure 4 shows the results for both nearest-neighbor and fixed bandwidths. Degrees ranging from 0 (local constant) to 3 (local cubic) are considered. Points on the right of these plots represent small bandwidths for which the fits have little bias but substantial variance.

Fits for the values of the smoothing parameters that minimize R are displayed in Figure 5. Local constant fitting is clearly unsatisfactory for both fixed and nearest-neighbor bandwidths. The smooths are very noisy because small bandwidths must be used to properly track the boundary effects. For local linear fitting, the improved boundary behavior enables larger bandwidths to be used, and a smoother curve results. However, even in this example with fairly modest curvature, the local quadratic and cubic fits perform better, with the fitted curves being substantially less noisy. Visually, there is not much difference between the fixed and nearest-neighbor fits for higher orders. The best fit, according to Figure 4, is the local quadratic fit with nearest-neighbor bandwidths, but fixed bandwidth selection performs well in this example.

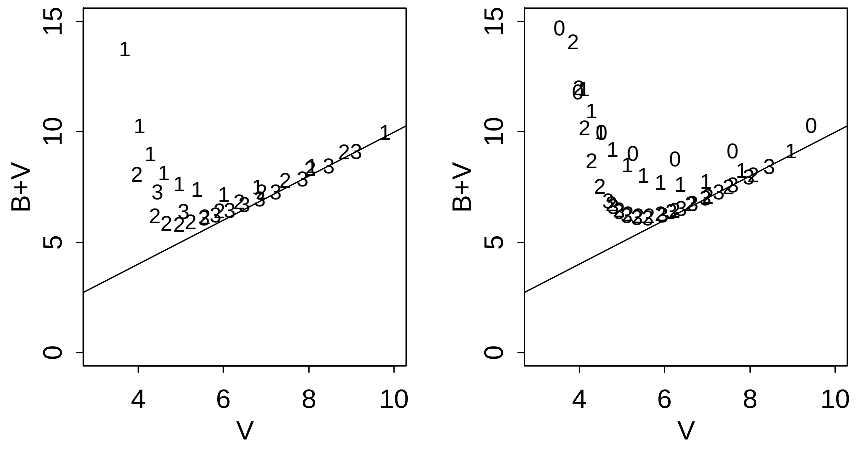


Figure 4: Plots of $R(\hat{f}, f)$ against $\mathrm{tr}(L^T L)$ for nearest neighbor bandwidths (left) and fixed bandwidths (right), for fitting of degrees 0 to 3. (Note 0 is entirely off the scale on the left figure).

It is important to reiterate that exclusive reliance in practice on a global criterion similar to that in the figure is unwise because a global criterion does not provide information about where the contributions to bias and variance are coming from in the design space. In many applications, high bias or high variance in one region may be more serious than in another. For example, a careful look at the right panels of Figure 5 reveals that for degrees one to three the y_i at the minimum and maximum values of the x_i have nearly zero residuals; the fits at these points are very nearly interpolating the data which results in unacceptably high variance. Nearest neighbor bandwidths go some way towards relieving the problem; this accounts for most of the advantage of nearest-neighbors observed in Figure 4 for local quadratic fitting.

As this example suggests, boundaries tend to dominate the fixed vs. nearest-neighbor comparison. Asymptotic comparisons provide no useful guidance in this case. For example, Gasser and Jennen-Steinmetz (1988) use the asymptotic characterization of the nearest-neighbor bandwidth as being proportional to the reciprocal of the density of the x_i in the design space; in our example $h(x) \approx \alpha/2\phi(x)$ where $\phi(x)$ is the standard normal density. At the left end-point $x = -2.736$ and $\alpha = 0.35$ for local linear fit-

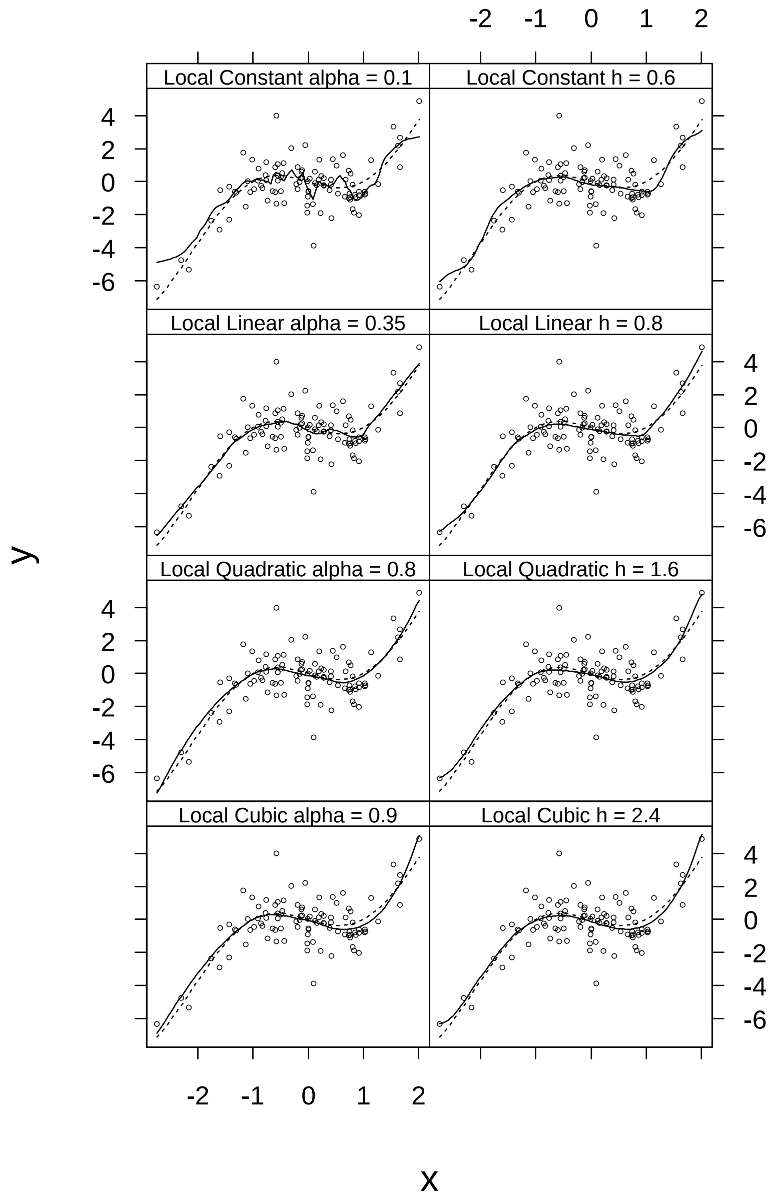


Figure 5: Local fits that minimize mean-square error for nearest-neighbor bandwidths (left) and fixed bandwidths (right). The solid curves are the estimates, and the dashed curves the true function.

ting. The asymptotic characterization yields $h = 18.5$ and all observations receive weights between 0.95 and 1! Clearly, this does not approximate reality, where only 34 observations receive non-zero weights, and many of the weights are small.

8.3 More on the Ozone Data

The ozone and wind speed data studied earlier are just two variables from a multivariate data set that has two other independent variables, temperature and solar radiation Cleveland (1993). The full dataset is 111 daily measurements of wind speed, temperature, radiation, and ozone concentration. Cleveland (1993) found that it is appropriate to fit models that are conditionally parametric in R and W. A C_p plot for this data is shown in Figure 6 for degrees 0 to 3 and a variety of bandwidths. Local constant fitting is vastly inferior to other degrees. Local linear also struggles; its minimum C_p is about 50% larger than for local quadratic. Local quadratic also slightly beats local cubic. The minimum C_p for the local quadratic fitting occurs at an α of 0.4; however, the C_p pattern is quite flat for quadratic fitting and the minimum C_p fit is fairly noisy, so Cleveland elected to use a larger α.

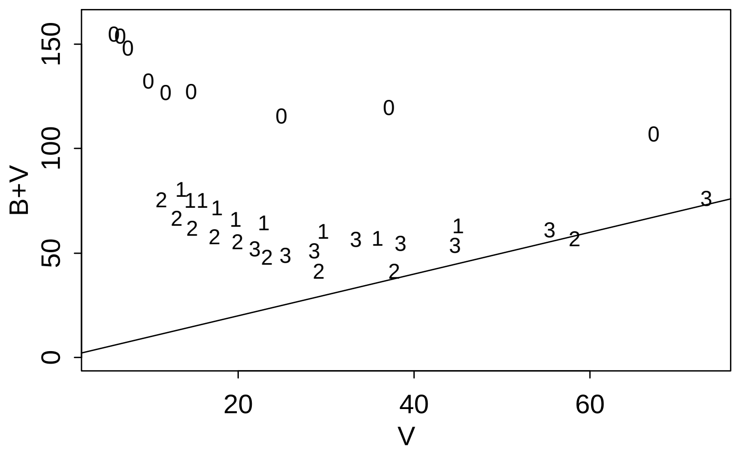


Figure 6: C_p plot for the environmental data, with trivariate predictors.

9 An Approach to Adaptive Fitting Based on Assessing Localization

In the previous sections we have discussed smoothing procedures for which the parametric family fitted locally does not change with x and for which the bandwidth selection parameters of fixed and nearest-neighbor smoothing do not change with x. But for some data sets, particularly those for which the amount of curvature in different broad regions of the design space is quite different, it can make sense to use adaptive methods. Many adaptive bandwidth methods have been suggested in the past (e.g., Friedman and Stuetzle, 1982). Varying the parametric family locally is less common (e.g., Fan and Gijbels, 1994b) but very promising.

We use a general approach to adaptive fitting based on an *assessment of parametric localization.* In Section 1, "parametric localization" was used to describe the basic approach that guides local regression — the approximation of a the true f locally by a parametric function. Using an assessment of localization, we can design methods for adaptive bandwidth selection (Cleveland and Loader, 1995) and adaptive parametric family selection (Cleveland, Hastie, and Loader, 1995).

The important principle is this. *For a particular x and a given neighborhood, the fit at x can be expected to perform well if the locally fitted parametric function adequately approximates the true function over the neighborhood.* Now one might object to this notion on the basis that it is possible for the fit at x to be good even if the fitted parametric function provides a poor approximation at places other than x; but further thought makes it clear that only in degenerate cases can we do well despite a poor approximation. For example, for equally-spaced data away from the boundary locally constant fitting will provide a good fit to a linear f. But this is solely due to the serendipitous result that local linear fitting in such a case happens to be the same as local constant fitting.

Now we will not assess the fit equally throughout the neighborhood. Instead we will weight our assessment based on the weight function used to compute the fit at x. In other words we will have greater tolerance for poor performance at positions far from x than at positions close to x.

With this notion of a weighted assessment of parametric localization as the approach, details of carrying it out are reasonably straightforward. For the fit at x we simply take an automatic selection criterion that we might have used for picking a global parameter such as the α for nearest-neighbors, and we make two changes. First, instead of applying the criterion over the x_i using the smoother, we the apply the criterion using the parametric function fitted at x. Second, we weight the criterion with the neighborhood weights used in the local fit at x. The next two examples illustrate this process.

9.1 Local Polynomial Mixing

We will now use the assessment of localization to develop an adaptive method for choosing the mixing degree for polynomial mixing. The method is discussed in detail by Cleveland, Hastie, and Loader (1995).

In Section 8 we used polynomial mixing and selected a single mixing degree for all x by the cross-validation sum of absolute deviations. We chose a value of p, fitted the mixed polynomial to get the fit $\hat{f}(x_i)$ at each x_i, and then studied the single-observation deletion effect by

$$\sum_{i=1}^{n} |[y_i - \hat{f}(x_i)]/[1 - h_{ii}(x_i)]|,$$

where h_{ii} is the i-th diagonal element of the hat matrix of the mixed polynomial smoother. The value $h_{ii}(x_i)$ comes from the hat matrix for the weighted least squares operator that fits the polynomial at x_i, which is why in the notation for the diagonal element we show the dependence on x_i. We will now localize this procedure to provide an adaptive fitting method.

For a given x we select a candidate p and carry out a weighted cross-validation of the mixed polynomial, $\hat{p}(x)$, fitted at x. Thus the criterion is

$$\frac{\sum_{i=1}^{n} w_i(x)|[y_i - \hat{p}(x_i)]/[1 - h_{ii}(x)]|}{\sum_{i=1}^{n} w_i(x)}.$$

Now $h_{ii}(x)$ is the i-th diagonal element of the hat matrix of the weighted least squares operator that fits the polynomial at x. We compute the localized criterion for a grid of p values from mixing degree 0 to mixing degree 3, and choose the mixing degree that minimizes the criterion.

We will illustrate the method with an example. The data are sickness rates for ages 29 to 79 reported by Spencer (1904b). We will study the percentage change in the rates by smoothing the first differences of the (natural) log rates as a function of age. Fits are shown in Figure 7. Each column is one value of the nearest-neighbor bandwidth parameter. The values increase logarithmically from 0.25 to 2. The top row is the adaptive fit with the mixing degree chosen locally by the above local method. The remaining rows show the fits for integer degrees 3 down to 0, in other words, local cubic fitting to local constant fitting. Figure 8 shows the residuals of these fits; superposed on each plot is a loess smooth with local linear fitting and $\alpha = 1/3$. Figure 9 shows the cross-validation score as a function of age for each of the five bandwidths, and Figure 10 shows the selected values of the mixing degree p, also as a function of age.

The four displays are quite informative. The residual plots in Figure 8 show the bandwidth at which each smoother just begins to introduce distortion (bias) in the fit. For degree 0, even the smallest bandwidth, $\alpha = 0.25$, has a slight distortion for the largest ages. For degree 1, it begins at $\alpha = 0.42$; for degree 2, it begins at $\alpha = 0.71$, and for degree 3 and the adaptive fit it begins at $\alpha = 1.19$. For no bandwidth does degree 0 produce a fit that has no distortion revealed in the residuals plots even though the fit for the smallest bandwidth is noisy. For degrees 1 to 3 the only distortion-free fits are a bit too noisy, although the fits are smooth and the distortion is quite minor for degree 1 with $\alpha = 0.42$, degree 2 with, $\alpha = 0.71$, and degree 3 with $\alpha = 1.19$. (No saving grace comes to degree 0 since it cannot produce a smooth fit without major distortion.) For the locally adaptive mixed fit with $\alpha = 0.71$

we get an excellent fit with almost no distortion and with requisite smoothness. And, of course, the fit has the attractive property that the parametric family was chosen automatically. The fit also has the interesting property that at about 45 years, just near the first age at which the pattern begins its nonlinear behavior, the local mixing method switches from degree 1 fitting to degree 3 fitting.

9.2 Adaptive Bandwidth Selection

We will now use the assessment of localization to develop an adaptive method for choosing the bandwidth for local linear polynomial fitting. The method is discussed in detail by Cleveland and Loader (1995).

For each x and bandwidth h, consider a localized goodness-of-fit criterion

$$\frac{\sum_{i=1}^{n} w_i(x) E(\hat{p}(x_i) - f(x_i))^2}{\sigma^2 \mathrm{tr}(W)}$$

where $w_i(x) = w(h^{-1}(x_i - x))$, $W = \mathrm{diag}(w_i(x))$, and $\hat{p}(x_i)$ are values of the local polynomial, $\hat{p}$, fitted at x. This is estimated by a localized version of C_p

$$C(h) = \frac{1}{\mathrm{tr}(W)}[2\mathrm{tr}(M_2) - \mathrm{tr}(W) + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n} w_i(x)(y_i - \hat{p}(x_i))^2]$$

where $M_2 = (X^T W X)^{-1}(X^T W^2 X)$, X is the design matrix, and $\hat{\sigma}^2$ is an estimate of σ^2 from a non-adaptive smooth with a small bandwidth. Roughly, for each fitting point, we choose a bandwidth h with small $C(h)$. Specifically, we choose an interval $[h_0, h_1]$, where h_0 is the distance from x to the kth nearest x_i for small k, and h_1 results in rejection of a goodness-of-fit test at a low significance level. Then, the interval $[h_0, h_1]$ is searched for the right-most local minimum of $C(h)$.

This algorithm is carried out at the vertices of a tree in a manner similar to that of loess (Cleveland and Grosse, 1991) but with a split rule suggested in Loader (1994). Suppose the tree has a cell with vertices $[v_k, v_{k+1}]$, and the adaptive procedure produces bandwidths h_k and h_{k+1} at these vertices. The cell requires further refinement if either bandwidth is small relative to the length of the cell; specifically a new vertex is added if $v_{k+1} - v_k > 0.7 \min(h_k, h_{k+1})$. The function estimate is constructed as a cubic interpolant using the local fits and local slopes at the vertices. Implementation involves only minor modification of existing algorithms, and our present code includes multivariate extensions, based on both rectangular and triangular cells. Computation of $C(h)$ is a straightforward by-product of the local fit; we do not rely on higher order fits or Taylor series expansions to estimate bias.

The top panel of Figure 11 shows a made-up data set with 2048 observations on $[0, 1]$; the response function is one of the four examples considered by

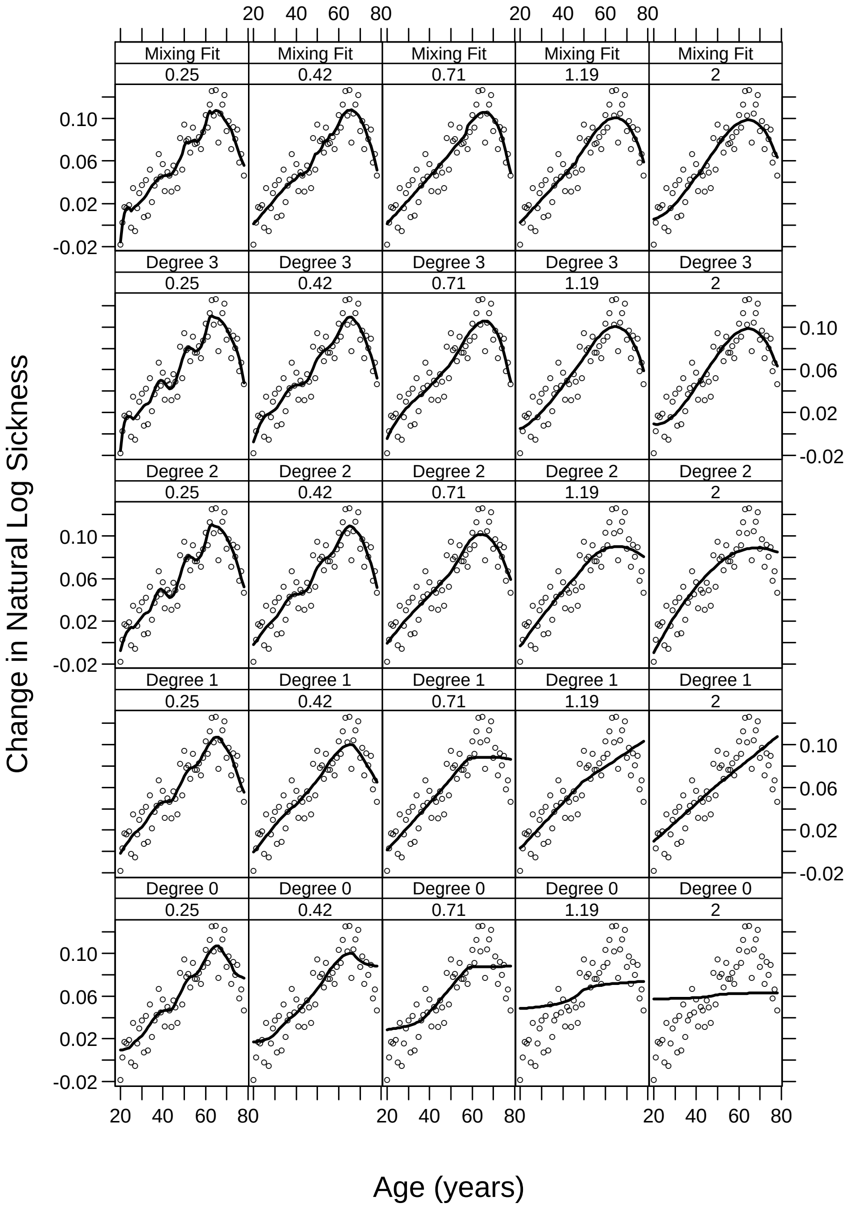


Figure 7: Fits for five nearest-neighbor bandwidths ($\alpha = 0.25$ to 2 in equal percentage steps) and five local fitting methods. The top row shows the fits for the adaptive mixing method, and the remaining rows show local polynomial fits for degrees 3 to 0.

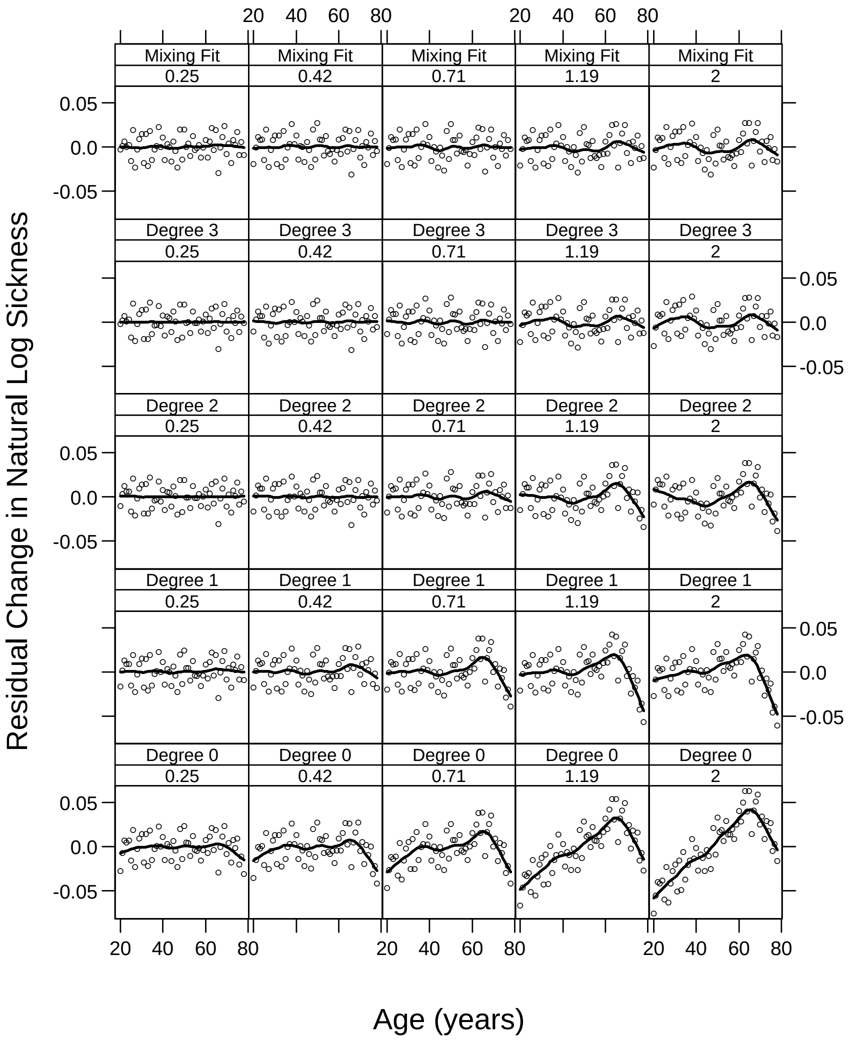


Figure 8: Residual plots for the fits in the previous figure.

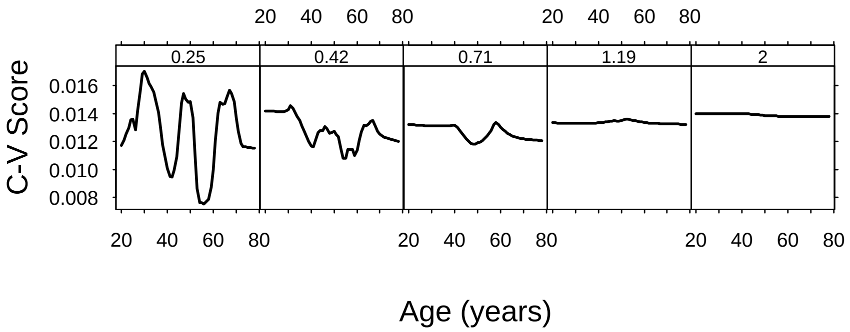


Figure 9: Cross-validation scores for the adaptive mixing fits.

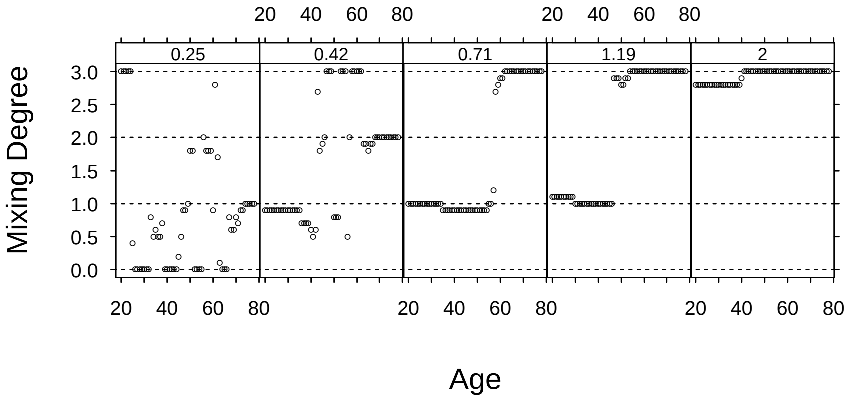


Figure 10: Mixing degrees for the adaptive mixing fits.

Donoho and Johnstone (1994) and Fan and Gijbels (1995). Let the x_i, which are equally spaced, be denoted so that x_i increases with i. We estimate σ^2 by

$$\frac{1}{6(n-2)}\sum_{i=2}^{n-1}(-y_{i-1}+2y_i-y_{i+1})^2.$$

In the second panel of Figure 11, the fit and truth are almost indistinguishable, apart from small amounts of roughness around the discontinuities. The fit appears better than those previously obtained for this problem, with sharper reproduction of the breaks and less noise. However, the residual plot in the third panel shows large residuals around most of the discontinuities, a clear indication that the fit is not perfect. As mentioned earlier, proper modeling of discontinuous functions requires appropriate choice of the weight function and basis functions (Loader, 1993; Speckman, 1995); this is how one would sensibly proceed for such data in practice. The plot in the fourth panel shows bandwidths at the knots determined by the tree algorithm.

10 Theory

10.1 Some Theory with Minimal Assumptions

Local regression with least squares as the criterion results in an estimate that is linear in the y_i provided model selection does not depend on the response. We will take x to be fixed. This is the standard for practice; that is, estimation and sampling distributions are conditional on x. Such conditioning is the sensible practice, as been argued by many (e.g., Cox, 1958).

Exact expressions for the bias and variance of $\hat{f}$ at x are easily obtained. The development here is similar to Katkovnik (1979). For simplicity we will treat the case of one independent variable and local polynomial fitting of degree p.

If f has a continuous second derivative, Taylor's theorem enables us to write

$$\begin{aligned} E\hat{f}(x)=\sum_{i=1}^{n} l_j(x)f(x_j) \quad &= \quad f(x)\sum_{i=1}^{n} l_j(x)+f'(x)\sum_{i=1}^{n}(x_j-x)l_j(x) \\ &\quad +\frac{1}{2}\sum_{i=1}^{n}(x_j-x)^2 l_j(x)f''(\theta_j) \qquad (3) \end{aligned}$$

where $(\theta_j-x)(\theta_j-x_j)\le 0$. For local regression,

$$l(x)^T=(l_1(x),\ldots,l_n(x))=c(x)^T(X^TWX)^{-1}X^TW$$

where $c(x)$ is a column vector of fitting functions $(1 \quad x \quad \ldots \quad x^p)^T$, X is the design matrix, with rows $c(x_i)$, and W is a diagonal matrix of the weights

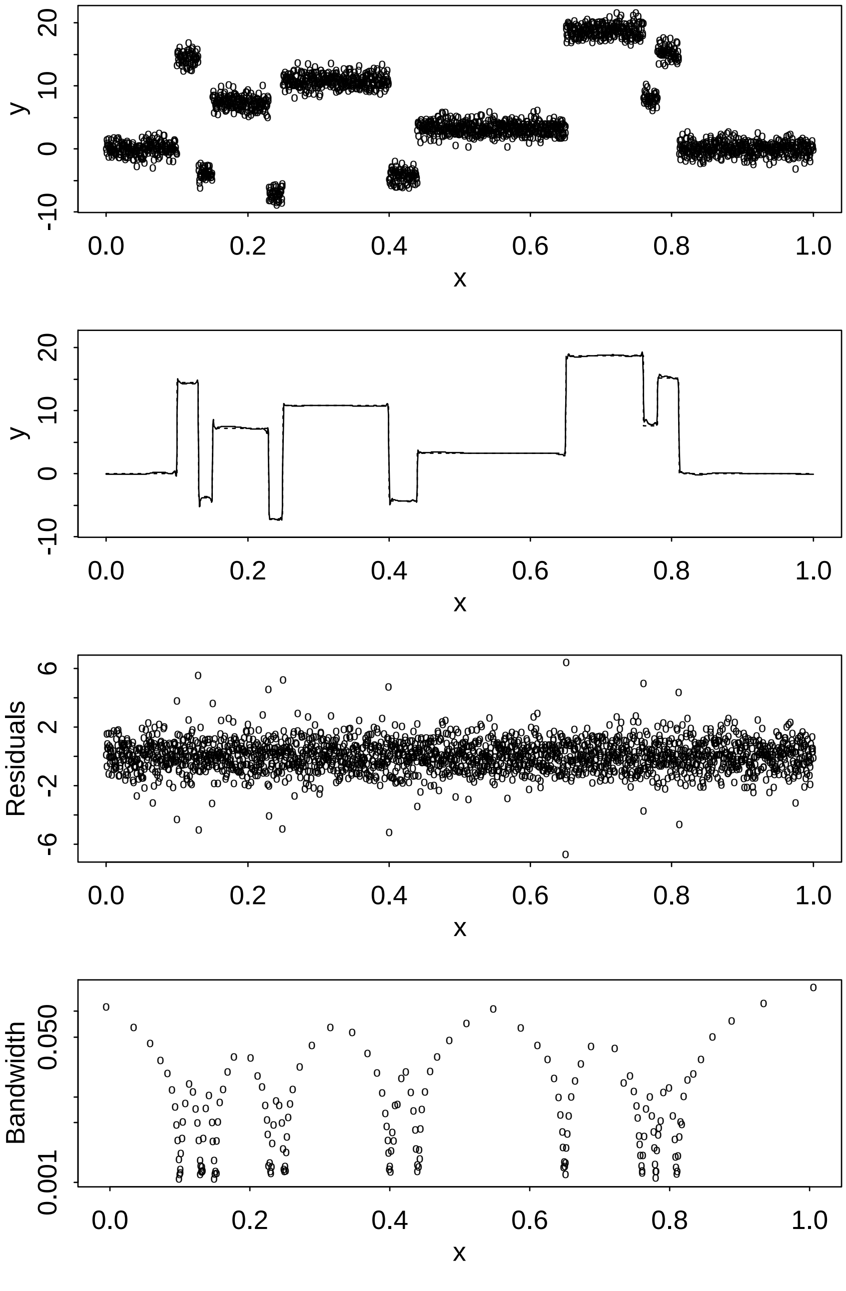


Figure 11: Adaptive bandwidth selection. From the top panel to the bottom the graph shows the data, the adaptive local linear fit, the residuals, and the local bandwidths at the fitted knots.

$w((x_i - x)/h(x))$. Note in particular, $l(x)^T X = c(x)^T$. This is just a mathematical statement of the obvious: If f is exactly one of the fitting functions, then f will be reproduced exactly. If local linear fitting is used, all linear functions are reproduced exactly. In particular, $\sum_{i=1}^n l_j(x) = 1$ and $\sum_{i=1}^n (x_j - x) l_j(x) = 0$. Hence

$$E\hat{f}(x) = f(x) + \frac{1}{2} \sum_{i=1}^{n} (x_j - x)^2 l_j(x) f''(\theta_j). \tag{4}$$

With fits of degree higher than 1, one can retain more terms in the Taylor series, leading to more precise bias estimates. The variance also has a simple closed form expression

$$\mathrm{var}\hat{f}(x) = \sigma^2 \|l(x)\|^2 = \sigma^2 c(x)^T (X^T W X)^{-1} X^T W^2 X (X^T W X)^{-1} c(x). \tag{5}$$

Expressions (4) and (5) are exact; no asymptotics are used. For fits with $p > 0$, terms in the bias expansion involving $f'(x)$ are *gone* for finite samples and *any* values of x_i in the design space, however bizarre. The expressions are easily computed and compared for various estimates in specific examples, and (5) is useful for assessing the variance of the estimate in practice. To make asymptotic analysis of performance and comparison with other estimates tractable, it is necessary to make further assumptions, as we do in the next section. However we emphasize, as has Stone (1980), *such additional assumptions are not required to retain the good properties of local regression.*

10.2 Asymptotic Theory

The results stated below for one independent variable are not new. Tsybakov (1986) and Müller (1987) are among the first to derive these for local regression, although similar expressions for kernel regression and density estimation have been known for much longer. Ruppert and Wand (1994) derive similar results for multivariate predictors.

We must make design assumptions, describing how the design behaves as a function of n. One common assumption is a random design: x_i are independent random variables with density $g(x)$. (This is simply a device, and should not be taken as a signal that in our data analyses we will model our data by random x_i.) Another is fixed design; $i - \frac{1}{2} = n \int_{-\infty}^{x_i} g(u)du$.

When fitting odd order polynomials of degree $p = 2k + 1$

$$\begin{aligned} E\hat{f}(x) - f(x) &= h^{p+1} f^{(p+1)}(x) B_p(w) + o(h^{p+1}) \\ \mathrm{var}\hat{f}(x) &= \frac{\sigma^2}{nhg(x)} V_p(w) + o((nh)^{-1}) \end{aligned}$$

at points where the design density $g(x)$ is continuous and non-zero. Here, $B_p(w)$ and $V_p(w)$ are constants depending only on p and the weight function

w. Letting $\Lambda_j = \int_{-1}^{1} c(x)c(x)^T w^j(x)dx$

$$\begin{aligned} B_p(w) &= \frac{1}{(p+1)!}[\Lambda_1^{-1}\int_{-1}^{1} x^{p+1}c(x)w(x)dx]_1 \\ V_p(w) &= [\Lambda_1^{-1}\Lambda_2\Lambda_1^{-1}]_{1,1} \end{aligned} \tag{6}$$

assuming w has support $[-1,1]$.

For fitting even order polynomials of degree $p = 2k$

$$\begin{aligned} E\hat{f}(x) - f(x) &= h^{p+2}(f^{(p+2)}(x) + (p+2)f^{(p+1)}(x)\frac{g'(x)}{g(x)})B_{p+1}(w) \\ &\quad + o(h^{p+2}) \\ \mathrm{var}\hat{f}(x) &= \frac{\sigma^2}{nhg(x)}V_p(w) + o((nh)^{-1}) \end{aligned}$$

at points where $g(x)$ is differentiable. It can be shown that $V_{2k} = V_{2k+1}$; the only difference between successive even and even orders is in the bias term.

A lower boundary point is typically defined as a point x_0 satisfying $g(x) = 0$ for $x < x_0$; $g(x) > 0$ for $x \geq x_0$ and $g(\,\cdot\,)$ is right continuous at x_0 (Fan and Gijbels, 1992). For example, $g(x) = I_{[0,1]}(x)$ has a lower boundary at 0, and an upper boundary at 1. At boundary regions, asymptotic results must be modified. Suppose x_0 is a boundary point and $x = x_0 + hv$ for fixed v. Then for even order fitting,

$$E\hat{f}(x) - f(x) = h^{p+1}f^{(p+1)}(x)B_{p,v}(w) + o(h^{p+1}).$$

The variance expansions and odd order bias have modified constants $B_{p,v}$, $V_{p,v}$ with forms similar to (6) but with integrals now taken over $[-v,1]$. The biggest difference is in the bias; even order fitting has bias $o(h^{p+1})$ rather than $o(h^{p+2})$. A less obvious difference is that now $V_{2k,v} \neq V_{2k+1,v}$; usually $V_{2k,v} < V_{2k+1,v}$.

How well do these asymptotics perform? Figure 12 studies the simulated data used previously in Figure 5 for fitting local constant through local cubic. The asymptotics suggest the variance (and hence fitted degrees of freedom) should be the same for orders $2k$ and $2k+1$; hence we use $h = 0.8$ for local constant and local linear, and $h = 2.4$ for local quadratic and local cubic. The first problem to note with the asymptotics is the weakness of the setting. Consider the problem of boundary bias. Under the definition of boundaries, the design density has compact support, and is bounded away from 0 on that support. By this definition, the case $x_i \sim N(0,1)$ does not have boundaries! But as we saw earlier, there are substantial boundary effects for this problem. For local constant and linear fitting, the asymptotics are mostly fine. The standard deviation approximation is slightly conservative in most cases. Note however one departure from the asymptotic theory: The local linear method is more variable than local constant near the boundaries. But for local quadratic and cubic fitting, the asymptotics perform very poorly. The bias approximations are completely misleading, with peaks and zeros having little relation to reality.

The limitations of asymptotics have been noted before in various settings. For example, Rosenblatt (1971) in a discussion of optimal kernels states:

> The arguments usually given have been of an asymptotic character and it is a mistake to take them too literally from a finite sample point of view.

Stoker (1993) goes even further and challenges the standard approach to asymptotics, suggesting instead that a bandwidth fixed as $n \to \infty$ gives a more realistic assessment of the performance of estimates.

The clear message then is one of caution: realistic problems are not well modeled by the asymptotic framework, and the asymptotic approximations can be quite poor. For example, it would be a mistake to consider two procedures as equivalent purely on the basis of having the same asymptotic expansions, or to use asymptotics alone to justify smoothing procedures.

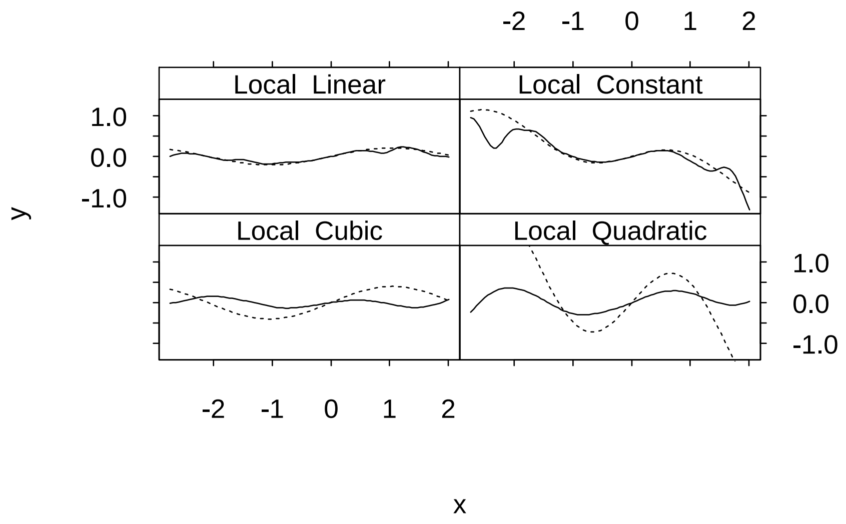
Unfortunately the use of asymptotics in the kernel smoothing literature over the past 10-15 years has not been good. Sweeping conclusions are drawn solely from asymptotics; in some cases, procedures are labeled as 'optimal' without any real justification. We present two examples in the next two sections — bandwidth choice and polynomial degree.

10.3 Bandwidth

Over the last decade, much effort has been expended devising methods to 'automatically' select a bandwidth and assessing how close the selected bandwidth is to a theoretically optimal bandwidth, with little regard for the possibility that the optimal bandwidth may be very inadequate. Recently, a major focus has been 'plug in' methods (Hall et al., 1991), in which one directly estimates the bias and variance (or asymptotic approximations to these) and choosing a bandwidth attempting to minimize either pointwise MSE or global criterion such as MISE.

The key problem here is bias estimation. For a heavily biased estimate, bias estimation is a reasonable problem. For the local linear estimate considered in Figure 5, bias estimation essentially amounts to estimating the curvature; this can be accomplished using a local quadratic or cubic fit (higher order kernels have usually been employed in the literature). A plug-in algorithm could work quite well here, if our criterion were how close the selected bandwidth is to an optimal bandwidth.

However, our criterion should not be the quality of the bandwidth selector, but the quality of the resulting estimate. Hence, the available information on curvature should go *directly into the estimate* $\hat{f}$ and not just into a bandwidth selector. The resulting smoother should not be heavily biased, and therefore its bias should be difficult to estimate. The problem with plug-in methods



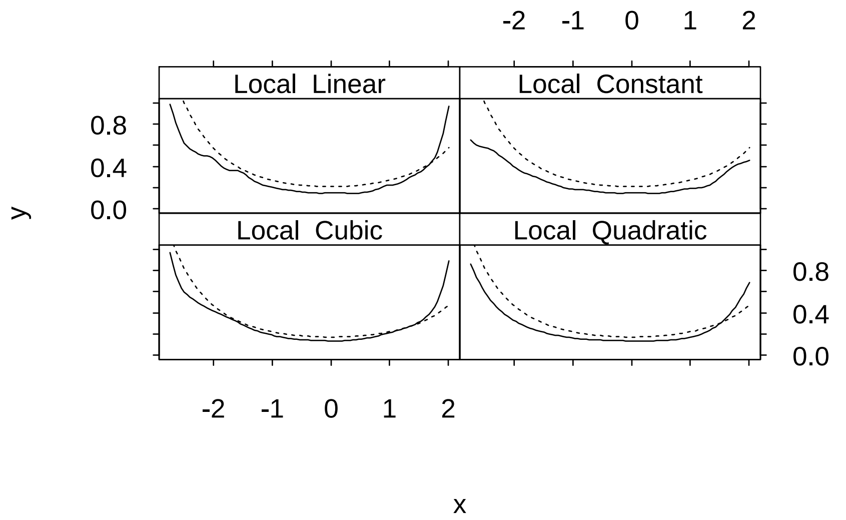


Figure 12: Biases (top panel), standard deviations (bottom panel) and their asymptotic approximations.

now becomes clear: For local quadratic and cubic fits, bias estimation essentially amounts to estimating fourth order derivatives, about which the data contains little or no information. In any case, the asymptotic expressions are very poor. To quote Katkovnik (1979)

> Attempts to select h from theoretical precision estimates are not very productive since they require assigning a type of information about the function $f(x)$ and the noise which usually does not exist apriori.

The weakness of plug-in methods can be seen (and occasionally has been acknowledged) from a comparison of the assumptions with work on optimal rates of convergence. Bandwidth selectors for second order kernel estimates are usually derived under an assumption of four or more derivatives; under this assumption the best possible rate on convergence of $\hat{f}(x)$ to $f(x)$ is $O_p(n^{-4/9})$ (Stone, 1980). But a second order kernel method achieves a rate no better than $O_p(n^{-2/5})$, regardless of how good the bandwidth selector is. While the difference between $2/5$ and $4/9$ is small, the advantage of going to higher order methods can be convincingly demonstrated through examples (e.g., Figure 5).

To anyone with a plug-in bandwidth selector for local constant or local linear regression, we suggest the following comparison: Fit a local quadratic regression, with the same variance as measured by equivalent degrees of freedom $\mathrm{tr}(L^T L)$ (or for pointwise bandwidth selectors, $\|l(x)\|^2$). Asymptotically, the local quadratic method will improve on the local linear, if anything more than existence of a second derivative is assumed. And improvement will also be observed in practice even in simple examples, such as that studied in Figure 4.

10.4 Degree of Fit

A question that has provoked much discussion is that of the relative merits of successive even and odd orders — local constant versus local linear fitting, and local quadratic versus local cubic fitting. The comparison is quite different at interior points and in 'boundary' regions, as can be seen either through examples or the asymptotics.

Consider the case of local constant versus local linear fitting. As noted earlier, for local constant fitting, bias is of size $O(h)$ at the boundaries, and $O(h^2)$ at interior points. For local linear fitting, the bias is $O(h^2)$ everywhere. A similar comparison holds at higher orders; fitting of order $2p+1$ has $O(h^{2p+2})$ bias everywhere, and fitting of order $2p$ has bias $O(h^{2p+1})$ at boundaries and $O(h^{2p+2})$ at interior points.

However, this observation alone *is not an argument in favor of odd order fitting*, and *does not imply odd orders have no boundary effects.* Variance has so far been ignored. Local polynomial fitting is much more variable at boundary regions, and this variance increase is more pronounced for odd orders, as shown for one example in Figure 12. What we can conclude is that boundary regions dominate the comparison. For local constant fitting, the boundary bias caused by slope is widely recognized as being unacceptable, and local linear is usually preferable. Our examples support this conclusion. The comparison of local quadratic and local cubic is much less clear cut.

11 Summary of Conclusions

We have had many detailed discussions in this paper with major conclusions embedded in them. Here we simply present the conclusions.

In carrying out local regression we use a parametric family just as in global parametric fitting, but we ask only that the family fit locally and not globally. This is parametric localization.

Four important aspects of local regression are the weight function, bandwidth selection, the local parametric family, and the fitting criterion. The choices of these aspects represent a modeling of the data that can be guided by automatic selection criteria and graphical diagnostics.

Polynomial mixing provides a continuous progression of polynomial fits beginning with local constant fitting and proceeding continuously though higher orders.

The early smoothing literature from the first four decades of the twentieth century provides a number of important insights about the choices (e.g., Spencer, 1904a; Henderson, 1916; Macaulay, 1931). For example it was nearly a given in this literature that for most applications the weight function needed to be smooth, that local constant fitting was inadequate, and that smoothers needed to reproduce exactly (and not just asymptotically) at least a quadratic. Despite this important widely professed intuition, much theoretical smoothing research during the 1980s and early 1990s focused on locally constant fitting and ignored higher order polynomial fitting until the papers of Fan (1993) and Hastie and Loader (1993) appeared. Fortunately, the success of local polynomial methods in practice (Stone, 1977; Cleveland, 1979; Friedman and Stuetzle, 1982; Cleveland and Devlin, 1988; Hastie and Tibshirani, 1990) established them and not local constant fitting as the standard approach for data analysis.

Of the two bandwidth specifications, fixed and nearest-neighbor, the latter has the attractive property that there are typically less radical swings in the variance of the smooth. Typically, neither method works well if there are radical changes in the smoothness of the function; in such a case, adaptive methods can be helpful.

We introduce the assessment of parametric localization as a general approach to adaptive fitting. We discuss methods of adaptive selection of mixing degree and for adaptive bandwidth selection. This leads to far better adaptive procedures than the pointwise bias estimation methods that have originated in the kernel literature.

Asymptotics for smoothing give only the roughest of indicators of what works when the number of observations is finite. One principal problem is that for finite samples, boundary regions are important because bandwidths can be large. Unfortunately, asymptotic results have sometimes been interpreted too much at face value. Two conclusions that have been drawn in the literature — (1) the fit of an odd order polynomial is better than the next lowest even order (2) fixed bandwidth smoothing is better than nearest-neighbor — do not hold up for finite samples.

REFERENCES

Brillinger, D. (1977). Discussion of a paper of Stone. *Ann. Statist.* **5**, 622-623.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assn.* **74**, 829-836.

Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, NJ, books@hobart.com.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assn.* **83**, 596-610.

Cleveland, W. S., Hastie, T., and Loader, C. (1995). Adaptive local regression by automatic selection of the polynomial mixing degree. In preparation.

Cleveland, W. S. and Grosse, E. H. (1991). Computational methods for local regression. *Statist. and Computing* **1**, 47-62.

Cleveland, W. S. and Grosse, E. H. and Shyu, M. J. (1992). Local regression models. *Statistical Models in S*, J. M. Chambers and T. Hastie, editors, pages 309–376. Chapman and Hall, New York.

Cleveland, W. S. and Loader, C. (1995). Computational methods for local regression from the 19th century to the present. In preparation.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372.

Daniel, C. and Wood, F. (1971). *Fitting Equations to Data.* Wiley, New York.

De Forest, E. L. (1873). On some methods of interpolation applicable to the graduation of irregular series. *Annual Report of the Board of Regents of the Smithsonian Institution for 1871*, 275-339.

De Forest, E. L. (1874). Additions to a memoir on methods of interpolation applicable to the graduation of irregular series. *Annual Report of the Board of Regents of the Smithsonian Institution for 1873*, 319-353.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008-2036.

FAN, J. and GIJBELS, I. (1994a). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assn.* **89**, 560-570.

FAN, J. and GIJBELS, I. (1994b). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. Unpublished manuscript, available by `ftp` from `stat.unc.edu`.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc,* Ser. B. **57**, 371-394.

FRIEDMAN, J. H. Multivariate adaptive regression splines (1991). *Ann. Statist.* **19**, 1-141.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assn.* **76**, 817-823.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report Orion 3, Dept. Statistics, Stanford University.

GASSER, TH. and JENNEN-STEINMETZ, C. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assn.* **83**, 1084-1089.

GRAM, J. P. (1883). Über Entwickelung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate. *J. Math.* **94**, 41-73.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263-269.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Oxford University Press, Oxford.

HASTIE, T. and LOADER, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statist. Science* **8**, 120-143.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HENDERSON, R. (1916). Note on graduation by adjusted average. *Actuarial Soc. Amer.* **17**, 43-48.

HENDERSON, R. (1924). A new method of graduation. *Actuarial Soc. Amer.* **25**, 29-39.

HOEM, J. M. (1983). The reticent trio: some little-known early discoveries in life insurance mathematics by L. H. F. Oppermann, T. N. Thiele and J. P. Gram. *Inter. Stat. Rev.* **51**, 213-221.

HJORT, N. L. (1994). Local Bayesian regression. Unpublished manuscript.

HJORT, N. L. and JONES, M. C. (1994). Locally parametric nonparametric density estimation. Unpublished manuscript.

KATKOVNIK, V. YA. (1979). Linear and nonlinear methods of nonparametric regression analysis. *Soviet Automatic Control* **5**, 25-34.

KENDALL, M. G. (1973). *Time Series.* Oxford University Press, Oxford.

KENDALL, M. G. and STUART A. (1976). *The Advanced Theory of Statistics, Vol. 3.* Hafner, New York.

LANCASTER, P. and SALKAUSKAS, K. (1981). Surfaces generated by moving least squares methods. *Mathematics of Computation* **37**, 141-158.

LANCASTER, P. and SALKAUSKAS, K. (1986). *Curve and Surface Fitting: An Introduction.* Academic Press: London.

LOADER, C. (1993). Change point estimation using local regression. Unpublished manuscript, available by `ftp` from `netlib.att.com`.

LOADER, C. (1994). Computing nonparametric function estimates. In preparation.

LOADER, C. (1995). Local likelihood density estimation. *Ann. Statist.* , to appear.

MACAULAY, F. R. (1931). *The Smoothing of Time Series.* National Bureau of Economic Research, New York.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

MALLOWS, C. (1974). Discussion of a paper of Beaton and Tukey. *Technometrics* **16**, 187-188.

MCDONALD, J. A. and OWEN, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28**, 195-208.

MÜLLER, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12**, 766-774.

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assn.* **82**, 231-238.

NADARAYA, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* **9**, 141-142.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065-1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42**, 1815-1842.

RUPPERT, D. and WAND, M. P. (1992). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, No. 3.

SHISHKIN, J, YOUNG, A. H., and MUSGRAVE, J. C. (1967). The X-11 variant of the Census Method II seasonal adjustment program. Technical Paper 15, U.S. Bureau of the Census.

SPECKMAN, P. (1995). Discussion of a paper of Donoho et al. *J. Royal Statist. Soc,* Ser. B. **57**, 337-338.

SPENCER, J. (1904a). On the graduation of the rates of sickness and mortality. *J. Inst. Act.* **38**, 334-347.

SPENCER, J. (1904b). Graduation of a sickness table by Makeham's hypothesis. *Biometrika* **3**, 52-57.

STANISWALIS, J. (1988). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assn.* **84**, 276-283.

STIGLER, S. M. (1978). Mathematical statistics in the early States. *Ann. Statist.* **6**, 239-265.

STOKER, T. M. (1993). Smoothing bias in density derivative estimation. *J. Amer. Statist. Assn.* **88**, 855-863.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595-620.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

STONE, M. (1974). Cross-validatory choice of assessment of statistical predictions (with discussion). *J. R. Statist. Soc. B* **36**, 111-47.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assn.* **82**, 559-567.

TSYBAKOV, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Prob. Inform. Trans.* **22**, 69-84.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhya* Ser. A **26**, 359-372.

WAHBA, G. (1990). *Spline Functions for Observational Data.* SIAM, Philadelphia.

WHITTAKER, E. T. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* **41**, 63-75.

WOOLHOUSE, W. S. B. (1870). Explanation of a new method of adjusting mortality tables, with some observations upon Mr. Makeham's modification of Gompertz's theory. *J. Inst. Act.* **15**, 389-410.