Local Regression Models

Bill Cleveland and Clive Loader
Bell Laboratories
Murray Hill, NJ 07974

September 8, 1998

Abstract

This is the first draft of the local regression chapter. It's rough. The writing needs editing, there are a few missing sections, more references are needed, and examples need to be inserted. Except for the examples, most problems should be cleared up by Wed. September 2 when we will make a revision available. We may have an example or two but are hoping that at the meeting Sept. 12/13 participants can point us to some exciting datasets. We have many, many applications available but most do not fall into the category of economic and business data. Those that we do have are boring.

1 Introduction

1.1 What Is Local Regression?

Regression models are used to study the dependence of a response variable Y on a vector of d predictor variables x. Suppose Y_i are observations of the response and x_i are observations of the predictors for $i = 1$ to n. Normal regression models take the form

$$Y_i = \mu(x_i) + \epsilon_i$$

where the ϵ_i are independent normal random variables with mean zero and variance σ^2, and $\mu(x)$ is the mean of Y given at x.

Regression models are often parametric. This means that μ is specified, as part of the model, to be a member of a parametric family, $\mu(x;\theta)$. In the normal linear model, $\mu(x;\theta)$ is a linear function of the predictors and the parameters are the coefficients of the function and σ^2. To fit this model

to the data, the parameters are estimated by least squares since the error terms are normal. The fitted regression function is

$$\mu(x, \hat{\theta})$$

where $\hat{\theta}$ are the least squares estimates of θ.

What if the data suggest that $\mu(x)$ is a complex function, say nonlinear with interactions among the predictors? One option is to specify a complex parametric family with a large number of parameters and hope the resulting class is sufficiently rich to approximate the true relation. We would still be in the domain of parametric regression.

An alternative is a local regression model. We still assume the regression model (1.1). But the mean function $\mu(x)$ is not constrained to have a simple parametric form. Instead, we assume that *locally*, in a specified neighborhood around each point x, μ can be well approximated by a member of a parametric class such as low order polynomials. We specify the parametric class and the neighborhoods, and then form an estimate, $\hat{\mu}(x)$, of the regression function using local regression fitting methods, which we will describe shortly.

Regression models are also commonly employed for data where the Y_i have a nonnormal distribution such as the binomial or Poisson models. Local regression models can be employed here too. For example, a standard parametric model for independent binary observations is

$$\frac{E(Y_i)}{1 - E(Y_i)} = \mu(x_i)$$

where μ is a member of a parametric family. In local regression we simply assume that μ can be well approximated locally by a member of a parametric class.

1.2 How Local Regression Models are Fitted to Data

The estimation of μ that arises from local regression models is terribly simple. This is part of its attraction. Let x be any point in the space of the predictor variables. Define a metric in this space; one possibility is to use Euclidean distance after dividing each independent variable by a sample scale estimate such as the standard deviation or the median absolute deviation. Next, determine a neighborhood of x by some method using this metric. Then, fit a member of the parametric class by estimating the parameters using only the observations in the neighborhood; the local fit at x is the

fitted parametric function evaluated at x. Almost always, we will want to incorporate a *neighborhood 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. We repeat this weighted fitting process for all those values x where we want an evaluation of the surface.

The method of estimation depends on the form of the distribution that is hypothesized for Y_i. If the distribution is normal then we use weighted least squares with the weights from $w(u)$. If the data are binary and we use the above logistic model, then the estimation is carried out by maximizing a weighted likelihood function.

1.3 Neighborhood Selection

To use local regression in practice to estimate μ we must define neighborhoods for any x where the estimate is to be computed. One simple choice is nearest neighbors. We find $x_{(r)}$, the rth closest x_i to x where $r \leq d$. The neighborhood is a d-dimensional sphere about x whose radius is the distance from x to $x_{(r)}$. Let $\alpha = r/d$, the fraction of points in the neighborhood. Then the local regression model specifies μ through the specification of α and the specification of the parametric family to be fitted locally.

1.4 Specification of Local Regression Models

The specification of a regression model consists of specifying the distribution of Y_i and then further specifying the function $\mu(x)$. Parametric and local regression models differ in how $\mu(x)$ is specified. For a parametric regression model, $\mu(x)$ is specified by a parametric family. For a local regression model, $\mu(x)$ is specified by the neighborhoods and the parametric family to be fitted locally. If we use nearest neighbors, then the neighborhood specification is a choice of α.

1.5 Why are Local Regression Models Useful: The Complementary Roles of Parametric and Local Models

There are two outcomes that can result from fitting a local regression model. One is that we adopt the local regression specification of $\mu(x)$ to provide the estimate of the function to fulfill the purpose of the analysis. Another is that the local regression estimate suggests modeling $\mu(x)$ by a parametric family, we switch to the parametric regression model, the new model passes the diagnostic checks, and we use the parametric estimate to to fulfill the purpose of the analysis. So the outcome can be that the local regression

model provides the final answer or the that the model is an exploratory tool that suggests a parametric model. In such a case, when the parametric model is parsimonious — that is, with a small number parameters — we choose the parametric regression model for simplicity of description of the behavior of the mean of $\mu(x)$.

But regression surfaces can be complex: nonlinear and with intricate interactions among variables. It is possible in some cases to develop parsimonious parametric models that fit the data. But nature continually confronts us with dependencies too complicated to describe by parsimonious parametric functions. In many of these cases, local regression can often provide an adequate fit because local regression can accommodate a very wide range of function shapes.

There are cases, too, where with enough time and energy we can build a parametric model, but the effort is far greater than building a local regression model, and the resulting subject matter consequences are no different. For complex regression functions, parametric model building is difficult because accommodating nonlinear functions with interactions can result in the consideration of many parametric families; there can be a large number of derived predictors from which to chose for inclusion in the model, such as powers and cross products of the original predictors, even when the number of original predictors is quite modest. By contrast, in local regression we search through a much more limited class of parametric families — usually, just polynomials of degree one or two. And typically, the neighborhood choice is reduced to the choice of a single parameter such as the α for nearest neighbors.

1.6 Similarities of Parametric and Local Regression Models

Parametric and local regressions have many similarities.

Interpretability

A quite simple parametric regression is not only parsimonious but typically has a high degree of interpretability if the parameters have incisive interpretations in terms of the subject matter. This is why we are ready to switch from a local to a simple parametric regression when the former suggests the latter.

But when parametric regression surfaces are complex — nonlinear and with interactions — it is almost never possible to interpret them through their parameters. The parameters of a parametric fit in such a case are

nothing more than a vehicle to compute the surface. Typically, we must use visual displays of $\mu(x)$ to comprehend their characteristics. We comprehend local regression surfaces, visualizing them in the same way. The important point is that for complex surfaces there is usually no enhanced interpretability to parametric surfaces.

Normal Distributions: Degrees of Freedom

For normal Y_i, local regression models have a notion of degrees of freedom, *df* just as parametric models do. There are a number of ways to implement the notion but here is one. The fitted values of a regression are $\mu(x_i\hat{\theta})$ for $i = 1$ to n. The least squares estimation of the fit results in fitted values that are linear combinations of the Y_i. Let $\mathbf{L}$ be the matrix that maps the measurements of the response, Y_i, to the fitted values. $\mathbf{L}$ depends only on the x_i. For local regression, $\text{tr}(\mathbf{L})$ can be taken to be the degrees of freedom, We can interpret this *df* as we do in the parametric case. *df* is a measure of how much variability out of n degrees of freedom are used up in the fitting. Suppose a local regression model of $\mu(x)$ uses nearest neighbor bandwidths with parameter α and polynomials of degree p. Then *df* tends to increase as p increases or as α decreases. In fact, a reasonable approximation of *df* for well-behaved x_i is is $1.2/\alpha$ for $p = 1$ and $3.6/\alpha$ for $p = 2$.

Normal Distributions: Sampling Distributions

For normal Y_i, a number of standard statistics are the basis of inferences for parametric models — χ^2, F, and t statistics. For normal local regression models, if we define these statistics in the same way, their sampling distributions are quite well approximated by the parametric sampling distributions with *df* defined in a manner similar to that above.

Exploration, Model Fitting, and Diagnostic Checking

In building regression models, we are rarely given the model specifications from information outside of the data, and we must look to the data to provide guidance. If we consider the entire modeling process — (1) exploratory analysis to determine an initial model for the data; (2) model fitting; (3) diagnostic checking to assess the modeling assumptions — then building a local regression model can employ many of the same tools as the building of parametric models, so we can bring to bear a wide range of powerful tools that have been developed for parametric fitting (Cleveland 1993).

1.7 Origins of Local Regression

Smoothing by local regression dates back at least to the early 19th century. For over 100 years, the major methodological developments came from researchers analyzing business and economic data. Application was restricted to one predictor, usually with equally spaced values, since the computational burden for other cases was typically prohibitive. Initial work began in actuarial studies where mortality and sickness rates were fitted as a function of equally spaced ages. In the early 20th century quite formidable applied mathematicians such as Henderson and Whittaker vastly enriched the theory and methods of smoothing. Henderson (1916) published results that gave considerable insight into the properties of local regression, showing that a number of smoothers that were widely used were local regression procedures with the parametric family equal to cubic polynomials. The local regression methodology moved into applications to business and economic time series. In 1931 the National Bureau of Economic Research published what would become a classic account of smoothing (Macaulay 1931).

In the statistics literature of the 1960s and early 1970s, local regression moved into more general regression studies, that is studies with more than one predictor and unequally spaced values. The first entry was the kernel smoother, local regression with the parametric family taken to be polynomials of degree zero (e.g. Watson 1964 and Nadaraya 1964) . From these initial efforts there arose an enormous theoretical literature focusing an asymptotic results in which n goes to infinity and neighborhood bandwidths go to zero. Initially, results were not refined enough to show what theory and practice had demonstrated in the early work of the 1910s, namely, that a polynomial family of degree zero represents a poor trade-off of theory and bias. (But finally, Fan showed the superior properties of local regression with higher degree polynomials).

In the 1970s, local regression with polynomials of degree higher than zero were introduced into the statistics literature (Stone 1977; Cleveland 1979; Katkovnik 1979). Implicit in these writings, and later explicitly in (Cleveland and Devlin 1988; Cleveland, Grosse, and Shyu 1992), was the notion that local regression analysis provides a model for the data in which we attempt to approximate as well as possible the underlying regression surface.

1.8 Computational Methods

Local regression has always taxed the computational capabilities of the time, whether in 1880 or today. Throughout, it has been necessary to balance computational efficiency with the statistical performance of smoothers. And throughout, it has been possible to develop computational methods that provide both good performance and computation fast enough for use in practice. This has resulted in widespread application of local regression tools.

From 1880, until computers appeared, summation formulas were employed that economized on manual computation; they consisted of a series of very arithmetically simple updating filters. Spencer (1904) developed summation formulas that became widely used and persisted in practice for decades.

Today, even with very fast computers we need computational methods. The prevalent method today is sparse localization and interpolation. The idea is quite simple — carry out local fits at a judiciously chosen set of compute points in the space of the predictors, and then use interpolation to get the regression surface elsewhere. This method was first employed for smoothing as a function of one predictor (Cleveland 1979) and was later extended to several independent variables (Cleveland and Devlin 1988). Recently, a number of new methods for sparse localization and interpolation have been developed and extended to a wide domain of local regression models (Loader 1998).

1.9 Software

Software for local regression consists of two distinct parts. The first is base software written in either C or Fortran or both, that carries out basic computations. The most widely-used base software for local regression is in the public domain. It uses the sparse localization methods described above. Users with an ability to program with low-level languages like C can employ this software. The second is interface software that calls the base software from other computing environments and substantially lightens the programming burden in carrying out local regression. Most of this interface software is in commercial packages.

LOWESS

This durable set of routines, written in Fortran in the late 1970s, are widely used. They smooth just as a function of one predictor for Y_i with normal

distributions or Y_i with long-tailed symmetric distributions (robust fitting). The statistics described above are not computed. The base software is available at

`http://netlib.bell-labs.com/netlib/go/lowess`

Other software systems have interfaces to LOWESS, including S-Plus, R, Systat, XploRE and Gauss. A SAS macro is available from

`http://hotspur.psych.yorku.ca/SCS/sssg/lowess.html`

LOESS

The loess base software is a collection of C and Fortran subroutines that carry out local fitting for normal and long-tailed distributions and for one or more predictors. It also computes the statistics described above. The base software is available at

`http://netlib.bell-labs.com/netlib/a/dloess`

and a postscript user's manual at

`http://netlib.bell-labs.com/netlib/a/cloess.ps`

The commercial system S-PLUS has a complete implementation of loess as part of its modeling language.

LOCFIT

The LOCFIT software, written in C, is by far the most ambitious. It computes fits and statistics for a wide range of distributions, including normal, long-tailed, binomial, exponential, and Poisson. It also computes density estimates. It can be used as stand-alone program with its own interface, or as a library in the S-Plus or R systems. The code is available at

`http://cm.bell-labs.com/stat/project/locfit/`

LOCFIT is described in more detail in a forthcoming book Loader (1998).

2 Local Regression Fitting

The components for a local fit are

1. a set of localizing weights;
2. a local model (often a polynomial);
3. a fitting criterion.

2.1 Localizing weights

For each fitting point x we need a vector of localizing weights $\{w_i(x)\}_{i=1}^n$. These weights should be chosen so that observations x_i 'close' to x receive the most weight, and observations far from x receive little or no weight. The most common specification of weights is in terms of three components:

1. A distance function $d(x_i, x)$. In one dimension, this is simply $|x_i - x|$. In multiple dimensions, the Euclidean distance $||x_i - x||$, or a weighted Euclidean distance with separate scales for each direction, are the most common choices.

2. A weight function $W(v)$, usually nonnegative, peaked at 0 and supported on $|v| < 1$. In this paper, we use $W(v) = (1 - |v|^3)_+^3$.

3. A bandwidth h, or bandwidth function $h(x)$. The bandwidth controls the size of the local neighborhood; we discuss this choice more later.

Given these three components, the localizing weights are

$$w_i(x) = W(\frac{d(x_i, x)}{h}).$$

2.2 The Local Model

The local model is often chosen to be a linear or quadratic polynomial. For example, in one dimension, a quadratic polynomial is

$$P(x_i - x) = a_0 + a_1(x_i - x) + \frac{a_2}{2}(x_i - x)^2.$$

More generally, a local polynomial can be written as the inner product

$$P(x_i - x) = \langle a, A(x_i - x) \rangle$$

where a is a vector of coeffiencients, and $A(x_i - x)$ is a vector of the fitting functions. In the quadratic example,

$$a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \qquad\qquad A(v) = \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix}.$$

Local quadratic fits are generally better in problems exhibiting high curvature, but can sometimes produce fits that are too noisy. Generally, there is little benefit (and increased computational complexity) to fitting local polynomials of degree higher than 2.

Local constant fitting (also known as a local average or kernel estimate) has been widely used in the statistics literature. But this is often inadequate. The most severe problem is boundary bias: if the true mean exhibits slope near ends of the observation interval, the local average cannot adequately track the data near boundaries, unless a smaller bandwidth is used. While ad-hoc corrections such as boundary kernels can provide correction for boundary bias in some circumstances, local linear (or higher order) methods provide a more direct and general solution to this problem (Hastie and Loader 1993).

2.3 Fitting Criterion

Many possible choices are available for the fitting criterion. The most common choice is local least squares:

$$\min_a \sum_{i=1}^{n} w_i(x)(Y_i - \langle a, A(x_i - x)\rangle)^2. \tag{1}$$

Let the minimizing vector be $\hat{a} = (\hat{a}_0, \hat{a}_1, \hat{a}_2)$. The local regresion estimate is simply the first component:

$$\hat{\mu}(x) = \hat{a}_0.$$

2.4 Bias and Variance

Since the local regression estimate $\hat{\mu}(x)$ is defined as the solution of a weighted least squares problem, much of the theory and methodology developed for use with linear regression models extends simply to the local regression case. In particular, one can easily derive an explicit form for the parameter estimates:

$$\hat{a} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}Y$$

where $\mathbf{X}$ is the design matrix; $\mathbf{W}$ is a diagonal matrix with elements $w_i(x)$, and Y is a vector of the responses $Y_1, \ldots, Y_n$. It also follows that $\hat{\mu}(x)$ is a linear estimate: that is, there exists a **weight diagram** vector $l(x) = \{l_i(x)\}_{i=1}^n$ such that

$$\hat{\mu}(x) = \sum_{i=1}^{n} l_i(x)Y_i = \langle l(x), Y\rangle = \langle l(x), Y\rangle\,.$$

There also exists a **hat matrix L** such that

$$\begin{pmatrix} \hat{\mu}(x_1) \\ \vdots \\ \hat{\mu}(x_n) \end{pmatrix} = \mathbf{L}Y.$$

Statistical properties of the local regression estimate, such as the mean and variance, can be derived in terms of the weight diagram and hat matrix. The mean of the local regression estimate is

$$E(\hat{\mu}(x)) = \sum_{i=1}^{n} l_i(x)\mu(x_i) = \langle l(x), E(Y) \rangle \tag{2}$$

and assuming the errors ϵ_i are independent with common variance σ^2, the variance of $\hat{\mu}(x_i)$ is

$$\text{var}(\hat{\mu}(x_i)) = \sigma^2 \sum_{i=1}^{n} l_i(x)^2 = \sigma^2 \|l(x)\|^2. \tag{3}$$

Note that $\hat{\mu}(x)$ is a biased estimate of the true mean $\mu(x)$. As the size of the local neighborhood (or the bandwidth h) increases, the local model becomes a poorer fit to the data, and the bias, or distortion of the data, becomes larger. On the other hand, as h decreases, there is less data in the local neighborhood, and the parameters of the local model become difficult to estimate, and the variance of $\hat{\mu}(x)$ increases. Thus, there is a trade-off between bias and variance: the bandwidth h must be small enough to control the bias of $\hat{\mu}(x)$, but large enough to control the variance of $\hat{\mu}(x)$.

The *degrees of freedom* of the local estimate are also defined in terms of the hat matrix. In fact, there is no unique definition, but two of the most useful are

$$\begin{aligned} \nu_1 &= \sum_{i=1}^{n} \text{infl}(x_i) = \text{tr}(\mathbf{L}) \\ \nu_2 &= \sum_{i=1}^{n} \|l(x_i)\|^2 = \text{tr}(\mathbf{L}^T\mathbf{L}). \end{aligned} \tag{4}$$

For a parametric regression model, the hat matrix **L** is symmetric and idempotent. In this case, the two definitions coincide, and usually equal the number of parameters, k, in the model. For local regression models, the two definitions are usually not equal; under fairly general conditions, $k \leq \nu_2 \leq \nu_1 \leq n$.

2.5 Henderson's Theorem

A result of Henderson (1916) brought together a diverse range of graduation rules. Henderson showed that essentially any graduation rule that reproduced cubic polynomials could be represented as a local cubic regression. This result, although deceptively simple, has profound consequences, even for present day statistical research. In particular, many alternative constructions of 'reduced bias' regression estimates have appeared in the statistics literature; related problems of bias estimation appear in confidence interval construction and bandwidth selection. Before discussing the consequences, we state the result in the generality of the modern local regression framework.

Theorem 1 *The weight diagram for a local polynomial fit of degree p has the form*

$$l_i(x) = W\left(\frac{x_i - x}{h(x)}\right)\langle \alpha, A(x_i - x)\rangle\,; \tag{5}$$

that is, the least squares weights multiplied by a polynomial of degree p. This representation is unique, provided $\mathbf{X}^T\mathbf{W}\mathbf{X}$ *is non-singular.*

Conversely, if a linear estimate reproduces polynomials of degree p, and the weight diagram has at most p sign changes, then the estimate can be represented as a local polynomial fit of degree p.

Later, we show the consequences of Henderson's theorem in confidence interval construction and bandwidth choice. For now, we just give a simple example involving bias correction by a 'double smoothing' procedure. By (2), the bias of the local regression estimate is

$$b(x) = \sum_{i=1}^{n} l_i(x)\mu(x_i) - \mu(x).$$

The bias is unknown, since it depends on $\mu(x)$. But we might substitute the estimate $\hat{\mu}(x)$ to obtain an estimate $\hat{b}(x)$, and thus a bias corrected estimate

$$\hat{\mu}(x) - \hat{b}(x) = 2\hat{\mu}(x) - \sum_{i=1}^{n} l_i(x)\hat{\mu}(x_i).$$

Now suppose $x_i = i; i = \ldots, -1, 0, 1, \ldots$, and $\hat{\mu}(x)$ is a local linear estimate. We claim the double smooth is quadratic reproducing:

$$\hat{\mu}(x_i) \quad = \quad \frac{\sum W(j/h)(i+j)^2}{\sum W(j/h)}$$

$$
\begin{aligned}
&= i^2 + c \\
\hat{\tilde{\mu}}(x_i) &= i^2 + 2c \\
\hat{b}(x_i) &= \hat{\tilde{\mu}}(x_i) - \hat{\mu}(x_i) = c \\
\tilde{\mu}(x_i) &= \hat{\mu}(x_i) - \hat{b}(x_i) \\
&= i^2.
\end{aligned}
$$

That is, by Henderson's theorem, the 'double smoothed' estimate is just a local quadratic estimate (the condition on sign changes can be checked in individual cases).

Similarly, other constructions of bias corrected estimates, for example, boundary kernels, higher order kernels, and methods based on derivative estimation, are attempts to ensure the estimate asymptotically satisfies higher order reproducing properties. Local quadratic regression achieves the same results, much more directly and with fewer assumptions.

3 Model Assessment and Selection

The main tools for assessing the performance of local fits are simple generalizations of the tools used in parametric modelling. These include graphical tools for looking at the fit and residuals, and formal model selection criteria such as cross validation and C_p (Mallows 1973).

Before proceeding, we should make some general comments about the use of model selection criteria. In general, *no model selection criteria can be expected to automatically produce the best fit.* One reason for this is simple: the 'best fit' depends not only on the data, but the questions being asked.

3.1 Residual Analysis

3.2 Cross Validation and CP

Cross validation methods attempt to estimate the predictive power of a local fit: How well does $\hat{\mu}(x_{\text{new}})$ predict a new observation Y_{new}? The prediction mean squared error is

$$
\text{PMSE}(\hat{\mu}) = E(Y_{\text{new}} - \hat{\mu}(x_{\text{new}}))^2. \tag{6}
$$

Clearly, PMSE($\hat{\mu}$) depends on assumptions made about x_{new}. For the present, we suppose the design points $x_1, \ldots, x_n$ are an independent sample from a density $f(x)$, and the new point x_{new} is sampled from the same density.

Leave-one-out cross validation estimates PMSE($\hat{\mu}$), by successively deleting each observation (x_i, Y_i) from the dataset, and obtaining the local regression estimate $\hat{\mu}_{-i}(x_i)$ from the remaining $n-1$ observations. Formally, the cross validation score is

$$\mathrm{CV}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{-i}(x_i))^2. \tag{7}$$

The leave-one-out cross validation criteria was introduced for parametric regression models by Allen (1974) as the PRESS (prediction error sum of squares) procedure. Model validation based on splitting datasets into *estimation data* and *prediction data* has a long history, discussed for example in Stone (1974) and Snee (1977). The generalized cross validation criterion was first proposed in the context of smoothing splines by Craven and Wahba (1979). This provides an approximation to cross validation, and is easier to compute:

$$\mathrm{GCV}(\hat{\mu}) = n\frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2}{(n-\nu_1)^2},$$

where ν_1 is the fitted degrees of freedom defined by (4).

The cross validation scores $\mathrm{CV}(\hat{\mu})$, or $\mathrm{GCV}(\hat{\mu})$, can be used to compare the predictive power of different estimates – obtained for example by varying smoothing parameters or local polynomial degree, or entering new variables into the model.

The cross validation methods are motivated by prediction error: how well does $\hat{\mu}(x)$ estimate new observations? Alternatively, one can consider estimation error: How well does $\hat{\mu}(x)$ estimate the true mean $\mu(x)$? One possible loss criterion is the sum of the squared error over the design points;

$$L(\hat{\mu}, \mu) = \sum_{i=1}^{n}(\hat{\mu}(x_i) - \mu(x_i))^2. \tag{8}$$

The CP statistic criterion, introduced by Mallows (1973) for parametric regression provides an unbiased estimate of $L(\hat{\mu}, \mu)$, in the sense that

$$E(\mathrm{CP}(\hat{\mu})) = E(L(\hat{\mu}, \mu)).$$

The CP statistic was extended to local constant fitting by Rice (1984) and to local regression by Cleveland and Devlin (1988). The CP estimate of risk for a local regression estimate $\hat{\mu}(x)$ is

$$\mathrm{CP}(\hat{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2 - n + 2\nu_1.$$

3.3 Using Cross Validation

A simple use of cross validation is to select the model with the lowest score $\mathrm{CV}(\hat{\mu})$ as the best fit. We don't generally recommend this approach. The reason is model uncertainty: different models, producing very different fits to the data, can have similar predictive power, and similar cross validation scores.

Figure 1 displays a simple example. For the ethanol dataset, we compute the GCV scores for local quadratic fits with a range of nearest neighbor smoothing parameters: $0.2 \leq \alpha \leq 0.8$. The resulting **cross validation plot** is very flat; as the smoothing parameter changes from $\alpha = 0.2$ to $\alpha = 0.6$, the fitted degrees of freedom decreases from 16.4 to 5.6, and the GCV score ranges from 0.107 to 0.127.

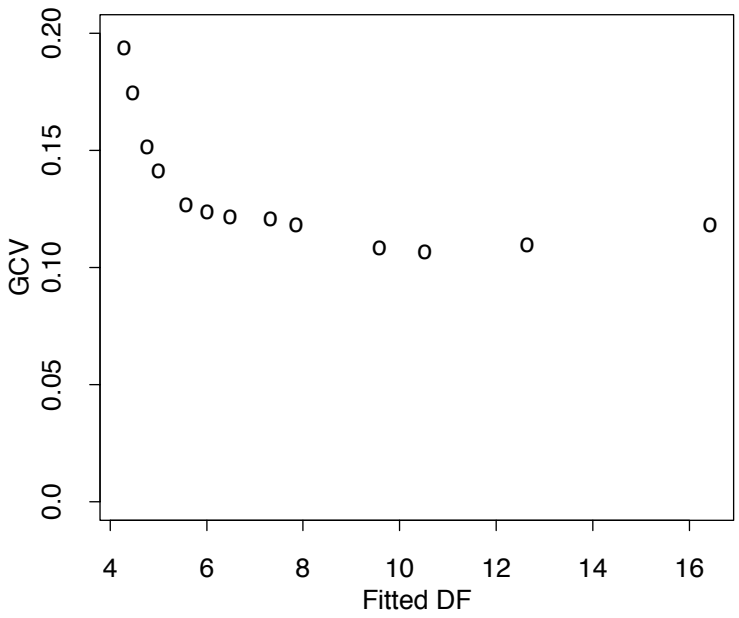

Figure 1: Generalized cross validation scores for the ethanol dataset. The horizontal axis shows fitted degrees of freedom; the corresponding smoothing parameters range from $\alpha = 0.8$ (on the left) to $\alpha = 0.2$ (on the right).

The cross validation plot in Figure 1 is similar in principle to the M plot (using the CP statistic) introduced in Cleveland and Devlin (1988). An important point in the construction is the use of the fitted degrees of

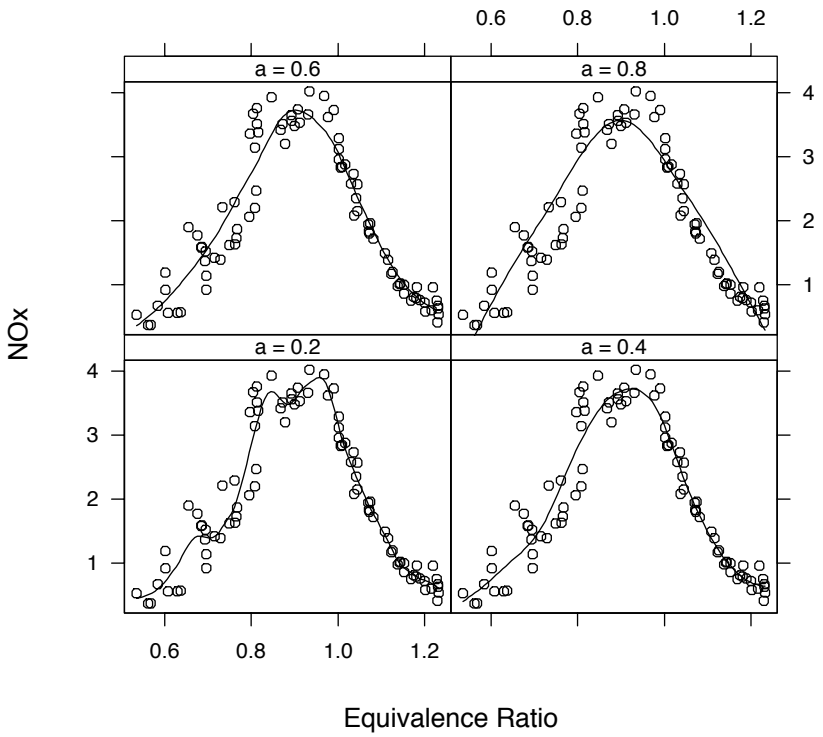

Figure 2: Smoothing the ethanol dataset: Local quadratic smoothing, with four choices of smoothing parameter.

freedom, rather than the smoothing parameter, as the horizontal axis. This aids interpretation: 4 degrees of freedom represents a smooth model with very little flexibility, while 16 parameters represents a noisy model showing many features. It also aids comparability; for example, we could easily compute cross validation scores for other polynomial degrees, or other smoothing methods, and add them to the plot.

The cross validation plot must be emphasized as *a graphical aid in choosing smoothing parameters*. To obtain full value from the plot, it must be used in conjunction with plots of the resulting fits. Figure 2 shows the fits for four of the smoothing parameters $\alpha = 0.8$, $\alpha = 0.6$, $\alpha = 0.4$ and $\alpha = 0.2$. The largest smoothing parameter, $\alpha = 0.8$, tracks the data poorly in several regions. This estimate clearly does not fit the data, and correspondingly the GCV plot rejects the fit.

On the other hand, making a definitive judgement between the remaining

fits is difficult. The fits show different features: for example, the bimodality at $\alpha = 0.2$. One can point to small clusters of observations that support bimodality, but it is unclear whether or not these represent a real feature. This is reflected by the flat GCV plot, which fails to make a definitive choice between the models. Flat plots, such as Figure 1, occur quite frequently, and any model with a GCV score near the minimum is likely to have similar predictive power. *The flatness of the plot reflects the uncertainty in the data, and the resultant difficulty in choosing smoothing parameters.*

A consequence of Figure 1 is that going to extensive lengths to minimize GCV is very data-sensitive and can produce an unsatisfactory fit. In general, minimizing GCV (or CP, or CV) is highly variable: two visually similar datasets could produce very different results. Most importantly, just minimizing GCV discards significant information provided by the whole profile of the GCV curve, as displayed by the cross validation plot.

At best, criterion such as GCV only provide a guide as to a reasonable range of smoothing parameters. The correct way to justify a choice of smoothing parameters in practice is to provide appropriate plots - both of the fitted curve and residuals - showing that the resulting fit reasonably tracks the data, without missing important features, and without showing undue noise.

We should emphasize that the points raised above are *not* a problem with cross validation, but a reflection of the difficulty of model selection. This point is discussed further in Loader (1999) and Chapter 10 of Loader (1998), where cross validation methods are compared with bandwidth selectors claimed to be less variable. Such selectors are found to reflect the model selection difficulty in other ways; in particular, missing features when applied to difficult smoothing problems.

4 Local Likelihood

The local least squares approach is most appropriate when the errors ϵ_i in are normally distributed, with a common variance. Local likelihood estimation (Tibshirani and Hastie 1987) provides a generalization when other assumptions are appropriate.

For example, in a binary model, the responses Y_i are 0 or 1, and the mean function is $\mu(x_i) = P(Y_i = 1) = 1 - P(Y_i = 0)$. The local log-likelihood for this model is

$$\sum_{i=1}^{n} w_i(x) \left(Y_i \log(\mu(x_i)) + (1 - Y_i) \log(1 - \mu(x_i)) \right).$$

One then imposes a local polynomial approximation. Since we know in advance $0 \leq \mu(x) \leq 1$, it is usual to use the **logistic link function**, so the local quadratic approximation becomes

$$\log\left(\frac{\mu(x_i)}{1-\mu(x_i)}\right) \approx a_0 + a_1(x_i - x) + a_2(x_i - x)^2.$$

In general, a likelihood regression model can be written

$$Y_i \sim f(y, \theta(x_i))$$

where $f(y, \theta)$ is a specified parametric family of distributions. The local log-likelihood, with a local polynomial approximation, is

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) \log f(Y_i, \langle a, A(x_i - x)\rangle). \tag{9}$$

Let $\hat{a}$ be the maximizer of the local likelihood. The local likelihood estimate of $\theta(x)$ is simply the constant coefficient:

$$\hat{\theta}(x) = \langle \hat{a}, A(0)\rangle = \hat{a}_0.$$

Under mild regularity conditions, the maximizer $\hat{a}$ of the local likelihood $\mathcal{L}_x(a)$ must be a solution of the system of **local likelihood equations**, obtained by differentiating (9):

$$\sum_{i=1}^{n} w_i(x) A(x_i - x) \dot{l}(Y_i, \langle \hat{a}, A(x_i - x)\rangle) = 0 \tag{10}$$

where $l(y, \theta) = \log f(y, \theta)$ and $\dot{l}(y, \theta)$ denotes the partial derivative with respect to θ.

When $f(y, \theta)$ is the Gaussian density, the local likelihood procedure is equivalent to local least squares regression. For most other likelihoods, there is no closed form for the local likelihood estimate, so iterative procedures must be used to solve (10). For many common likelihoods and link functions, $l(y, \theta)$ is a concave function. It follows that $\mathcal{L}_x(a)$ is also concave,[1] and the maximizer $\hat{a}$ exists and is unique.

Exact distribution theory for the local likelihood estimate is unavailable, except in the Gaussian case. But useful approximations can be derived using a linearization of the likelihood around the true parameter $\theta(x)$. This

[1] excluding singularities in very small samples

is exactly analagous to the methods used for parametric generalized linear models in McCullagh and Nelder (1989) and elsewhere.

Let $\tilde{a}$ be the coefficients of the Taylor series of $\theta(x)$, up to order p. Linearizing the local likelihood equations gives

$$\begin{aligned} 0 &= \mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\hat{a}) \\ &= \mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a}) - \mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{X}(\hat{a} - \tilde{a}) + o(\|\hat{a} - \tilde{a}\|). \end{aligned}$$

where $\mathbf{V}$ is a diagonal matrix with elements $-\ddot{l}(Y_i, \theta_i)$. Solving for $\hat{a}$ gives the representation

$$\hat{a} \approx \tilde{a} + (\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a}). \tag{11}$$

Let $\tilde{\theta}_i$ be the components of $\mathbf{X}\tilde{a}$. The next step is to expand the terms $\dot{l}(Y_i, \tilde{\theta}_i$ around the true parameter $\theta(x_i)$:

$$\begin{aligned} \dot{l}(Y_i, \tilde{\theta}_i) &= \dot{l}(Y_i, \theta(x_i)) + (\tilde{\theta}_i - \theta(x_i))\ddot{l}(Y_i, \theta(x_i)) + o(\tilde{\theta}_i - \theta(x_i)) \\ &= \dot{l}(Y_i, \theta(x_i)) - \frac{(x_i - x)^{p+1}}{(p+1)!}\theta^{(p+1)}(x_i)\ddot{l}(Y_i, \theta(x_i)) + o(h^{p+1}). \end{aligned}$$

Substituting this back into (11) gives the decomposition

$$\begin{aligned} \hat{a} \approx\ & \tilde{a} + \mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \theta) \\ & -\frac{\theta^{(p+1)}(x)}{(p+1)!}\mathbf{J}_1^{-1}\sum_{i=1}^{n} w_i(x)(x_i - x)^{p+1}A(x_i - x)\ddot{l}(Y_i, \theta(x_i)) \end{aligned} \tag{12}$$

where $\mathbf{J}_j = \mathbf{X}^T\mathbf{W}^j\mathbf{V}\mathbf{X}$. In particular, this yields the approximate variance for the parameter vector $\hat{a}$:

$$\mathrm{var}(\hat{a}) \approx \mathbf{J}_1^{-1}\mathbf{J}_2\mathbf{J}_1^{-1}.$$

4.1 Local Quasi-Likelihood

One of the assumptions made when motivating a least squares procedure is that of a constant residual variance. If this assumption is violated, can modify the local least squares criterion (1) to

$$\sum_{i=1}^{n} \frac{w_i(x)}{\sigma_i^2}(Y_i - \langle a, A(x_i - x)\rangle)^2 \tag{13}$$

where $\sigma_i^2 = \mathrm{var}(Y_i)$. If these variances are known, one can easily solve the least squares problem (13), yielding the local parameter estimates

$$\hat{a} = (\mathbf{X}^T\mathbf{V}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}Y \tag{14}$$

where $\mathbf{V}$ is a diagonal matrix with elements $1/\sigma_i^2$.

The estimate (14) is only useful if the variances σ_i^2 are known, at least up to a multiplicative constant. More generally, we have to impose some structure on the variance function. The quasi-likelihood model, introduced by Wedderburn (1974), assumes a functional relationship between the variance and mean function:

$$\text{var}(Y_i) = \sigma^2 V(\mu),$$

where $V(\mu)$ is a known function. This model can be fitted iteratively: Initialize $\mathbf{V} = \mathbf{I}$, then alternately estimate $\hat{a}$ (and hence $\hat{\mu}(x_i)$) and $\mathbf{V} = \text{diag}(1/V(\hat{\mu}(x_i)))$ until convergence. The limiting $\hat{a}$ (assuming it exists) must be a stationary point of (14), or equivalently, satisfy the system of equations

$$\mathbf{X}^T\mathbf{W}\mathbf{V}(Y - \mathbf{X}\hat{a}) = 0.$$

These equations look very much like the local likelihood equations (10), with the equivalence

$$\dot{l}(Y_i, \mu_i) = \frac{1}{\sigma^2 V(\mu_i)}(Y_i - \mu_i),$$

Family	Mean	Variance
Gaussian	μ	σ^2
binomial	μ	$\sigma^2\mu(1-\mu)$
Poisson	μ	$\sigma^2\mu$
gamma	μ	$\sigma^2\mu^2$
geometric	μ	$\sigma^2\mu(\mu+1)$

Table 1: Likelihood families and their Quasi-variance functions.

Table 1 lists some common likelihood models, and their corresponding quasi-variance functions.

4.2 Robust Fitting

Local least squares estimates can be motivated as local likelihood estimates when the errors ϵ_i are assumed to have a Gaussian distribution. However, when the ϵ_i have a long tailed distribution, a least squares estimate can be overly sensitive to extreme observations. To make the local least squares more robust, we need to identify and downweight these extreme observations.

Such a scheme was proposed as part of the LOWESS procedure of Cleveland (1979). The algorithm used by Cleveland is:

1. Assign all observations a prior weight $v_i = 1$.
2. Smooth the data, with prior weights v_i.
3. Compute the residuals $\hat{\epsilon}_i = Y_i - \hat{\mu}(x_i)$, and estimate the scale s as the median of the absolute values of the residuals.
4. Assign observations prior weights
$$v_i = (1 - \hat{\epsilon}_i^2/(6s^2))_+^2.$$
5. Repeat steps 2, 3 and 4 until convergence.

The robustness arises from the downweighting at the fourth step. An observation with $\hat{\epsilon}_i = 0$ receives robustness weight 1, while an observation with $|\hat{\epsilon}_i| > \sqrt{6}s$ receives robustness weight 0.

LOWESS performs the robustness iterations in a global sense: the entire local regression curve is computed, then the robustness weights computed, and so on. An alternative is to perform the iterations locally for each fitting point. This amounts to solving the system of equations

$$\sum_{i=1}^{n} w_i(x) A(x_i - x) \hat{\epsilon}_i B(\hat{\epsilon}_i/s) = 0$$

where B is the 'downweighting' function, and s is a local scale estimate; for example, the median absolute residual over all observations with $w_i(x) > 0$.

Katkovnik (1979), Katkovnik (1985) and Tsybakov (1986) considered versions of M-estimation for local regression. The least squares criterion is replaced by a criterion of the form

$$\sum_{i=1}^{n} w_i(x) \rho\left(\frac{Y_i - \langle a, A(x_i - x)\rangle}{s}\right) \tag{15}$$

where $\theta_i = \langle a, A(x_i - x)\rangle$; $\rho(\cdot)$ is a symmetric concave function, and s is a scale parameter. Ordinary local regression corresponds to the case $\rho(z) = z^2$. The minimizer of the criterion (15) satisfies the system of equations

$$\sum_{i=1}^{n} w_i(x) A(x_i - x) \rho'(\hat{\epsilon}_i/s) = 0. \tag{16}$$

This shows M estimation and local downweighting are equivalent with appropriate choice of criteria. The system (16) can be solved by iterated least squares.

Intuitively, the function $\rho(v)$ should be symmetric, and increasing on $[0, \infty)$. Since we want the result to be less sensitive to outliers than the least squares estimate, $\rho(v)$ should increase more slowly than a quadratic. One possible choice is

$$\begin{aligned} \rho_c(v) &= x^2 I(|v| < c) + (2c|v| - c^2) I(|v| \geq c) \\ \rho_c'(v) &= -2cI(v \leq -c) + 2vI(|v| < c) + 2cI(v \geq c) \\ \rho_c''(v) &= 2I(|v| < c) \end{aligned} \tag{17}$$

where c is a prespecified constant. For a fixed dataset, the resulting estimate converges to the local least squares estimate as $c \to \infty$, and to the local L_1 estimate (Katkovnik 1985, Section 7.3; Wang and Scott 1994

If the errors ϵ_i have a symmetric density $g(v)$, the variance of the estimate $\hat{\mu}(x)$ is approximately

$$\text{var}(\hat{\mu}(x)) \approx \frac{\int \dot{\rho}(v)^2 g(v) dv}{(\int \ddot{\rho}(v) g(v) dv)^2} \|l(x)\|^2. \tag{18}$$

This follows from a derivation similar to the local likelihood variance. The optimal choice of ρ is $\rho(v) = -\log(g(v))$; in this case, the local robust regression becomes local likelihood estimation. Tsybakov (1986) gave results similar to (18).

Since the density $g(v)$ will usually be unknown, one must choose a function $\rho(v)$ that exhibits good behavior over some class of densities $g(v)$. For $\rho(v) = v^2$, (18) reduces to $\sigma^2 \|l(x)\|^2$, which is poor for heavy tails. Choosing $\rho(v) = |v|$ leads to $\|l(x)\|^2/(4g(0)^2)$, which is poor when $g(0)$ is small; for example, a bimodal density. The choice (17), with a robust estimate for the scale parameter, is a compromise between these two extremes.

4.3 Density Estimation

By choosing an appropriate likelihood function, the local likelihood method can be extended beyond simple regression models. We consider density estimation here; other models are discussed in Loader (1998).

Density estimation assumes we have an independent sample $X_1, \ldots, X_n$ from an unknown density $f(x)$. The object is to estimate $f(x)$. Since the density is non-negative, it is natural to use a local polynomial approximation

	$\rho_1(v)$	$\rho_2(v)$	$\rho_3(v)$	$\rho_4(v)$
Normal	1.000	1.571	1.313	1.010
Double Exp.	2.000	1.000	1.179	1.607
Cauchy	∞	2.467	2.000	4.679

Table 2: Robustness variance factors for three densities, and four robustness functions: $\rho_1(v) = v^2/2$; $\rho_2(v) = |v|$; $\rho_3(v) = -\log(1+x^2)$ and $\rho_4(v)$ defined by (17).

for $\log f(x)$. The local likelihood for this model is

$$\begin{aligned}\mathcal{L}_x(a) &= \sum_{j=1}^{n} W\left(\frac{X_j - x}{h}\right)\langle a, A(X_j - x)\rangle \\ &\quad -n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)\exp(\langle a, A(u-x)\rangle)du.\end{aligned}$$

Letting $\hat{a}$ be the maximizer of the local likelihood, the local likelihood density estimate is

$$\hat{f}(x) = \exp(\hat{a}_0).$$

As in the local likelihood regression setting, differentiating the local likelihood yields the system of local likelihood equations:

$$\frac{1}{n}\sum_{j=1}^{n} A(X_j - x)w_j(x) = \int_{\mathcal{X}} A(u-x)W\left(\frac{u-x}{h}\right)\exp(\langle \hat{a}, A(u-x)\rangle)du. \quad (19)$$

These equations have a very simple and intuitive interpretation. The left hand side of (19) is simply a vector of localized sample moments up to order p, while the right hand side is localized population moments using the log-polynomial density approximation. The local likelihood estimate simply matches localized sample moments with localized population moments.

If local constants are fitted, (19) consists of the single equation

$$\frac{1}{n}\sum_{j=1}^{n} W\left(\frac{X_j - x}{h}\right) = \int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)\exp(\hat{a}_0)du,$$

yielding the closed form for the density estimate

$$\hat{f}(x) = \exp(\hat{a}_0) = \frac{1}{nh\int W(v)dv}\sum_{j=1}^{n} W\left(\frac{X_j - x}{h}\right). \quad (20)$$

This is the kernel density estimate considered by Rosenblatt (1956); Whittle (1958) and Parzen (1962).

The kernel density estimate has been widely studied since its introduction in the 1950's; see for example the books by Silverman (1986) or Scott (1992). Being based on a local constant approximation, it suffers from the same problems as local constant regression, such as trimming of peaks. An additional problem occurs in the tails, since increasing bandwidths for data sparsity can lead to severe bias. This problem was investigated more fully by Loader (1996) where relative efficiencies of kernel and local log-polynomial methods were compared.

5 Statistical Inference

In this section we study inferential issues for local regression: variance estimation; nonhomogeneous variance; confidence intervals and bands; and testing goodness of fit. Although these topics are only vaguely related, some tools are used repeatedly and so treating them together is appropriate.

Although we formulate the results for local regression, the methods and tests can be applied to local likelihood models.

5.1 Variance Estimation

The residual variance σ^2 represents variation in the response variable that cannot be explained by the predictors. This represents a lower bound on the prediction mean squared error for future observations: whatever estimate is used for $\hat{\mu}(x)$, the squared prediction error for future observations will be at least σ^2:

$$E((Y_{\text{new}} - \hat{\mu}(x_{\text{new}}))^2) = \sigma^2 + E((\hat{\mu}(x_{\text{new}}) - \mu(x_{\text{new}}))^2).$$

It is thus of fundamental importance to estimate σ^2. The variance also arises in many diagnostic statistics, including the CP statistic introduced earlier, and confidence bands and the goodness of fit tests introduced later in this section.

In analogy with parametric regression, the natural variance estimate is the normalized residual sum of squares:

$$\hat{\sigma}^2 = \frac{1}{n - 2\nu_1 + \nu_2} \sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2. \tag{21}$$

The expected value is

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{1}{n - 2\nu_1 + \nu_2} \sum_{i=1}^{n} \text{bias}(\hat{\mu}(x_i))^2.$$

In particular, $\hat{\sigma}^2$ is unbiased when the estimate $\hat{\mu}(x)$ is unbiased. This of course is rarely true in the nonparametric regression setting, although if the bandwidth is not too large, it may be reasonable to assume the bias is negligible.

The variance estimate can be written as a quadratic form:

$$\hat{\sigma}^2 = \frac{1}{\text{tr}(\mathbf{\Lambda})} Y^T \mathbf{\Lambda} Y \tag{22}$$

where $\mathbf{\Lambda} = (\mathbf{I} - \mathbf{L})^T(\mathbf{I} - \mathbf{L})$. Thus we can find the distribution by applying the general theory of quadratic forms. In particular, if the errors ϵ_i are normally distributed with variance σ^2 and $\hat{\sigma}^2$ is unbiased (i.e. $\mu^T \mathbf{\Lambda} \mu = 0$), the distribution of the quadratic form is

$$\frac{1}{\sigma^2} Y^T \mathbf{\Lambda} Y \stackrel{\mathcal{L}}{=} \sum_{j=1}^{n} \lambda_j Z_j$$

where λ_j are the eigenvalues of $\mathbf{\Lambda}$ and Z_j are independent χ^2_1 random variables.

For a parametric regression model, the λ_j are all 0 or 1 (i.e. $\mathbf{\Lambda}$ is idempotent), and $Y^T \mathbf{\Lambda} Y$ has a χ^2 distribution. For local regression variance estimates, this simplification no longer holds. The exact distribution of quadratic forms can be found by inverting characteristic functions (Imhof 1961, Davies 1980).

A simple approximate distribution, useful for inferential purposes, is provided by Satterthwaite (1946). This involves matching the mean and variance of the quadratic form to those of a χ^2 distribution. The approximation was applied to local regression problems by Cleveland (1979); simulations studying the accuracy are found in Cleveland and Devlin (1988).

The approximation is based. on the mean and variance of the quadratic form. If ϵ is a multivariate normal distribution with the identity covariance,

$$\begin{aligned} E(\epsilon^T \mathbf{\Lambda} \epsilon) &= \delta_1 \\ \text{var}(\epsilon^T \mathbf{\Lambda} \epsilon) &= 2\delta_2 \end{aligned}$$

α	δ_1	δ_2	δ_1^2/δ_2
0.7	82.52	82.24	82.79
0.3	75.51	75.23	75.79
0.1	45.67	43.81	47.62

Table 3: One-moment and two-moment chi-square approximations. Comparing degrees of freedom for the ethanol dataset.

where $\delta_1 = \mathrm{tr}(\mathbf{\Lambda})$ and $\delta_2 = \mathrm{tr}(\mathbf{\Lambda}^2)$. Thus, letting $\nu = \delta_1^2/\delta_2$, and assuming the unbiasedness of $\hat{\sigma}^2$,

$$\begin{aligned} E(\frac{\nu\hat{\sigma}^2}{\sigma^2}) &= \nu \\ \mathrm{var}(\frac{\nu\hat{\sigma}^2}{\sigma^2}) &= 2\nu. \end{aligned}$$

Thus, we use the distributional approximation

$$\frac{\nu\sigma^2}{\sigma^2} \sim \chi^2_\nu.$$

The computation of the degrees of freedom (and in particular δ_2) in the chi-square approximation can be expensive, especially for large sample sizes. In light of (22) it is very tempting to use $\delta_1 = \mathrm{tr}(\mathbf{\Lambda})$ as the degrees of freedom for the approximating chi-square distribution. While this is dangerous for general quadratic forms, it is usually an acceptable approximation for the residual sum of squares. Table 3 presents a small comparison of the degrees of freedom for the residual sum of squares for the ethanol dataset, using the one and two moment approximations. Three different smoothing parameters are used. Only the very small $\alpha = 0.1$ shows large discrepancy between the two approximations.

So far, we have assumed the residual variance σ^2 is a constant. A more sophisticated model assumes the variance $\mathrm{var}(Y_i)$ varies in a smooth manner with the covariates x_i:

$$\mathrm{var}(Y_i) = \sigma^2(x_i).$$

This model can be fitted simply by smoothing squared residuals from a local regression fit. Since the variance is a scale parameter, it makes sense to use a local likelihood fit with the gamma family.

5.2 Pointwise Confidence Intervals

A local regression estimate may show different features in a dataset, such as multiple peaks or regions of strong dependence. Moreover, features seen in a fit are often strongly dependent on the bandwidth used, and from the estimate alone one may be unsure which features are real.

Interval estimates for the mean function $\mu(x)$ attempt to make confidence statements about the true mean and hence help provide a more formal basis for judging what features in a dataset are real. An interval estimate has the form $(L(x), U(x))$. The limits $L(x)$ and $U(x)$ are data based quantities, chosen so that $L(x) \leq \mu(x) \leq U(x)$ with high confidence. The interval $(L(x), U(x))$ is a $(1-\alpha)100\%$ pointwise confidence interval for $\mu(x)$ if

$$\sup_{\mu \in \mathcal{F}} P_\mu(L(x) \leq \mu(x) \leq U(x)) \geq 1 - \alpha. \tag{23}$$

Here, $\mathcal{F}$ denotes a suitable class of smooth functions.

Assuming ϵ_i are independent Gaussian random variables with mean 0 and variance σ^2, a local polynomial estimate $\hat{\mu}(x)$ has the distribution

$$\frac{\hat{\mu}(x) - E\mu(x)}{\sigma \|l(x)\|} \sim N(0,1).$$

If we assume no bias, so $E\mu(x) = \mu(x)$, confidence intervals may take the form

$$I_1(x) = (\hat{\mu}(x) - c\hat{\sigma}\|l(x)\|, \hat{\mu}(x) + c\hat{\sigma}\|l(x)\|), \tag{24}$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution.

Of course, the set of functions for which $E\hat{\mu}(x) = \mu(x)$ is quite small. With a small bandwidth, it may be reasonable to assume $E\hat{\mu}(x) \approx \mu(x)$, and $I_1(x)$ is an approximate confidence interval for $\mu(x)$. This 'undersmoothing' provides the easiest and most practical solution to the bias problem; however, one needs to beware that small bandwidths imply $\text{var}(\hat{\mu}(x))$ is large, and wide confidence intervals may result.

An alternative is to adjust the intervals to allow for bias. If $b(x) = E\hat{\mu}(x) - \mu(x)$, a bias corrected confidence interval is

$$I_2(x) = (\hat{\mu}(x) - b(x) - c\hat{\sigma}\|l(x)\|, \hat{\mu}(x) - b(x) + c\hat{\sigma}\|l(x)\|).$$

Since $b(x)$ is unknown, we need to use a bias estimates $\hat{b}(x)$ to form an estimated confidence intervals $\hat{I}_2(x)$. But some thought - and application of Henderson's theorem - shows this approach doesn't solve the bias problem.

Bias estimates attempt to remove leading terms from asymptotic expansions of $E(\hat{\mu}(x)$ – either by double smoothing, as in Section 2.5, or by derivative estimation. But Henderson's theorem implies these constructions, at least asymptotically, simply increase the order of the estimate. In this case an estimated $\hat{I}_2(x)$ is just an undersmoothed interval centered around the higher order estimate $\hat{\mu}(x) - \hat{b}(x)$. One has the additional problem that $\text{var}(\hat{\mu}(x) - \hat{b}(x))$ may be larger than $\text{var}(\hat{\mu}(x))$.

Another approach to bias adjustment is to focus on the class of smooth functions $\mathcal{F}$ in (23). For example, if $\mathcal{F}_\delta$ is defined to be the class of functions for which $|b(x)| \leq \delta$, then

$$I_3(x) = (\hat{\mu}(x) - \delta - c\hat{\sigma}||l(x)||, \hat{\mu}(x) + \delta + c\hat{\sigma}||l(x)||)$$

is a confidence interval for $\mu(x)$. Note the difference between $I_2(x)$ and $I_3(x)$: while $I_2(x)$ attempts to recenter the bands to allow for bias, $I_3(x)$ expands the bands. This type of expansion was used by Knafl, Sacks, and Ylvisaker (1985) to construct simultaneous confidence bands. Sharper results, in which one attempts to adjust c rather than expanding by δ, were considered in Sun and Loader (1994).

5.3 Simultaneous Confidence Bands

Many questions depend on more than just single values of $\mu(x)$. For example, we may be interested in comparison of mean responses at different levels of the covariate x, or choosing a level of the covariates to maximize the mean response. Thus, we are also interested in constructing simultaneous confidence bands over a set $\mathcal{X}$, where $\mathcal{X}$ is typically taken to be a set bounding the predictors x_i. The band $\{(L(x), U(x)); x \in \mathcal{X}\}$ is a $(1-\alpha)100\%$ simultaneous confidence band if

$$\sup_{\mu \in \mathcal{F}} P_\mu(L(x) \leq \mu(x) \leq U(x) \, \forall \, x \in \mathcal{X}) \geq 1 - \alpha.$$

The construction of simultaneous confidence bands is similar to confidence intervals: begin with a band $\{I_1(x); x \in \mathcal{X}\}$ that is valid under the assumption of no bias, and then adjust the bands to allow for bias. The issues involved in bias estimation and adjustment are the same for simultaneous bands as they are for the pointwise intervals. Thus, our focus is on construction of the 'no bias' bands.

Under the no bias assumption, a confidence band $\{I_1(x); x \in \mathcal{X}\}$ will cover the true mean $\mu(x)$ if, and only if,

$$M_\sigma = \sup_{x \in \mathcal{X}} \frac{|\hat{\mu}(x) - \mu(x)|}{\sigma||l(x)||} \leq c.$$

To find the critical value c, we need to find the distribution of M_σ. For linear regression, Scheffé (1959) showed the distribution of M_σ is related to an F distribution when $\mathcal{X} = \mathcal{R}^d$. For local regression the exact distribution of M_σ is quite intractable and approximations must be used.

The results we used are based on approximations for tail probabilities for Gaussian stochastic processes and random fields. When $\mathcal{X}$ is one dimensional, results can be derived by counting the number of upcrossings $N(c)$ the process $(\hat{\mu}(x) - \mu(x))/\sigma\|l(x)\|$ makes over the level c. For large values of c, the probability of multiple upcrossings is small, and

$$P(M_\sigma \geq c) \approx E(N(c)).$$

Remarkably, for Gaussian processes, the expected number of upcrossings has a simple closed form expression. The first result of this kind was derived by Rice (1939); see Section 7.2 of Leadbetter, Lindgren, and Rootz'en (1983) for more discussion and references.

To state the result in the confidence band setting, let $T(x) = l(x)/\|l(x)\|$. We assume $\{T(x); x \in \mathcal{X}\}$ is continuous (this is satisfied for local regression models, provided the weight function $W(v)$ is continuous). Then

$$P(M_\sigma \geq c) \leq 2(1 - \Phi(c)) + \frac{\kappa_0}{\pi} e^{-c^2/2}$$

where κ_0 is the length of the path $\{T(x); x \in \mathcal{X}\}$. Explicitly, if $\{T(x)\}$ is differentiable and $\mathcal{X} = [a, b]$,

$$\kappa_0 = \int_a^b \|T'(x)\| dx.$$

A discretized version of this result was given by Knafl, Sacks, and Ylvisaker (1985). In multiple dimensions the simple upcrossing approach no longer works, and more sophisticated geometric arguments must be used. Results from Sun (1993) and Sun and Loader (1994) yield a series approximation of the form

$$\begin{aligned} P(M_\sigma \geq c) \quad \approx \quad & \kappa_0 \frac{\Gamma(\frac{d+1}{2})}{\pi^{(d+1)/2}} P(\chi^2_{d+1} > c^2) + \frac{\zeta_0}{2} \frac{\Gamma(\frac{d}{2})}{\pi^{d/2}} P(\chi^2_d > c^2) \\ & + \frac{\kappa_2 + \zeta_1 + m_0}{2\pi} \frac{\Gamma(\frac{d-1}{2})}{\pi^{(d-1)/2}} P(\chi^2_{d-1} > c^2) + \ldots \end{aligned} \tag{25}$$

where κ_0, ζ_0, κ_2, ζ_1 and m_0 are certain geometric constants. In particular, κ_0 represents the area, or volume of the set $\mathcal{I} = \{T(x) : x \in \mathcal{X}\}$. Explicitly, we can write

$$\kappa_0 = \int_{\mathcal{X}} \det{}^{1/2}[\langle T_i(x), T_j(x) \rangle] dx,$$

where $T_i(x)$ denotes the partial derivative of $T(x)$ in the ith direction, and $[\cdot]$ denotes a matrix with the given (i,j)th elements. ζ_0 is the volume of the boundary of $\mathcal{X}$ and m_0 is related to the corners of $\mathcal{X}$. κ_2 and ζ_1 are more complicated constants related to the curvature of $\mathcal{X}$ and the boundary of $\mathcal{X}$ respectively. When $d = 2$ it is known that $\kappa_0 + \kappa_2 + \zeta_1 + m_0 = 2\pi$ for simple sets $\mathcal{X}$, such as rectangles.

When σ is unknown but estimated by $\hat{\sigma}^2$ with ν degrees of freedom, we get the approximation

$$\begin{aligned}\alpha \approx\ & \kappa_0 \frac{\Gamma(\frac{d+1}{2})}{\pi^{(d+1)/2}} P(F_{d+1,\nu} > \frac{c^2}{d+1}) + \frac{\zeta_0}{2} \frac{\Gamma(\frac{d}{2})}{\pi^{d/2}} P(F_{d,\nu} > \frac{c^2}{d}) \\ & + \frac{\kappa_2 + \zeta_1 + m_0}{2\pi} \frac{\Gamma(\frac{d-1}{2})}{\pi^{(d-1)/2}} P(F_{d-1,\nu} > \frac{c^2}{d-1}).\end{aligned}$$

5.4 Goodness of Fit Testing

Suppose we are interested in testing the adequacy of a linear model for $\mu(x)$:

$$\begin{aligned}\mathcal{H}_0 : &\quad \mu(x) = a + bx \\ \text{vs}\mathcal{H}_1 : &\quad \text{otherwise}\end{aligned}$$

The general idea of a goodness of fit test is to fit models under both the null and alternative hypotheses, and compare the results. Under the null hypothesis, we fit a parametric least squares estimate. Under the alternative hypothesis, we fit a local polynomial model.

The problem of goodness of fit testing has been widely studied, and we do not attempt to review all available methods here. Instead, we present two quite different techniques. First we consider tests based on the residual sum of squares and likelihood ratios, using ideas similar to Cleveland and Devlin (1988). Second class of tests is based on maximal deviations of tests based on higher order coefficients of local polynomial expansions. Other references for goodness of fit testing include Cox, Koh, Wahba, and Yandell (1988), Azzalini, Bowman, and Härdle (1989), Raz (1990), Azzalini and Bowman (1993), Hjellvik, Yao, and Tjøstheim (1996) and Hart (1997).

Many methods have been proposed for approximating significance levels of tests. For our purposes, we use relatively simple approximations, based on Satterthwaite's method and on the upcrossing and tube methods. Formally, these methods are only justified under normality assumptions. But the reasons for selecting the simple approaches are two-fold. First, testing

goodness of fit is typically only a small part of the process of analyzing a data set, and applying expensive simulation methods solely to compute p-values seems excessive. Second, while many *claims* have been made about the performance of more sophisticated approaches and the assumptions such methods don't make, the *evidence* to support these claims is very weak.

We can easily form an F type test statistic based on a local polynomial regression estimate. Let $\hat{\alpha}_0 + \hat{\alpha}_1 x$ be the parametric least squares fit and $\hat{\mu}(x)$ be a local polynomial fit. Under these fits, we compute the residual sums of squares:

$$\begin{aligned} \mathrm{RSS}_0 &= \sum_{i=1}^{n} (Y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i))^2 \\ &= Y^T (\mathbf{I} - \mathbf{L}_0)^T (\mathbf{I} - \mathbf{L}_0) Y \\ \mathrm{RSS}_1 &= \sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2 \\ &= Y^T (\mathbf{I} - \mathbf{L}_1)^T (\mathbf{I} - \mathbf{L}_1) Y \end{aligned}$$

where $\mathbf{L}_0$ and $\mathbf{L}_1$ are the hat matrices for the global and local fits respectively.

$$F = \frac{(RSS_0 - RSS_1)/(\nu_0 - \nu_1)}{\hat{\sigma}^2} \tag{26}$$

where $\nu_j = \mathrm{tr}(\mathbf{I} - \mathbf{L}_j)^T (\mathbf{I} - \mathbf{L}_j)$ for $j = 0, 1$, and $\hat{\sigma}^2$ is computed under the local model. The distribution of the ratio (26) using either the one-moment or Satterthwaite approximation. Test statistics of this form were considered in Cleveland and Devlin (1988).

Test statistics based on quadratic forms are not always informative. Even if the test rejects goodness of fit of the null model, one gets no information as to what form the lack of fit takes. Maximal deviation tests attempt to address this problem by determining whether individual features are significant.

Suppose we are interested in testing the null hypothesis that $\mu(x)$ is constant. We estimate $\mu(x)$ using a local polynomial (linear or higher order) fit. Under the null hypothesis, the local slope estimate $\hat{\mu}'(x)$ has mean 0. As a test statistic, we can consider the scaled maximum of $\hat{\mu}'(x)$:

$$M = \sup_{x \in \mathcal{X}} \frac{|\hat{\mu}'(x)|}{\sqrt{\mathrm{var}(\hat{\mu}'(x))}} \tag{27}$$

for a suitable set $\mathcal{X}$. The problem of computing critical values for this test statistic is closely related to the simultaneous confidence band problem.

The statistic (27) provides a legitimate test for any bandwidth h. Different h will provide tests with different power. The tools used previously to assess bandwidths, such as cross validation, are not necessarily appropriate in the present setting. An informal solution to the bandwidth problem is to look at test statistics for multiple bandwidths, although this will not preserve a specified significance level.

A more formal procedure is to take the bandwidth as another 'dimension' in the random field. That is, consider the test statistic

$$M' = \sup_{x \in \mathcal{X}, h_0 \leq h \leq h_1} \frac{|\hat{\mu}'(x,h)|}{\sqrt{\text{var}(\hat{\mu}'(x,h))}}$$

for a suitable range $h_0 \leq h \leq h_1$ of bandwidths. Critical values for M' can be computed using 25, except the integral defining κ_0 must now be taken over both x and h. This approach was studied for a continuously observed process by Siegmund and Worsley (1995).

6 Extensions of Local Regression

The local regression method can be extended and modified in numerous ways. In this section, we introduce some of the most important.

6.1 Additive Models

The definition of multivariate local regression extends to any number of dimensions. But beyond two or three dimensions, a local regression model is difficult to fit, both due to the rapid increase in the number of parameters in the local model, and the sparsity of data in high dimensional space. In addition, visualization of a high dimensional surface is difficult.

Because of these problems, a number of simplified models have been proposed. Typically, these methods build a fitted surface by applying local regression (or other smoothers) to low dimensional projections of the data.

Additive models assume the regression surface is an additive function of predictors. With two predictors p_1 and p_2, the additive model is

$$\mu(p_1, p_2) = \mu_1(p_1) + \mu_2(p_2)$$

where $\mu_1(p_1)$ and $\mu_2(p_2)$ are smooth functions. The backfitting algorithm can be used to fit the model and alternately estimates the components.

1. Initialize $\hat{\mu}_1 = \hat{\mu}_2 = 0$.

2. Estimate $\hat{\mu}_1$ by smoothing $Y_i - \hat{\mu}_2(p_{2,i})$ against $p_{1,i}$.

3. Estimate $\hat{\mu}_2$ by smoothing $Y_i - \hat{\mu}_1(p_{1,i})$ against $p_{2,i}$.

4. Repeat 2) and 3) until convergence.

A thorough account of additive models and the backfitting algorithm can be found in Hastie and Tibshirani (1990). Opsomer and Ruppert (1997) discuss theoretical properties of the backfitting algorithm.

A special case of the additive model is a semiparametric model,

$$\mu_{x_1,x_2} = \mu_1(x_1) + \langle \beta, x_2 \rangle . \tag{28}$$

This model is particularly attractive since the backfitting algorithm has a closed form limit:

$$\hat{\beta} = \left(\mathbf{X}_2(\mathbf{I} - \mathbf{L}_1)\mathbf{X}_2^T \right)^{-1} \mathbf{X}_2^T (\mathbf{I} - \mathbf{L}_1) Y, \tag{29}$$

where $\mathbf{X}_2$ is the design matrix for the parametric component, and $\mathbf{L}_1$ is the hat matrix for the smooth component. See Hastie and Tibshirani (1990), page 118.

Using a slightly different approach, Eubank and Speckman (1993) arrive at a modified form of (29), using $(\mathbf{I} - \mathbf{L}_1)^T(\mathbf{I} - \mathbf{L}_1)$ in place of $\mathbf{I} - \mathbf{L}_1$. Other references on semiparametric models include Engle, Granger, Rice, and Weiss (1986), Green (1987) and Severini and Staniswalis (1994).

6.2 Conditionally Parametric Models

The conditionally parametric fit, introduced in Cleveland, Grosse, and Shyu (1992) and Cleveland (1994), provides a generalization of the semiparametric model (28), without the complexity of a full local fit. A conditionally quadratic model is

$$\hat{\mu}(p_1, p_2) = a_0(p_1) + a_1(p_1)p_2 + a_2(p_1)p_2^2$$

where a_0, a_1 and a_2 are smooth functions of p_1. Like the semiparametric model, for fixed p_1, the conditionally parametric model is a parametric function of p_2. But unlike the semiparametric model, all coefficients, and not just the intercept, are allowed to vary as functions of x_1.

The conditionally parametric model is particularly simple to fit: one simply drops the parametric predictor p_2 from the distance function used in computing the neighborhood weights. That is,

$$w_i(x) = W\left(\frac{p_{1,i} - p_1}{h}\right).$$

As p_2 changes with p_1 fixed, the smoothing weights, and hence the fitted local model, do not change. This ensures the fit is conditionally parametric in p_2.

6.3 Seasonal Components

Sequences of observations frequently exhibit both long term trends and short term seasonal effects. For example, observations may exhibit annual or daily cycles.

To enforce periodicity in the local regression fit, we define the smoothing weights and fitting functions in a periodic manner. First, define the distance function

$$d(x_1, x_2) = 2|\sin((x_1 - x_2)/(2s))| \tag{30}$$

where s is a scale parameter. Note some properties of this distance function. If $x_1 = x_2$, then $d(x_1, x_2) = 0$. If $x_1 - x_2$ is small, then $\sin((x_1 - x_2)/(2s)) \approx (x_1 - x_2)/(2s)$, so $d(x_1, x_2) \approx |x_1 - x_2|/s$. If $x_1 - x_2 = \pi s$, then $d(x_1, x_2) = 2$; the maximum value If $x_1 - x_2 = 2\pi s$, then $d(x_1, x_2) = 0$. Thus the distance function is periodic in nature, with period $2\pi s$. The smoothing weights are defined as

$$w_i(x) = W\left(\frac{d(x_i, x)}{h}\right). \tag{31}$$

Then, in place of the local polynomial model, we fit the circular model

$$\mu_x(x_i) = a_0 + a_1 s \sin((x_i - x)/s) + a_2 s^2(1 - \cos((x_i - x)/s)) \tag{32}$$

using the weights $w_i(x)$, to get local parameter estimates $\hat{a}_0, \hat{a}_1, \hat{a}_2$. The smooth estimate is $\hat{\mu}(x) = \hat{a}_0$.

Fitting the circular model defined by the weights (31) and fitting functions (32) produces a purely periodic model. A more general model includes both a long term trend $\hat{\mu}_0(x)$ and a periodic component $\hat{\mu}_1(x)$:

$$\hat{\mu}(x) = \hat{\mu}_0(x) + \hat{\mu}_1(x)$$

One estimates the long term trend $\hat{\mu}_0(x)$ using a local polynomial smoother with a large bandwidth, and the periodic component with a small bandwidth periodic smoother.

An even more general procedure is to fit as a bivariate model. For example, suppose one wants to model a sequence exhibiting annual effects. If local regression is applied directly, the local neighborhood for a date x would consist only of nearby dates in the same year. But in the bivariate model, the local neighborhood defines closeness to take into account nearby days over a range of years.

6.4 Correlated Errors

7 Understanding Model Selection

As we have seen earlier, many model selection issues arise in local fitting: variable selection, smoothing parameter selection, choice of the local polynomial degree or local model. We have introduced a number of criteria, such as cross validation, that can help guide the choice of these components.

Successful use of these criteria requires a careful understanding of the model selection problems, and in particular what makes the problems difficult. An ideal goal might be automatic bandwidth and model selection: an algorithm that takes the data as input, and magically produces the best local polynomial fit as output. Unfortunately (and despite numerous claims in statistical journals) this goal is unattainable, since the data is often inconclusive as to what the best fit is.

In Figure 2, we showed four local quadratic fits to the ethanol dataset, each computed with a different smoothing parameter. As discussed in Section 3, the largest smoothing parameter, $\alpha = 0.8$, fairly clearly oversmooths. Choosing among the remaining fits is indecisive: while changing the smoothing parameter from $\alpha = 0.6$ to $\alpha = 0.2$ results in fits showing different structure in the data, making a definitive statement as to which fit is best is quite impossible. That is, there is a wide uncertainty in the data.

Figure 3 shows another example using a simulated dataset. Here, we have fitted two local quadratic regressions. On the left, the bandwidth has been selected using the CP method, producing $h = 0.0618$ (cross validation produced very similar results). On the right, $h = 0.2878$ is selected by a plug-in approach, described below.

At first glance, the plug-in approach appears to produce a much better fit; the CP approach looks substantially undersmoothed. A closer look may leave us less certain. Between 0.6 and 0.7, several observations provide some suggestion of a peak. Whether these are sufficient to indicate a significant feature is unclear from the data alone. The point to be made here is that any model selection critria should reflect this uncertainty: we should be told

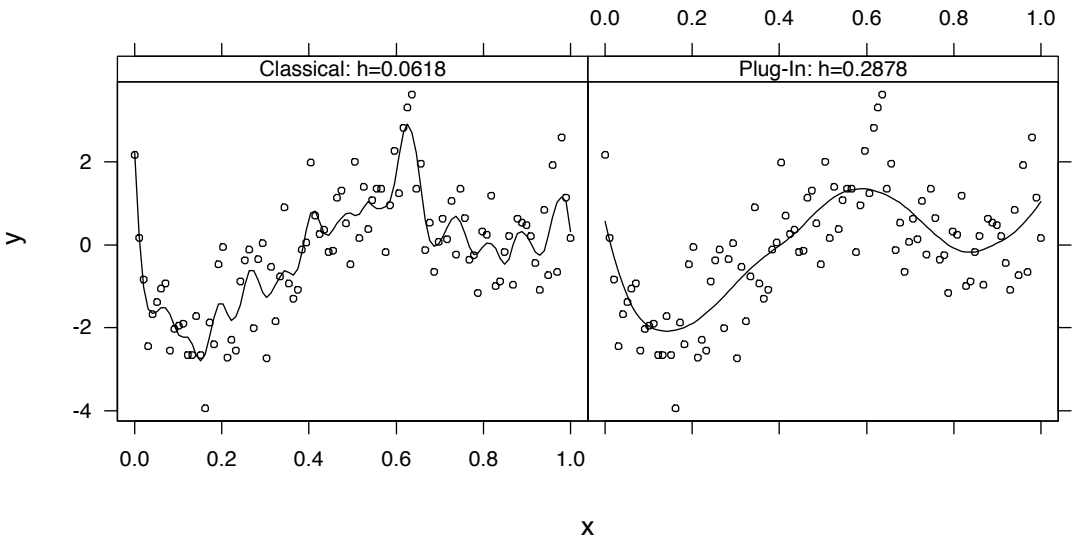

Figure 3: Two local quadratic fits to a simulated dataset. On the left, CP selected $h = 0.0618$. On the right, a plug-in approach selected $h = 0.2878$.

that there is little to choose between models that perform similarly. On the other hand, model selection criteria should firmly reject models that clearly do not fit the data.

This is precisely the behavior indicated in the GCV plot in Figure 1. The largest bandwidths (smallest degrees of freedom) $\alpha = 0.8$ does not fit the data (Figure 2), and this fit should be rejected. The comparison between the other three fits in Figure 2 is less clear, and the GCV criterion reflects this. The GCV scores are $0.184, 0.123, 0.109$ and 0.111 for $\alpha = 0.8, 0.6, 0.4$ and 0.2 respectively, which is almost flat for the smaller smoothing parameters. This *reflects the uncertainty in the data*: inadequate fits are rejected, while those we aren't sure about produce similar results.

7.1 Plug-in Bandwidth Selection

Over the past decade, a huge literature has been devoted to the problem of bandwidth selection, particularly searching for selectors that are less variable than classical approaches such as cross validation and CP. Although originally developed for kernel density estimation (Woodroofe 1970, Park and Marron 1990, Sheather and Jones 1991, Jones, Marron, and Sheather 1996), there have been a number of recent extensions to the regression set-

ting (Gasser, Kneip, and Köhler 1991, Ruppert, Sheather, and Wand 1995).

The beginning point for plug-in bandwidth selection is the mean integrated squared error (MISE):

$$\text{MISE}(\hat{\mu}, \mu) = \int E((\hat{\mu}(x) - \mu(x))^2) dx.$$

This can be decomposed into bias and variance components:

$$\text{MISE}(\hat{\mu}, \mu) = \int b(x)^2 dx + \int v(x) dv$$

where $v(x) = \text{var}(\hat{\mu}(x))$ and $b(x) = E\hat{\mu}(x) - \mu(x)$. Expressions for the mean and variance are given by (2) and (3); in the plug-in literature, it is usual, (although not strictly necessary) to use asymptotic approximations. We develop the methods here for local linear estimates; similar expressions, but involving higher order derivatives of $\mu(x)$, can be derived for higher order estimates.

For a local linear estimate, the simplest asymptotic approximation to the bias is

$$\begin{aligned} E\hat{\mu}(x) - \mu(x) &\approx \frac{\mu''(x)}{2} \sum_{i=1}^{n} (x_i - x)^2 l_i(x) && (33) \\ &\approx \frac{h^2}{2} \mu''(x) \int v^2 W(v) dv. && (34) \end{aligned}$$

Similarly, the variance is approximately

$$\begin{aligned} \text{var}(\hat{\mu}(x)) &= \sigma^2 \sum_{i=1}^{n} l_i(x)^2 \\ &\approx \frac{\sigma^2}{nhf(x)} \int W(v)^2 dv. && (35) \end{aligned}$$

Here, $f(x)$ is the design density; we assume the x_i are an independent sample from $f(x)$. With these approximations,

$$\text{MISE}(\hat{\mu}, \mu) \approx \frac{h^4}{4} \left(\int v^2 W(v) dv \right)^2 \int \mu''(x)^2 dx + \frac{\sigma^2}{nh} \int W(v)^2 dv \int f(x)^{-1} dx. \quad (36)$$

The asymptotic MISE (36) can easily be minimized over h, to yield the 'asymptotically optimal' bandwidth

$$h^5_{\text{opt}} = \frac{\sigma^2}{n} \frac{\int W(v)^2 dv \int f(x)^{-1} dx}{(\int v^2 W(v) dv)^2 \int \mu''(x)^2 dx}.$$

This depends on three unknowns: σ^2; the design density $f(x)$; and the second derivative $\mu''(x)$. The first two can be estimated relatively simply.

The idea of plug-in selection is to substitute a 'pilot' estimate for the second derivative $\mu''(x)$. But this is extremely problematic: the pilot estimate itself involves a bandwidth, and the whole bandwidth selection problem is assumed away the bandwidth problem. While second derivative estimates have been proposed, these all involve bandwidths themselves. In fact, one can adapt Henderson's theorem to derivative estimation: Essentially any second derivative estimate amounts to fitting a local quadratic estimate, and taking the curvature term. If the pilot estimate smooths out real features in the data, then so will the resulting estimate.

Several solutions to the problem of choosing a pilot bandwidth have been proposed. One approach, adopted in Sheather and Jones (1991) and Gasser, Kneip, and Köhler (1991), among others, is an assumed relation. For example, Gasser, Kneip, and Köhler (1991) assume the selected bandwidth h should be $n^{-1/10}$ times the pilot bandwidth k. They select a pilot bandwidth for which the plug-in method actually satisfies this relation.

Alternatively, Ruppert, Sheather, and Wand (1995) use the CP method to select a pilot estimate, based on a crude 'blocked' form of local regression. With this type of algorithm, it then becomes important to identify the contributions of both the CP and plug-in steps.

8 Regression Examples

We take $\mu(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4 + c\exp(-1000(x-0.62)^2)$. Figure 4 displays the results of 1000 simulations for $c = 0$ and $c = 3$. We fit local quadratic models; the pilot estimates for the plug-in selectors are local quartic. At $c = 0$, this should be favorable to plug-in selectors, since the pilot local quartic estimate has no bias, and large pilot bandwidths can be used. This is reflected in Figure 4, where GKK and RSW are substantially less variable.

Are CP and GCV too variable? In Figure 4 with $c = 0$, the bandwidths selected range from 0.05 (the programmed lower bound) to over 0.4. To further understand this variability, we take a closer look at one of the 'worst' samples generated in the simulations of Figure 4. The dataset is shown in Figure 3; the bandwidth selectors selected $h = 0.0618$ (CP); $h = 0.0616$ (GCV) and $h = 0.288$ (GKK). The RSW method selected one block at the CP step, leading to the bandwidth $h = 0.353$.

If one looks just at the selected bandwidth, the GKK method, and RSW

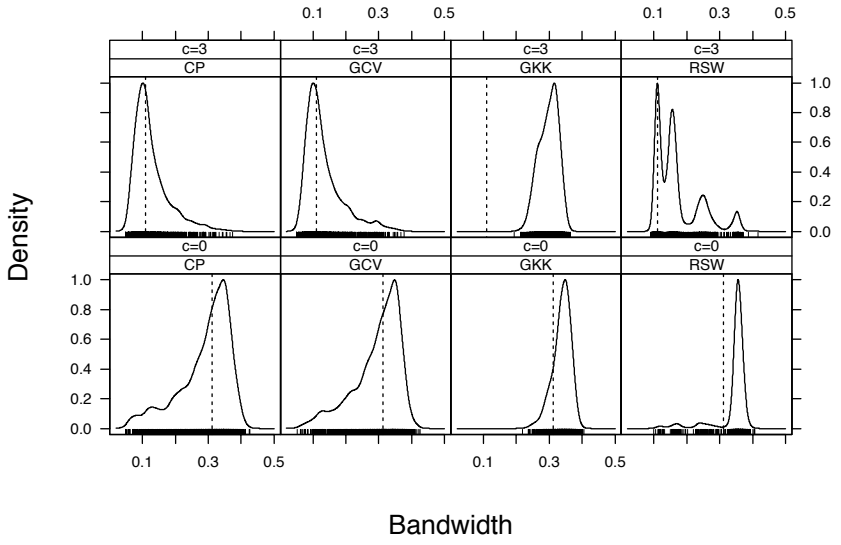

Figure 4: Selected bandwidths for local quadratic regression.

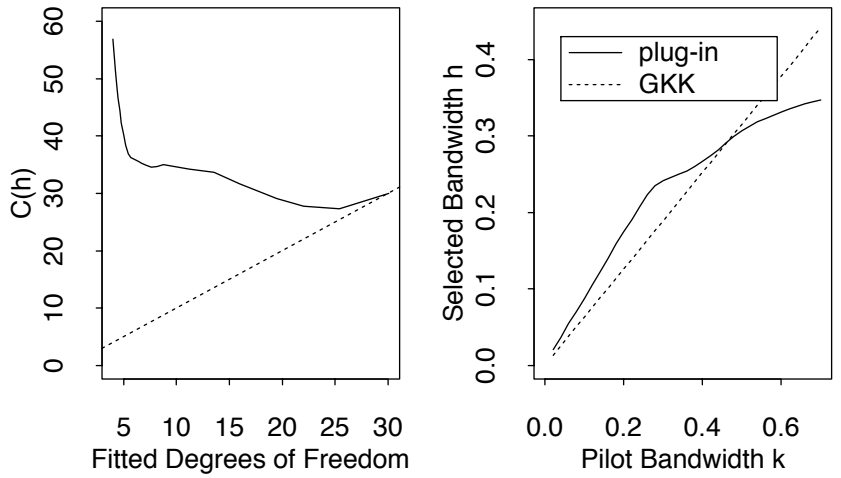

Figure 5: Informative bandwidth assessment for the 'bad' dataset. The flatness of the $C(h)$ plot (left) reflects the uncertainty in the data. The GKK method selects the larger bandwidth (right) without any suggestion of the uncertainty.

(by virtue of its initial C_p step) have got this dataset right, and $C(h)$) and GCV have got it wrong. But the plug-in methods (particularly GKK) pay a price for this: oversmoothing and missing the structure on the more difficult problem in Figure 4, when $c = 3$. But looking at the whole criteria, rather than just the selected bandwidth, produces a much more valuable assessment. The $C(h)$ plot in the left panel of Figure 5 correctly reflects the uncertainty in the dataset, with two local minima and a nearly flat plot from 5 to 30 degrees of freedom. GCV (not shown) produces a similar flat plot. GKK selects the larger bandwidth, with no hint of uncertainty at the smaller bandwidth. The result is catastrophic failure when the bump is real, as demonstrated in Figure 4.

The conclusion here is simple. Variability of CP and GCV is not the problem, but a symptom of how difficult purely data-based bandwidth selection is. It is easy to 'fix' the variability of $C(h)$ to give better results on the dataset in Figure 3; for example, by taking the left-most local minimum rather than the global minimum. But this type of fix fails to address the difficulty of bandwidth selection, and will lead to failure in difficult problems, similar to GKK in Figure 5.

8.1 Asymptotic Results

Some of the strongest arguments in favor of plug-in bandwidth selectors have been based on asymptotic studies. In particular, the rates of convergence of cross validation and similar selectors is $O_p(n^{-1/10})$, while plug-in selectors achieve much faster rates: The algorithm of Sheather and Jones (1991) achieves a rate $O_p(n^{-5/14})$, and other plug-in algorithms achieve the rate $O_p(n^{-1/2})$ (Hall, Sheather, Jones, and Marron 1991). The rates here refer to the rate of convergence to 0 or

$$\frac{\hat{h} - h_0}{h_0}$$

where $\hat{h}$ is the selected bandwidth, and $h_0 = h_0(n)$ is the true minmizer of the MISE.

At a glance, these results appear to provide compelling evidence that plug-in selectors must be better, at least asymptotically. But in light the preceding discussion and examples, the relevance of asymptotic arguments is already questionable. The rates are based on asymptotic expansions of the MISE around h_0; this says nothing about the ability of the methods to distinguish between competing models or identify real features from random noise.

However, the problems with the asymptotic results for bandwidth selection are much deeper. In particular, the plug-in results make stronger assumptions about the smoothness of the underlying mean function, which make the resulting estimate asymptotically inefficient, irrespective of the selected bandwidth. More particularly, the higher order estimate implicit at the pilot stage of the plug-in method beats the resulting estimate. Thus, one is better off omitting the plug-in steps entirely, and using the pilot estimate as the final estimate!

These points are discussed further in Loader (1995). Brown, Low, and Zhao (1997) and Cleveland and Loader (1996) explicitly show how plug-in based estimates can be beaten: simply choose a higher order estimate, and match bandwidths to equate (or reduce) the variance or fitted degrees of freedom. With this bandwidth matching, the higher order fit has, asymptotically, no bias, so must beat the kernel estimate.

An important related paper is Gu (1998), who studies the statistical relevance of a number of issues related to smoothing parameter selection. In particular, he questions the validity of bandwidth asymptotics and a number of other commonly used measures. He also has several examples relating datasets to selected smoothing parameters.

9 Computational Methods

Computational methods have always been necessary to make local regression feasible in practice. And there has always had to be a compromise between the ideal method from the point of view of statistical theory and methods of sufficiently efficient computation to make methods feasible in practice. From the early 19th century until computers became widespread in the 1960s, local regression applications were largely confined to one predictor with equally spaced values; even at that, computational methods were necessary. The solution was summation formulas. When computers came along, this special case became very easy, allowing ideal methods to replace those previously used in practice. But as often happens with computational advances, the goals became far more ambitious. It now was in the domain of feasibility to bring local regression to the general regression setting: one or more predictors and predictor values that were not necessarily equally spaced. New computational methods were needed and in this case the solution has been sparse localization.

We will review the work both in summation formulas and sparse localization. The former is not relevant as a computational method for practice

today, but it is extremely instructive in another way. There is a beautiful interplay between theory and practice; theorists were connected to practice, and the consequence was theoretical results that contributed today. The work in this field had not fallen prey to mathematicism in the sense of (**?**). As we will see in the next section, in the 1960s when local regression moved into general regression studies an asymptotic theory evolved that was not closely linked to practice, resulting the inevitable reduced relevance to practice.

9.1 Summation Formulae

In the 19th century, new local regression methods were developed by researchers applying them to actuarial data: sickness and mortality rates as a function of equally spaced ages (Cleveland and Loader 1996; Loader 1998). This is the domain within which summation formulas were developed. Since the predictor has equally spaced values, we will take $x_i = i$. Smoothers were linear in the y_i. The fitted values were almost always of the form

$$\hat{y}_k = \sum_{j=-a}^{a} c_j y_{k+j}$$

for $k = a+1 \ldots n-a$ where the c_j are symmetric. We will refer to the c_j as the *smoothing weights* and adopt the convention that $c_j = 0$ for $|j| > a$. End effects were addressed by supposing that a smoother did not need to go to the ends of the data, or that if it did, it was possible to extend the data forward and backward. Thus end effects were not directly addressed in most accounts, and we will not do so either in this section.

Through time, in the 19th century, tacit standards evolved for the performance of a smoother. First, effects needed to be accurately tracked, that is, a smoother needed to have acceptable bias properties. The actuaries couched this issue in terms of functions that were reproduced. In most work, the reproduction standard was cubic polynomials. The standard measure of smoothness was the expected sum of squares of the third differences of the fitted values, or, equivalently, the sum of squares of the third differences of the smoothing weights. As we will see, an optimization goal evolved — find a smoother that minimizes the sum of squares of the third differences subject to reproduction of cubics and subject to the constraint of fast computation.

One early method, two successive moving averages of length 5, has been attributed to Finlaison in 1929. Finlaison's method was simple to compute; equal-weight moving averages can be easily computed by updating. But

Finlaison's method did not perform well; it could not reproduce polynomial behavior greater than degree zero. Woolhouse (1870) published a method that had good performance; it yielded a reasonably smooth result and reproduced cubics. The fit at $x_i = i$ is the mean of five fitted polynomials evaluated at i. The rth polynomial, for $r = -2, -1, 0, 1, 2$ is the least-squares fit of quadratic at five positions: $i - r + 5k$ for $k = -2, -1, 0, 1, 2$. At first, the method was unattractive because the computational method that was first developed was tedious to carry out by hand. Thus in the 1870s, there was the computationally expensive method of Woolhouse with good performance and the computationally efficient method of Finlaison with poor performance. This set the stage for the invention of summation formulas, which began appearing in the 1880s.

Most summation formulas can be described by

$$\hat{y}_k = Gy_k$$

where G is a difference operator defined as follows. Let

$$\begin{aligned}
[2m+1]y_k &= \sum_{j=-m}^{m} y_{k-j}, \\
[2m]^+ y_k &= \sum_{j=-m/2+1}^{m/2} y_{k-j}, \\
[2m]^- y_k &= \sum_{j=-m/2}^{m/2-1} y_{k-j}, \\
\gamma_r y_k &= y_{k-r} + y_{k+r}.
\end{aligned}$$

Then for q, r, and s odd,

$$G = \frac{[q][r][s]}{qrs}\{1 + 2(a+b+c) - a\gamma_1 - b\gamma_2 - c\gamma_3\}.$$

The values of q, r, and s have been taken to be odd integers to center the fitted values at integer values of the independent variable $x_k = k$, but this could also be achieved by two even integers, one with a $[\,]^+$ summation and the other with a $[\,]^-$ summation and one odd integer. An algebra of the forward shift operator, $Fy_k = y_{k+1}$, was developed, and using this algebra, it is easy to see (e.g., Henderson 1916) that the above summation formula reproduces cubics provided

$$a + 4b + 9c = \frac{q^2 + r^2 + s^2 - 3}{24}.$$

In the discussion of a paper of (**?**), Hardy pointed out that Woolhouse's method could be expressed as a summation formula,

$$\frac{[5]^3}{125}\{1 - 3\gamma_1\},$$

and thus could also be easily computed.

Spencer (1904) developed a number of summation formulas that became quite popular because they did a good job on the criteria cited above. His 21-point rule is

$$\frac{[5]^2[7]}{175}\{1 + \gamma_1/2 - \gamma_3/2\} = \frac{[5]^2[7]}{350}\{2 + \gamma_1 - \gamma_3\}.$$

The 21 smoothing weights c_k are symmetric and the values of $350c_k$ for $k = -10$ to 0 are

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

The summation formulas can be reduced to a sequence of very simple arithmetic operations amenable even to hand calculation. For example, Spencer's 21-point rule can be carried out by the following where v is an integer chosen to maintain sufficient decimal places of accuracy and to enable the computation to begin with integers:

$$\begin{aligned}
S_1 &= 10^v y_k \\
S_2 &= \frac{1}{7}S_1 \\
S_3 &= [3]S_2 \\
S_4 &= \gamma_3 S_2 \\
S_5 &= S_2 + S_3 - S_4 \\
S_6 &= [7]S_5 \\
S_7 &= \frac{1}{5}S_6 \\
S_8 &= [5]S_7 \\
S_9 &= [5]S_8 \\
S_{10} &= \frac{1}{10}S_9
\end{aligned}$$

Each moving sum can be carried out by updating, which requires an addition and a subtraction for each position beyond the first.

The summation formulas reproduced cubics. Furthermore, because of the repeated moving averages, they yielded smooth results. (A single, long, equal-weight moving average would not do so.) But what can be said about the smoothness, and why is it reasonable to place these summation formulas in the category of local regression? Henderson (1916) carried out a beautiful piece of theoretical work, answering both of these questions.

Suppose we are carrying out weighted local cubic fitting of $2m+1$ values. Let w_k for $k = -m, \ldots, m$ be the neighborhood weight given to (k, y_k) for the fit at $x_i = i$. Henderson (1916) proved the theorem described in Section 2; in this case, the theorem implies 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's theorem also implies that if the smoothing weights, c_k, of a cubic-reproducing summation formula have no more than three sign changes, then the formula can be represented as local cubic smoothing with neighborhood weights $w_k > 0$ and a cubic polynomial $\phi(k)$ such that $\phi(k)w_k = c_k$. For example, for Spencer's 21-point rule we can take

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

and take w_k for $k = -10, \ldots, 0$ 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 (1916) 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 smoothing weights, or equivalently, the expected sum of squares of the third differences of the fit. To simplify the formula, suppose the neighborhood for the fit at i extends from $i - (m - 2)$ to $i + (m + 2)$. The solution for the smoothing weights $c_k, k = -(m-2), \ldots, (m-2)$ is

$$\frac{((m-1)^2 - k^2)(m^2 - k^2)((m+1)^2 - k^2)((3m^2 - 16) - 11k^2)}{8m(m^2 - 1)(4m^2 - 1)(4m^2 - 9)(4m^2 - 25)}.$$

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 function $(1-x^2)^3$.

Henderson did not appear to put forward his optimal smoother for use in practice because it would have been too computationally complex. But it did provide a benchmark for judging summation formulas. For example, one can check how close a particular summation formulas comes to the ideal. Suppose the neighborhood covers 15 points so that $m = 9$ in the above formula. Figure 6 shows the smoothing weights for Henderson's optimal smoother, Spencer's 15-point rule, Woolhouse's formula, and local cubic regression with a uniform neighborhood weight function. The sum of squares of the third differences of the smoothing weights multiplied by 1000 are the following: Henderson = 3.58, Spencer = 5.51, Woolhouse = 8.58, Cubic = 128.3. Spencer's formula does quite well compared with the optimal, and so, in practice, was a far more attractive choice given the computational resources in the early 1900s. The use of a uniform weight function results in quite poor results; the discontinuity at $\pm$ 1 results in local roughness in the smoothed values.

9.2 Sparse Localization

The methods we describe next are related to what is known in the numerical analysis and engineering literature as finite element methods. Evaluating a surface at a large number of points may be computationally prohibitive; in the local regression case, this requires a least squares fit at each evaluation point. For local likelihood methods, a non-linear optimization may be required at each point. But if the underlying surface is smooth, evaluation of the surface at a point provides some information about the surface in the neighborhood of the point.

The hierarchy of methods described in this section exploit smoothness of the surface by fitting at a small number of points. Interpolation methods are then used to define a complete surface over the domain. Since the interpolant at any point will be defined in terms of the fit at a small number of evaluation points, this can result in substantial savings over direct fits.

The computational methods described here can be viewed as bridging the gap between local fitting methods and spline methods. In the final representation, the fit resulting from the LOESS and LOCFIT methods is very similar to that which results from a spline fit; indeed, the explicit

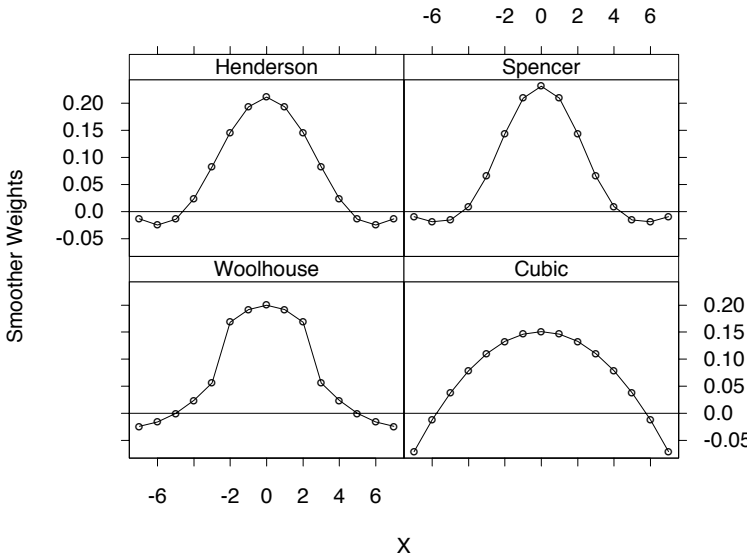

Figure 6: Smoothing weights for Henderson's optimal and three other smoothers.

two stage procedures have also been used in spline software. This type of representation has significant advantages over direct evaluation.

On the other hand, the use of local fitting rather than the global approaches yields significant computational advantages, particularly in the multidimensional setting.

LOWESS

The LOWESS software (Cleveland 1979) provides an early attempt at a sparse localization method. The description is included here for historical completeness; subsequent methods provide more general and more complete computational methods.

LOWESS implements local linear regression with one predictor and computes fitted values at observation points. Rather than compute at each observation, a subset of the observations are chosen and linear interpolation used. The specific algorithm, designed by Paul Tukey, proceeds as follows: First, order the x_i from smallest to largest. The fit at x_1 is computed. Then a step is taken to $x_1 + \delta$. If there are x_i between x_1 and $x_1 + \delta$, the fit is computed at the largest such x_i and the fit at values between x_1 and this largest value are computed by linear interpolation. This stepping out by δ

is carried out repeatedly; each step goes out from the point of the last full computation of local fitting.

LOWESS also incorporates a number of other features:

- Nearest neighbor bandwidths; The bandwidth $h = h(x)$ is varied to cover a fixed fraction α of the observations. This is necessary for nonuniform designs, to avoid sparse regions where the estimate is not well defined. Even for uniform designs the nearest neighbors are usually advantageous, since they help control boundary variability problems.

- Tricube weight function. When fitting at a point x, observations are assigned weights $W((x - x_i)/h)$ with $W(u) = (1 - |u|^3)^3_+$. The use of smooth weights - rather than rectangular weights - produces a smoother curve estimate; a point that has been well appreciated for over a century (Woolhouse ???).

- Robustness iterations.

LOESS

The LOESS software employs substantially new computational methods (Cleveland and Devlin 1988; Cleveland and Grosse 1991; Cleveland, Grosse, and Shyu 1992). This enables many significant enhancements in the applicability compared to LOWESS. For example, LOESS allows multivariate fitting with two or more predictors; local linear and quadratic fitting; conditionally parametric fitting, and computing of quantities for inference.

The specific partition used by LOESS is a k-d tree, first introduced by Friedman, Bentley, and Finkel (1977) for finding nearest neighbors. This is a partition into rectangular cells, with each cell of the final tree containing the same number of points. Thus, the structure adapts to design density in a similar manner to the nearest neighbor bandwidths used by LOESS; the cells are smallest, and hence the most vertices, are in regions where the bandwidth is smallest.

The local polynomial fit is computed at the vertices of the k-d tree, and both the fitted values and local slopes are stored for use at the prediction stage.

The LOESS computations are carried out in two distinct stages. In the first stage, local polynomial fitting of degree one or two is carried out at a small number of points (typically much smaller than the sample size), determined by a k-d tree algorithm. The fitted values and partial derivatives of

the locally fitted polynomials are stored in a data structure. In the second stage, piecewise cubic blending functions are used to construct the estimate across the cells of the k-d tree. That is, it is the second stage that provides the evaluation of the fit at any desired set of points. This approach works because having a fitted value and derivative at a point x of a smooth surface provides substantial information about the fit in a neighborhood of x. When the local fitting at the vertices is linear, the algorithm preserves linear functions; when the local fitting at the vertices is quadratic, the algorithm preserves quadratic functions provided cross derivatives are not set to zero.

For the second, or evaluation, stage of LOESS, it is only the evaluation data structure that is required; neither the original data nor any other information is needed. In other words, the fit is parsimoniously characterized. Evaluation is $O(g)$ where g is the number of evaluation points and is exceedingly fast. Note that is does not depend on n the sample size. Such a structure is suitable for evaluation of the surface at any set of points in the design space — single points (for prediction), a grid (for plotting the surface), the original data (for diagnostics such as residual plots), or any other collection of points. And the two-step computational approach can serve as a basis for local fitting procedures in more complicated settings. For example, Loader (1994) applies the model directly to local likelihood density estimation.

9.3 LOCFIT

The computational model for LOCFIT extends the ideas of LOESS. The predictor space is again divided into cells, with direct computations performed at vertices of the partition. The most significant difference is in how the split rules are determined: rather than using the number of points, an edge requires splitting if the length of the edge is large relative to the bandwidth used at either end of the edge.

Initially, the data is bounded by an interval (cell) $[z_0, z_1]$, and the local fit is computed at these two points. In addition to the local coefficients, the bandwidths h_0 and h_1 actually used at these points are saved. A score is then assigned to the cell:

$$\rho(z_0, z_1) = \frac{|z_0 - z_1|}{\min(h_0, h_1)}. \tag{37}$$

If this score is large, then interpolation across the cell is likely to be unsatisfactory, and refinement is necessary. A new vertex is then added at the midpoint $(z_0+z_1)/2$. Two new cells $[z_0, z_2]$ and $[z_2, z_1]$ are added to the tree.

The local fit is evaluated at z_2; again, the local coefficients and bandwidth are saved. The process is then repeated recursively until none of the leaf cells need further splitting.

At this stage we have said nothing about selection of the bandwidths h_j, and so the computational algorithm described can be used with any selection rule: fixed, nearest neighbor or adaptive methods. Also, the definition of the evaluation structure depends only on the bandwidth and is otherwise completely independent of the fitting procedure. Hence, the computational method can be adapted for use with a wide variety of local fitting procedures. The present LOCFIT implementation supports local least-squares regression; local likelihood with a variety of likelihoods; density estimation and hazard rate estimation.

In multiple dimensions, the computational methods follow the same ideas. One begins by bounding the data in a d dimensional hypercube, and computes the fit, and bandwidths, at each of the 2^d vertices. The cube is then split at the midpoint of the longest edge, generating two new hypercubes. The process is continued recursively, until no edges of the cube require splitting, as determined by the rule (37).

References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics* **16**, 125–127.

Azzalini, A. and A. W. Bowman (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, Series B* **55**, 549–557.

Azzalini, A., A. W. Bowman, and W. Härdle (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Brown, L. D., M. G. Low, and L. H. Zhao (1997). Superefficiency in nonparametric function estimation. *The Annals of Statistics* **25**, 2607–2625.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W. S. (1993). *Visualizing Data.* Summit, New Jersey: Hobart Press.

Cleveland, W. S. (1994). Coplots, nonparametric regression, and conditionally parametric fits. In T. W. Anderson, K. T. Fang, and I. Olkin (Eds.), *Multivariate Analysis and its Applications*, pp. 21–36. Hayward: Institute of Mathematical Statistics.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

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

Cleveland, W. S., E. H. Grosse, and W. M. Shyu (1992). Local regression models. In J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S*, pp. 309–376. Pacific Grove: Wadsworth and Brooks/Cole.

Cleveland, W. S. and C. R. Loader (1996). Smoothing by local regression: Principles and methods. In W. Härdle and M. G. Schimek (Eds.), *Statistical theory and computational aspects of smoothing*, Heidelberg, pp. 10–49. Physica-Verlag.

Cox, D. D., E. Koh, G. Wahba, and B. S. Yandell (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics* **16**, 113–119.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Davies, R. B. (1980). ASS 155: The distribution of a linear combination of chi-squared random variables. *Applied Statistics* **29**, 323–333.

Engle, R., C. Granger, J. Rice, and A. Weiss (1986). Nonparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–320.

Eubank, R. L. and P. L. Speckman (1993). A bias reduction theorem with applications in nonparametric regression. *Scandinavian Journal of Statistics* **18**, 211–222.

Friedman, J. H., J. L. Bentley, and R. A. Finkel (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* **3**, 209–226.

Gasser, T., A. Kneip, and W. Köhler (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* **86**, 643–652.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**, 245–260.

Gu, C. (1998). Model indexing and smoothing parameter selection in nonparametric regression. *Statistica Sinica* **??**, To Appear.

Hall, P., S. J. Sheather, M. C. Jones, and J. S. Marron (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–270.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.

Hastie, T. J. and C. R. Loader (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science* **8**, 120–143.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

Henderson, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America* **17**, 43–48.

Hjellvik, V., Q. Yao, and D. Tjøstheim (1996). Linearity testing using local polynomial approximation. Technical Report UKC/IMS/96/19, Institute of Mathematics and Statistics, University of Kent at Canterbury.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419–426.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401–407.

Katkovnik, V. Y. (1979). Linear and nonlinear methods of nonparametric regression analysis. *Soviet Automatic Control* **5**, 35–46.

Katkovnik, V. Y. (1985). *Nonparametric Identification and Smoothing of Data (in Russian)*. Moscow: Nauka.

Knafl, G., J. Sacks, and D. Ylvisaker (1985). Confidence bands for regression functions. *Journal of the American Statistical Association* **80**, 683–691.

Leadbetter, M. R., G. Lindgren, and H. Rootz'en (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag.

Loader, C. R. (1994). Computing nonparametric function estimates. In *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface*, pp. 356–361.

Loader, C. R. (1995). Old Faithful erupts: Bandwidth selection reviewed. Technical report, Bell Laboratories, Murray Hill, NJ.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics* **24**, 1602–1618.

Loader, C. R. (1998). *Local Regression and Likelihood*. New York: Springer-Verlag.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *The Annals of Statistics* **27**, To Appear.

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics* **25**, 186–211.

Park, B. U. and J. S. Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* **85**, 66–72.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**, 1065–1076.

Raz, J. (1990). Testing for no effect when estimating a smooth regression function by nonparametric regression. *Journal of the American Statistical Association* **85**, 132–138.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics* **12**, 1215–1230.

Rice, S. O. (1939). The distribution of the maxmima of a random curve. *AMJM* **61**, 409–416.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* **27**, 832–837.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110–114.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: John Wiley & Sons.

Severini, T. A. and J. Staniswalis (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89**, 501–511.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* **53**, 683–690.

Siegmund, D. O. and K. J. Worsley (1995). Testing for a signal with unknown location and scale in a stationary gaussian random field. *The Annals of Statistics* **23**, 608–639.

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

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics* **19**, 415–428.

Spencer, J. (1904). On the graduation of rates of sickness and mortality. *Journal of the Institute of Actuaries* **38**, 334–343.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statistics* **5**, 595–645.

Stone, M. (1974). Cross-validating choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 111–47.

Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *The Annals of Probability* **21**, 34–71.

Sun, J. and C. R. Loader (1994). Simultaneous confidence bands in linear regression and smoothing. *The Annals of Statistics* **22**, 1328–1345.

Tibshirani, R. J. and T. J. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–567.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmition* **22**, 69–84.

Wang, F. T. and D. W. Scott (1994). The L_1 method for robust nonparametric regression. *Journal of the American Statistical Association* **89**, 65–76.

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

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society, Series B* **20**, 334–343.

Woodroofe, M. (1970). On choosing a delta sequence. *The Annals of Mathematical Statistics* **41**, 1665–1671.

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. *Journal of the Institute of Actuaries* **15**, 389–410.