Literature Review for Local Polynomial Regression

Matthew Avery

Abstract

This paper discusses key results from the literature in the field of local polynomial regression. Local polynomial regression (LPR) is a nonparametric technique for smoothing scatter plots and modeling functions. For each point, x_0, a low-order polynomial WLS regression is fit using only points in some "neighborhood" of x_0. The result is a smooth function over the support of the data. LPR has good performance on the boundary and is superior to all other linear smoothers in a minimax sense. The quality of the estimated function is dependent on the choice of weighting function, K, the size the neighborhood, h, and the order of polynomial fit, p. We discuss each of these choices, paying particular attention to bandwidth selection. When choosing h, "plug-in" methods tend to outperform cross-validation methods, but computational considerations make the latter a desirable choice. Variable bandwidths are more flexible than global ones, but both can have good asymptotic and finite-sample properties. Odd-order polynomial fits are superior to even fits asymptotically, and an adaptive order method that is robust to bandwidth is discussed. While the Epanechnikov kernel is superior is an asymptotic minimax sense, a variety are used in practice. Extensions to various types of data and other applications of LPR are also discussed.

1 Introduction

1.1 Alternative Methods

Parametric regression finds the set of parameters that fits the data the best for a predetermined family of functions. In many cases, this method yields easily interpretable models that do a good job of explaining the variation in the data. However, the chosen family of functions can be overly-restrictive for some types of data. Fan and Gijbels (1996) present examples in which even a 4th-order polynomial fails to give visually satisfying fits. Higher order fits may be attempted, but this leads to numerical instability. An alternative method is desirable.

One early method for overcoming these problems was the Nadaraya-Watson estimator, pro-

posed independently and simultaneously by Nadaraya (1964) and Watson (1964). To find an estimate for some function, $m(x)$, we take a simple weighted average, where the weighting function is typically a symmetric probability density and is referred to as a kernel function. Gasser and Müller (1984) proposed a similar estimator:

$$\widehat{m}_h(x) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(u-x)du \tag{1}$$

where $s_i = (X_i + X_{i+1})/2$, $s_0 = -\infty$, and $s_{n+1} = \infty$. This estimator is able to pick up local features of the data because only points within a neighborhood of x are given positive weight by K_h. However, the fit is constant over each interval, (s_i, s_{i+1}), and a constant approximation may be insufficient to accurately represent the data. A more dynamic modeling framework is desired.

1.2 Local Polynomial Regression (LPR)

In local polynomial regression, a low-order weighted least squares (WLS) regression is fit at each point of interest, x using data from some neighborhood around x. Following the notation from Fan and Gijbels (1996), let the (X_i, Y_i) be pairs of data points such that

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \tag{2}$$

where $\epsilon_i \sim N(0,1)$, $\sigma^2(X_i)$ is the variance of Y_i at the point X_i and X_i comes from some distribution, f. In some cases, homoskedastic variance is assumed, so we let $\sigma^2(X) = \sigma^2$. It is typically of interest to estimate $m(x)$. Using Taylor's Expansion:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \ldots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \tag{3}$$

We can estimate these terms using weighted least squares. Minimze:

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{p}\beta_j(X_i - x_0)^j\right]^2 K_h(X_i - x_0). \tag{4}$$

In (4), h controls the size of the neighborhood around x_0, and $K_h(\cdot)$ controls the weights, where $K_h(\cdot) \equiv \frac{K(\frac{\cdot}{h})}{h}$, and K is a kernel function. Denote the solution to (4) as $\widehat{\boldsymbol{\beta}}$. Then $\widehat{m}^{(r)}(x_0) = r!\widehat{\beta}_r$. It is often simpler to write the weighted least squares problem in matrix notation. Therefore, let $\boldsymbol{X}$ be the design matrix centered at x_0:

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 - x_0 & \dots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n - x_0 & \dots & (X_n - x_0)^p \end{pmatrix}. \tag{5}$$

Let W be a diagonal matrix of weights such that $W_{j,j} = [K_h(X_i - x_0)]$. Then the minimization problem:

$$\underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{XB}) \tag{6}$$

is equivalent to (4), and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{WX})^{-1}\boldsymbol{X}^T\boldsymbol{Wy}$. (Fan and Gijbels, 1996) We can also use this notation to express the conditional mean and variance of $\widehat{\boldsymbol{\beta}}$:

$$E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{WX})^{-1}\boldsymbol{X}^T\boldsymbol{Ws} \tag{7}$$

$$Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{WX})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma X})(\boldsymbol{X}^T\boldsymbol{WX})^{-1}, \tag{8}$$

where $\boldsymbol{s} = (m(X_1), \dots, m(X_2)) - \boldsymbol{X\beta}$ and $\boldsymbol{\Sigma} = diag\{K_h^2(X_i - x_0)\sigma^2(X_i)\}$. There are three critical parameters whose choice can effect on the quality of the fit. These are the bandwidth, h, the order of the local polynomial being fit, p, and the kernel or weight function, K (often denoted K_h to emphasize its dependence on the bandwidth). While we focus mainly on estimation of $m(x)$,

many of these results can be used for estimating the rth derivative of $m(x)$ with slight modification. The remainder of this section discusses early work on the subject of LPR, and Section 2 covers some general properties. Section 3 discusses the choice of bandwidth, Section 4 covers the choice of order and the kernel function, Section 5 discusses options for fast computation, and Section 6 details some extensions.

1.3 Early results for local polynomial regression

Stone (1977) introduced a class of weight functions used for estimating the conditional probability of a response variable, Y given a corresponding value for X. Particularly, Stone suggests a weight function that assigns positive values to only the k observations with X-values closest to the point of interest, x_0, where "closest" is determined using some pseudo-metric, p, which is subject to regularity conditions. A "k nearest neighbor" (kNN) weight function is defined as follows. For each x_0, let $W_i(x)$ be a function such that $W_i(x) > 0$ if and only if $i \in I_k$, where I_k is an index set defined such that $i \in I_l$ if and only if fewer than k of the points $X_1, X_2, \ldots, X_n$ are closer to x_0 than X_i using the metric p. Otherwise, let $W_i(x) = 0$. Then $W_i(x)$ is a kNN weight function. Moreover, the sequence of kNN weight functions, W_m is consistent if $k_m \to \infty$ and $k_m/m \to 0$ as $m \to \infty$. Stone uses a consistent weight function to estimate the conditional expectation of Y using a local linear regression. The proposed equation is equivalent to the linear case of (4).

Cleveland (1979) expanded upon this idea, suggesting an algorithm to obtain an estimated curve that is robust to outliers. As in Stone (1977), we fit a p-degree local polynomial for each Y_i using weights $w_j(X_i)$ and note the estimate, $\widehat{Y}_i$. To get robust estimates, we find new weights according the size of the estimated residuals, $e_i = Y_i - \widehat{Y}_i$, and letting $\delta_j = B(e_j/6s)$, where s is a scaling factor equal to the median of the e_i's, and $B(\cdot)$ is a weight function. (Cleveland suggests using a bisquare weight function, see Section 4.2.) Finally, we compute the robust estimators by fitting the weighted polynomial regression model for each point X_i using $\delta_j w_j(X_i)$ as the new weights. The combined weights in this estimator ensure that "near-by" points remain strongly

weighted, but points with high associated first-stage residuals have less influence over the final fit. This keeps estimates near "outlier" points from being highly biased while still ensuring a smooth fit that picks up local features of the data.

An early attempt at describing the distributional properties of the local polynomial regression estimator is given in Cleveland (1988). Building on the methodology described above in Cleveland (1979), they note that the estimated mean function, $\widehat{m}(x_0)$, can be written as a linear combination of the Y_i:

$$\widehat{m}(x_0) = \sum_{i=1}^{n} l_i(x_0) Y_i. \tag{9}$$

Since we are assuming that the ϵ_i are normally distributed, it is clear that $\widehat{m}(x_0)$ also has a normal distribution with associated variance $\widehat{\sigma}^2(x_0) = \sigma^2 \sum_{i=1}^{n} l_i^2(x_0)$. These results are similar to what we would have for standard polynomial regression and suggest that results from the standard case may hold for LPR. Some relevant examples are given in Cleveland (1988).

2 Properties of Local Polynomial Regression estimators

2.1 Conditional MSE

Fan and Gijbels (1992) establish some asymptotic properties for the estimator described in (4). In particular, they give an expression for the conditional bias and conditional variance of the estimator for $\widehat{m}(x)$ found by minimizing:

$$\sum_{j=1}^{n} \left(Y_n - \beta_0 - \beta_1(x - X_j)\right)^2 \alpha(X_j) K\left(\frac{x - X_j}{h_n}\alpha(X_j)\right). \tag{10}$$

Note that the linear ($p = 1$) case of (4) is a equivalent to (10) when $\alpha(X_j) = 1$. The conditional bias and variance are important because they allow us to look at the conditional MSE, which is

important for choosing an optimal bandwidth. (See Section 3

The results from Fan and Gijbels (1992) are limited to the case where the X_i's are univariate. Ruppert and Wand (1994) give results for multivariate data, proposing the following model:

$$Y_i = m(\boldsymbol{X_i}) + \sigma(\boldsymbol{X_i})\epsilon_i, \quad i = 1, \ldots, n. \tag{11}$$

where $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, ϵ_i are iid with mean 0 and variance 1, and $\sigma^2(\boldsymbol{x}) = Var(Y|\boldsymbol{X} = \boldsymbol{x}) < \infty$. A solution to the problem comes from slightly modifying (6). Consider the case of local linear regression ($p = 1$). We now let

$$\boldsymbol{X} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x}_0) & \ldots & (\boldsymbol{X}_1 - \boldsymbol{x}_0)^T \\ \vdots & \vdots & & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x}_0) & \ldots & (\boldsymbol{X}_n - \boldsymbol{x}_0)^T \end{pmatrix} \tag{12}$$

and denote $W = diag\{K_H(\boldsymbol{x}_1 - \boldsymbol{x}_0), \ldots, K_H(\boldsymbol{x}_n, \boldsymbol{x}_0)\}$, where K is a d-dimensional kernel and $K_H(u) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}u)$, where $\boldsymbol{H}^{1/2}$ is the bandwidth matrix, analogous to h for the univariate case. Often $\boldsymbol{H}$ will be given a simple diagonal form, and then $\boldsymbol{H} = diag(h_1^2, \ldots, h_d^2)$.

Using similar assumptions to the univariate case, we can give expression for the conditional bias and conditional variance of $\widehat{m}_H(\boldsymbol{x})$. We work with the conditional bias and variance (given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) because, by conditioning on the data, the moments of $\widehat{m}_H(\boldsymbol{x})$ exist with probability tending to 1. The asymptotic properties of this estimator will depend on whether we are looking at an interior point or a point near the boundary. For some interior point x_0, we have the following:

$$E(\widehat{m}_H(\boldsymbol{x}_0) - m_H(\boldsymbol{x}_0)|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{2}\mu_2(K)tr(\boldsymbol{H}\boldsymbol{H}_m(\boldsymbol{x})) + o_P(tr(\boldsymbol{H})) \tag{13}$$

$$Var(\widehat{m}_H(\boldsymbol{x}_0)|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \{n^{-1}|\boldsymbol{H}|^{-1/2}R(K)/f(\boldsymbol{x})\}\sigma^2(\boldsymbol{x})(1 + o_P(1)). \tag{14}$$

In the above, $\mathcal{H}_m$ denotes the Hessian matrix ($d \times d$-dimensional) of m and $R(K)$ is the square integral of $K(u)$. In the lead term of the bias, note that $\boldsymbol{H}\mathcal{H}_m(\boldsymbol{x}_0)$ is the sum over each direction of the product of the bandwidth times the curvature of m at $\boldsymbol{x}_0$. If there is very high curvature, an estimator with a large bandwidth will struggle to approximate it accurately, which leads to a high bias as (13) would suggest. The first part of the expression in (14) can be thought of as the inverse of the effective sample size used for the fit. So as we would expect, the variance increase as the effective sample size decreases. The relationship in these two expressions is similar to what we see in the univariate case: the larger the neighborhood, the larger the bias. Conversely, when the neighborhood becomes smaller, the variance will be large. (This bias/variance tradeoff is discussed in Section 3.1.)

2.2 Minimax Efficiency

Fan (1993) showed that the local linear model using the Epanechnikov kernel optimizes the linear minimax risk. Minimax risk is a criterion used to benchmark the efficiency of an estimator in terms of the sample size necessary to obtain a certain quality of results. For example, if an estimator $\widehat{m}(x)$ is 95% efficient when compared to the "optimal" estimator, $\widehat{m}_{opt}(x)$, then an estimate based on $n = 100$ data points using $\widehat{m}(x)$ will have similar asymptotic properties to an estimate based on 95 observations using $\widehat{m}_{opt}(x)$. Fan et al. (1997) extended this result to LPR with order p as well as the case of derivative estimation. So local polynomial regression is the best linear smoother in this minimax sense for interior points.

2.3 Performance at the boundary

One advantages of LPR over other smoothers is its relatively good performance near the boundary. For many nonparametric smoothers, estimates of points near the boundary of the support of the data behave differently from those on the interior. Let f be the marginal distribution function for the $\boldsymbol{X}_i$. Denote the support of f by $supp(f)$. We say $\boldsymbol{x}$ is an interior point if

$$\{z : \boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{z}) \in supp(K)\} \subset supp(f), \tag{15}$$

where $supp(K)$ is the support of $K_H(\boldsymbol{x} - \bullet)$. So $\boldsymbol{x}$ is an interior point if the neighborhood around $\boldsymbol{x}$ as defined by $\boldsymbol{H}$ does includes points outside the support of f.

Fan and Gijbels (1992) note that previous estimators, such as the Nadaraya-Watson and Gasser-Müller estimators described in Section 1.1 converge more slowly at the boundary. However, they show that the convergence rate of the estimator they propose (see Section 3 is the same for boundary points and interior points. Ruppert and Wand (1994) note a similar results for the multivariate case. However, in both situations, the conditional variance is larger in practice for points on the boundary than for points on the interior. Fan and Gijbels attributed this to the lower number of data points being used for estimations near the boundary, but Ruppert and Wand also note that the estimates for the intercept and slope parameters are not asymptotically orthogonal as they are for interior point estimations. Finally, Cheng et al. (1997) show that no linear estimator can beat LPR on the boundary in a minimax sense in terms of MSE. Rather than by showing directly that other proposed boundary corrections are inferior, they show that the local polynomial estimator is optimal in this minimax sense, and therefore any other estimator cannot give a substantial improvement in efficiency on. So LPR is minimax efficient for both interior and boundary points. For futher discussion, also see Hastie and Loader (1993)

3 Bandwidth selection

3.1 The bias-variance tradeoff

The choice of bandwidth, h, is of critical importance for local polynomial regression. The bandwidth controls the complexity or how "jagged" the fit is. Smaller values for h will result in less smoothing while larger values produce a curve with fewer sharp changes. Additionally, there is a tradeoff between variance and bias. Larger values for h will reduce the variance, since more

points will be included in the estimate. However, as h increase, the average distance between these "local" points and x_0 will also increase. This can result in a larger bias. A natural way to choose a bandwidth and balance this tradeoff is by minimizing the mean squared error (MSE). (Fan and Gijbels, 1996) In local regression settings, we must also choose whether to find a bandwidth, h, that is optimal for the full range of our data (a global bandwidth) or choose an h_x that is optimal at each point but varies depending on x. This latter choice is referred to as a variable bandwidth. We focus on global bandwidths first.

3.2 Global bandwidth selection

Integrating the conditional MSE over the parameter space gives an expression for Mean Integrated Squared Error (MISE). Minimizing MISE is a common method for choosing an optimal bandwidth. (Ruppert et al., 1995, Xia and Li, 2002, Fan and Gijbels, 1992) Estimating the minimizer, h_{opt}, can either be done empirically using cross-validation (CV) techniques or asymptotically using expressions for the asymptotic bias and variance as described in 2.1. These results give us an expression for the conditional MSE, but this expression includes unknown terms (particularly, $m''(x)$, $\sigma(x)$, and $f(x)$) that we must estimate. There are many approaches for finding estimates for these unknowns, varying from simple "rules of thumb" to complex, multi-stage methods, but they are unified in that they "plug in" estimates for these unknown terms to solve for h_{opt}. CV methods simply choose the value for h (generally from some grid of possible values) that minimizes the CV error, typically using leave-one-out CV. We discuss CV methods first.

Fan and Gijbels (1992) describe a simple method for estimating h using cross-validation. To find an estimate for the global optimal bandwidth, we minimize:

$$\sum_{j=1}^{n}(Y_j - \widehat{m}_{-j}(X_j))^2, \tag{16}$$

where $\widehat{m}_{-j}(\cdot)$ denotes the estimated mean function leaving out the jth term. Note that the depen-

dence of $\widehat{m}(\cdot)$ on h is suppressed. Xia and Li (2002) add a weight function to reduce boundary effects, solving:

$$\widehat{h}_n = \underset{h}{\text{argmin}} CV(h); \;\; CV(h) = n^{-1} \sum_{j=1}^{n} (Y_j - \widehat{m}_{-j}(X_j))^2 G(x_t). \tag{17}$$

The resulting estimator, $\widehat{h}_n$ is asymptotically optimal with respect to MISE:

$$\lim_{n \to \infty} \left(\frac{MISE(\widehat{h_n})}{\inf_h MISE(h)} \right) = 1, \tag{18}$$

Additionally, $\widehat{h}_n$ is asymptotically normal, centered about the true optimal bandwidth (as defined by MISE), h_{opt}. This estimator also has good finite sample properties, as demonstrated via simulation. Stable estimates of h_{opt} can be obtained through the use of higher-order polynomial fits, particularly when sample sizes are large ($n > 200$). Chapter 3 of Wand and Jones (1995) also provides a good description of CV techniques in this context. Li and Racine (2004) derive the rates of convergence for bandwidths chosen through cross-validation and show that the resulting estimators are asymptotically normal about the true value. Unlike Xia and Li (2002), these results can be applied to multivariate problems.

The other school of thought for obtaining global bandwidths "plugs in" estimates of the unknown terms in an expression for the asymptotic MSE and minimizes the resulting function. Ruppert et al. (1995) provide a global bandwidth selection algorithm that performs well relative to cross-validation estiamtors in terms of both asymptotics and practical performance. Expressions for estimating the unknowns are given, and three "plug-in"-type estimators are proposed: a simple, "rule of thumb" estimator, $\widehat{h}_{ROT}$, a "direct plug-in" estimator, $\widehat{h}_{DPI}$, and a "solve-the-equation" estimator based on solving a system of equations, $\widehat{h}_{STE}$. The simplest is $\widehat{h}_{ROT}$, which estimates the mean function and variance by dividing the interval into blocks and fitting quartic functions. The direct plug-in estimator is a two-stage estimator which uses the same quartic estimates from

$\widehat{h}_{ROT}$ to obtain first-stage estimates. Finally, $\widehat{h}_{STE}$ is computed by solving a system of equations derived using estimates from the the ROT and DPI estimators.

In terms of theoretical performance, $\widehat{h}_{ROT}$ is based on an inconsistent estimator and thus has no consistency properties. However, the other two estimators are covergent to the MISE-optimal bandwidth. In simulation, all three estimators performed well, although $\widehat{h}_{ROT}$ had a tendency to "undersmooth," and $\widehat{h}_{STE}$ occasionally chose bandwidths that were larger than optimal (twice out of 3000 trials). While the differences between $\widehat{h}_{DPI}$ and $\widehat{h}_{STE}$ were small, the DPI estimator performed best in all but one setting. Another global plug-in estimator is proposed by Fan and Gijbels (1995a) is discussed in the next section.

3.3 Variable bandwidth selection

Variable bandwidths provide a compelling alternative to global bandwidths since they are more flexible and can respond to the local properties of the data. An optimal estimator will choose smaller bandwidths for points where the nearby data is jagged and larger bandwidths where the data is smoother and more linear.

Using the model given in (10), Fan and Gijbels (1992) attempt to find a function $\alpha_{opt}(x)$ to minimize Average Mean Integrated Squared Error (AMISE). The resulting expression for $\alpha_{opt}(x)$ is:

$$\alpha_{opt}(x) = \begin{cases} \left(\frac{f_X(x)[m''(x)]^2}{\sigma^2(x)}\right)^{1/5} & if\, W(x) > 0, \\ \alpha^*(x) & if\, W(x) = 0. \end{cases} \tag{19}$$

where $W(x)$ is a nonnegative weight function and $\alpha^*(x)$ can take any values greater than 0. (This result is given as Theorem 3 in Fan and Gijbels (1992).) Note that the AMISE for a global bandwidth is obtained by setting $\alpha(\cdot) = 1$. Comparing the AMISE for the optimal constant and optimal variable bandwidths, we see that

$$AMISE_{v,opt} \leq AMISE_{c,opt}. \tag{20}$$

So asymptotically, the variable bandwidth estimator is better than the global bandwidth estimator by mean integrated squared error. Also, Fan and Gijbels (1992) show that the "plug-in" estimator, $\widehat{m}(x, \widehat{\alpha}_{opt})$ is asymptotically equivalent to $\widehat{m}(x, \alpha_{opt})$, which allows us to show that the plug-in estimator is asymptotically normal about the true mean function.

As in the case of global plug-in estimators, the quality of our estimate, $\widehat{\alpha}_{opt}(\cdot)$ will depend on the quality of our estimates for the unknown functions, $f(x)$, $m''(x)$, and $\sigma^2(x)$. Cross-validation is suggested as a method to obtain estimates for $f(x)$ and $m''(x)$, while an estimate for $\sigma^2(x)$ can be obtained using the residuals, $\widehat{Y}_j = Y_j - \widehat{m}(X_j)$. These estimates are plugged into (19) to give us $\widehat{\alpha}_{n,opt}(\cdot)$, which in turn is used to calculate $\widehat{m}(\cdot, \widehat{\alpha}_{n,opt})$.

Schucany (1995) also proposes a variable bandwidth selector for kernel regression which can be extended for local linear regression. This estimator is based on the case where the values for X_i are equally-spaced, (Fan and Gijbels (1992) assumed a continuous, random distribution for the data with bounded support) leading to the nonparametric regression model:

$$Y_i = m(i/n) + e_i, \qquad i =, \ldots, n; \;\; e_i \overset{iid}{\sim} N(0, \sigma^2). \tag{21}$$

An expression for the optimal bandwidth is given by:

$$h_{opt}^{SCH}(x) = \left(\frac{\sigma^2 A}{2pnB(x)^2} \right)^{1/(2p+1)}, \tag{22}$$

where A is a constant dependent upon the kernel and $B(x)$ is an approximation for the bias. To practically estimate $h_{opt}(x)$, we need to find estimates σ^2 and B_t. Schucany (1995) suggest an estimator for $B(x)$ that is calculated using a pilot bandwidth, so the quality of our final estimator, $\widehat{h}_{opt}^{SCH}$ will depend on the choice of this "pilot bandwidth." To estimate σ^2, any $\sqrt{n}$-consistent

estimator is sufficient. It can be shown that $\widehat{h_t}/h_{opt}(x)$ converges to 1 in probability, but the rate of convergence depends on the choice of pilot bandwidths. This local estimator compared favorably to a global bandwidth estimator in simulation.

Fan and Gijbels (1995a) and Fan et al. (1996) propose two-stage, "data-driven" global and variable bandwidth estimators and flesh out their asymptotic and finite sample properties. First, we choose a pilot bandwidth using a residual squares criterion (RSC). RSC is defined thus:

$$RSC(x_0; h) = \widehat{\sigma}^2(x_0)[1 + (p + 1)V], \tag{23}$$

where V is the first diagonal element of $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$. We can now choose a global bandwidth selector based on integrated RSC:

$$IRSC = \int_{[c,d]} RSC(y; h)dy \tag{24}$$

Multiplying the minimizer of (24) by adjustment factor determined by the kernel, K (for details, see Fan and Gijbels (1995a)) gives us a pilot estimate of the global optimal bandwidth denoted $\widehat{h}^{RSC}$. (Note that $\widehat{h}^{RSC}$ depends on the choice of p and can be generalized for estimating the rth derivative.) Using this, Fan and Gijbels (1995a) now propose both global and variable bandwidth estimators. The global estimator is a simple refinement using cross-validation:

$$\widehat{h}_2^{RSC} = \underset{h}{\text{argmin}} CV(h), \left\{ \int_{[c,d]} \widehat{MSE}_{p,r}(y; h)dy \right\} \tag{25}$$

A variable bandwidth selector is chosen by breaking the support of the data into k subintervals, denoted I_k. For each interval, we minimize

$$IRSC(h) = \int_{I_k} RSC(y; h)dy, \tag{26}$$

to find a pilot bandwidth for each subinterval. Smoothing over the resulting step function by locally averaging lets us fit an order-$(p+2)$ polynomial over the support of the data. We use the estimates from this fit as pilot estimates and for each I_k, solve (25). Smoothing the step function resulting from these refined estimates, we get an estimate for the optimal variable bandwidth function.

Simulation showed that the refinement produced substantial gains in terms of speed of convergence. Also, the variable bandwidth estimator had good properties when compared against the global bandwidth estimator, including the case where the underlying function was linear and hence a constant bandwidth was actually the optimal choice.

Fan et al. (1996) built on the preceding by proving some additional theoretical results. Particularly, they give an asymptotic expansion of the conditional bias and variance, which allows us to determine their rates of convergence, showing that $\widehat{h}^{RSC}$ converges to h_{opt}. These results are also extended for derivative estimation.

Prewitt and Lohr (2006) propose a variable bandwidth estimator that reduces the need to estimate unknown equations. Fan and Gijbels (1996) treated f as unknown and were forced to estimate it, while Schucany (1995) assumed equally-spaced data. Rather than use an estimate of f, Prewitt and Lohr use the eigen values of $M_p = n^{-1}X_t^T W_t X_t$ to construct consistent estimators of the conditional variance and conditional bias. They construct their estimator for $m(x)$ in two stages. The first stage estimates $\widehat{h}_1$ by minimizing the $AMSE(t,h)$ at the point x. This preliminary estimator is consistent but can be improved upon substantially. Substituting $\widehat{h}_1$ for h_{opt} and using the expressions for conditional bias and variance derived using the eigen value representation, they construct a second-stage estimator for AMSE, $\widehat{AMSE}_2(x,h)$. Both additive parts of this second stage estimator have the local variance $\sigma^2(x)$ as a common factor, so we do not need to estimate $\sigma^2(x)$ when minimizing $\widehat{AMSE}_2(t,h)$. Thus, the second-stage estimator is not directly dependent on estimating either the variance function or the distribution of the data.

While this method is asymptotically equivalent to the previous methods presented, it seems to perform better in finite sample. Prewitt and Lohr (2006) compared their two-stage method to the

variable method suggested by Fan and Gijbels (1995a). The eigenvalue method showed substantial improvement over the "global over subinterval" method of Fan and Gijbels in terms of integrated squared error (ISE). In application, the estimated mean curve appeared somewhat jagged, and a five-point moving average was suggested for a smoother-looking curve.

In this section, we have seen a variety of different methods for choosing bandwidths. In practice, few of the variable methods are used due to computational difficulty. And while the "plug-in" methods can be superior to CV methods, CV methods are often far simpler to implement. Indeed, most existing software for R uses CV methods. (See Section 5.)

4 Other model specifications: choosing p and K

4.1 Choosing P

In addition to choosing the optimal bandwidth, it is also important to choose the appropriate order of polynomial to fit. As when choosing a bandwidth, there is a tradeoff between bias and variance. Higher-order polynomials allow for precise fitting, meaning the bias will be small, but as the order increases, so does the variance. However, this increase is not constant. The asymptotic variance for $\widehat{m}(x)$ only increases whenever p goes from odd to even. There is no loss when going from $p = 0$ to $p = 1$, but going from $p = 1$ to $p = 2$ will increase the asymptotic variance. This strongly suggests only considering odd-ordered polynomials, since the gain in bias appears to be "free", with no associated cost in variance. (Fan and Gijbels, 1995a, Ruppert and Wand, 1994)

Fan and Gijbels (1995b) suggest an adaptive method for choosing the correct order of polynomial based on local factors, allowing p to vary for different points in the support of the data. The resulting estimator has the property of being robust to bandwidth. This means that if the chosen bandwidth is too large, a higher order of polynomial will be chosen to better model the contours of the data. If the chosen bandwidth is too small, a lower order polynomial will be fit to help make the estimates numerically stable and reduce the variance. The algorithm for adaptive order fitting

is outlined thus:

Construct a grid of points, $\{x_j : j = 1, \ldots, n_{grid}\}$ and choose a maximum order to be considered, p_{max}. Fit a standard polynomial regression of order $p_{max} + a$ in order to obtain "pilot" estimates for $\widehat{\beta}^*_{p_{max}}, \ldots, \widehat{\beta}^*_{p_{max}+a}$. Using these, estimate the MSE of the fit at each grid point for each order up to p_{max} and smooth across the grid points to get an estimate of the MSE as a function of x for each candidate p. Denote this function $\widehat{MSE}_p(x_0)$. Then for every grid point, x_j, choose the p that minimizes $\widehat{MSE}_p(x_j)$, and denote this p_j.

In simulation, this method demonstrated the "robust to bandwidth" property. Estimates using the adaptive bandwidth selector were essentially the same across a variety of bandwidths differing by at most a factor of 3. Moreover, the adaptive algorithm outperformed the local linear regression in terms of mean absolute deviation error (MADE). Particularly, the adaptive order fit chose mostly linear fits except in regions of high curvature, which is where a higher-order fit would be desirable. This method also didn't overfit, performing well in the case where the true function was a straight line and the true optimal fit was linear everywhere.

Although adaptive order fitting is robust to bandwidth, consideration should still be given to choosing h. Fan and Gijbels (1995b) suggest a simple rule of thumb for computational efficiency.

4.2 Choosing K

Most of the results discussed in previous sections require assumptions about K. All assume that K is a symmetric, unimodal, and most assume the existence of some moments. Fan and Gijbels (1992) require all moments to exist while others only require a finite number. It is also common to assume a bounded support for K and that K be smooth, but these are not universal. The most commonly used kernel functions are the standard normal density and kernels of the form:

$$K(x, q) = \left(2^{2q+1} B(q+1, q+1)\right)^{-1} (1 - x^2) I_{\{|x|<1\}}, \tag{27}$$

where $B(\cdot,\cdot)$ is the beta function. For $q = 0, 1, 2, 3$ respectively, these are called the uniform, Epanechnikov, biweight, and triweight kernels. The triangular kernel, $K(x) = (1 - |x|)I_{\{|x|<1\}}$ is also used occasionally, but it lacks the smoothness property. (Wand and Jones, 1995) In Section 1.3, kNN weighting schemes were discussed, and each of the above kernels can be modified to act as a kNN weight function by defining the bandwidth in the appropriate manner. Since the kernels mentioned above meet all of the standard assumptions, choosing kernel that is optimal in some sense may be desirable. The Epanechnikov kernel is optimal in the sense that it attains the minimum AMISE most quickly in terms of sample size, but the others, including the Gaussian are not very much slower. Thus, the decision may come down to the preference of the practitioner. For example, the Epanechnikov kernel has discontinuous first derivatives which may be undesirable, so the Gaussian kernel may be chosen instead. (Wand and Jones, 1995)

5 Efficient computational methods for LPR

Local polynomial regression is more computationally complex than standard regression techniques, since a model must be fit for each observed data point. With "brute force" methods, it would take approximately n times longer to fit a local linear regression than it would take to fit a "global" linear regression even if a uniform kNN weighting function was used. When we add in kernel evaluations and complex algorithms for choosing bandwidths and orders, the problem has the potential to get computationally difficult quickly. Many methods for choosing h and p rely on pilot estimates or cross-validation. (Fan and Gijbels, 1995a, Prewitt and Lohr, 2006, Fan et al., 1996) This necessitates solving the LPR minimization repeatedly. Therefore, a good way to reduce overall computation time is to find a quick method for solving this minimization.

Fan and Marron (1994) propose two methods for solving a local polynomial regression problem. The first of these, the "updating" method, is most easily explained by imaging the data to come from an equally-spaced grid of design points. Recall that the Nadaraya-Watson estimator is essentially a weighted average over the design points in the neighborhood of x_0. Once we have

computed $\widehat{m}(X_j)$ for the N-W estimator, we can obtain $\widehat{m}(X_{j+1})$ by removing Y_{j-i_h} and adding in Y_{j+i_h+1}. This is a reduction from $O(n^2h)$ operations to just $O(n)$. This idea can be easily generalized to non-equally-spaced designs provided we keep the uniform kernel, and variable bandwidths can be accommodated without difficulty. Non-uniform kernels are more complex, but many (such as the Epinechnikov kernel) can be expanded in ways that lends themselves to updating.

The second method proposed is a "binning" procedure. This significantly reduces the number of kernel evaluations necessary to fit a local polynomial regression. The first step is to create an equally-spaced grid of g points, denoted $x_1^*, \ldots, x_g^*$. Each data point is then mapped to the nearest grid point, $X_i \mapsto x_{j(i)}$, and an index set corresponding to each grid point is created: $I_j = \{i : X_i \mapsto x_j^*\}$. Then for each grid point, we can "summarize" the corresponding binned data with the bin average, $\bar{Y}_{j(i)}$, and the bin count, $c_j = \#\{X_i : i \in I_j\}$. Now, we can estimate $m(\cdot)$ using approximations form the bins. While there is some reduction in the number of calculation since there are now $g < n$ points on which to evaluate, the major gains are due to the equally-spaced grid. Many of the remaining kernel evaluations will end up being the same. Let $\Delta = x_j - x_{j-1}$ be the "bin width." Then $x_j - x_{j-k} = k\Delta$ for every j. This leads to large computation savings, since $K_h(x_{j'} - x_j) = K_h(\Delta(j' - j))\ \forall j$.

The net result is a decrease from $O(n \cdot g)$ kernel evaluations to $O(g)$. For some, the reliance on the approximation by mapping the data to grid points may be too rough. An elegant solution to this is available through "linear binning". Instead of mapping data points to the nearest grid point, we "split" each data point, with one part going to each of the two nearest grid points along with an associated weight describing the distance the original data point was to that grid point. So if X_0 is directly between x_j and x_{j+1}, a weight of $1/2$ will be given to each of the two bins. A general function for this weight can be written thus:

$$W_{I,j} = \left(1 - \frac{|X_i - x_j|}{\Delta}\right)_+ \tag{28}$$

The new bin counts and bin "averages" can now be written:

$$c_j = \sum_{i=1}^{n} w_{I,j} \text{ and } \bar{Y}_j^* = \sum_{i=1}^{n} w_{I,j} Y_i. \tag{29}$$

Linear binning has the same computational requirements as simple binning, and since it is more precise (see Hall and Wand (1996)), it is clearly preferred to simple binning. Numerous packages exist for computing LPR in R, particularly, the $locfit$ and $KernSmooth$ packages (Loader, 2007, original by Matt Wand. R port by Brian Ripley., 2009), and the $loess$ function. (R Development Core Team, 2009) Another useful reference for computational issues in LPR is Seifert et al. (1994).

6 Extensions of local polynomial regression

Algorithms have been developed to apply LPR to difficult types of data. Cleveland (1979) constructed a LPR estimator robust to outliers. (See 1.3.) Functions with jumps in their derivative, referred to as "changepoints" can also be difficult to estimate using traditional methods. It is possible to get a good estimator using local methods by choosing a variable bandwidth such that the "change points" are not included in the local fit. Spokoiny (1998) chooses the largest interval for each point, x_0, such that the residuals from the resulting estimator are sufficiently small. This is checked using a test statistic that becomes large and enters the rejection region when the residuals are large. Intervals containing changepoints will have test statistics in the rejection region with probability close to 1, so the estimated function, $\widehat{m}(x)$ will be based on intervals on which the true function is smooth.

The variance for standard estimators can blow up if an insufficient number of data points are given positive weight (ie, if the chosen bandwidth is small), as can be the case for sparse or clustered data. The ridging estimator proposed by Seifert and Gasser (2000) deals with this problem by adding a shrinkage term to the estimator, ensuring that the conditional variance remains bounded. LPR can also be applied to derivative estimation. Li et al. (2003) propose a method for estimating

the expectation of the derivative of a mean function. This is done using a sample average of the estimated derivative function. The asymptotic distribution is derived, and the estimator is compared with existing techniques.

Due to the difficulty in the implementation nonparametric models for multivariate data, an additivity assumption may be imposed. For d-dimensional data, we have:

$$E(Y|X = x \equiv (x_1, \ldots, x_d)) = m(x) = m_1(x_1) + \cdots + m_d(x_d) \tag{30}$$

A popular method for fitting such a model is the backfitting algorithm, proposed by Buja et al. (1989) . In the context of local polynomial fitting, Opsomer and Ruppert (1997) give sufficient conditions for convergence of the backfitting algorithm and give the asymptotic properties of the estimators for the $d = 2$ case. Existence and uniqueness for $\widehat{m}_1$ and $\widehat{m}_2$ is proved given a few standard assumptions.

LPR also has applications beyond smoothing. Alcala et al. (1999) use LPR to test whether a mean function belongs to a particular parametric family. Under the null hypothesis that $m(x)$ belongs to the specified family, both parametric regression and LPR give consistent, unbiased estimates. A test statistic using these is constructed, and if the discrepancy is too great, H_0 is rejected and we conclude that the function is not in the specified family.

Kai et al. (2010) propose an alternative to LPR in the form of local composite quantile regression (CQR). While LPR is the best linear smoother (see Section 2.2), CQR is not a linear estimator, so it may still be an improvement. Indeed, for many common error distributions, this method appears to be more efficient asymptotically than LPR. LCQR can also be applied to derivative estimation.

References

JT Alcala, JA Cristobal, and W Gonzalez-Manteiga. Goodness-of-fit test for linear models based on local polynomials. *Statistics & Probability Letters*, 42(1):39–46, Mar 15 1999.

Andreas Buja, Trevor Hastie, Robert Tibshirani, and Trevor Hastieand Rober Tibshirani. R.: Linear smoothers and additive models. *Annals of Statistics*, pages 453–510, 1989.

Ming-Yen Cheng, Jianquing Fan, and JS Marron. On automatic boundary corrections. *Annals of Statistics*, 25(4):1691–1708, 1997.

William Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

William Cleveland. Regression by local fitting methods, properties, and computational algorithms. *Journal of Econometrics*, 37(1):87–114, 1988.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.

Jianqing Fan and Irene Gijbels. Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):371–394, 1995a.

Jianqing Fan and Irne Gijbels. Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, 20(4):2008–2036, 1992.

Jianqing Fan and Irne Gijbels. Adaptive order polynomial fitting: Bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, 4(3):213–227, Sep. 1995b.

Jianqing Fan and Irne Gijbels. *Local polynomial modelling and its applications*. Chapman & Hall, 1996.

Jianqing Fan and James S. Marron. Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3(1):35–56, Mar. 1994.

Jianqing Fan, Theo Gasser, Irne Gijbels, Michael Brockmann, and Joachim Engel. Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency. *Annals of the Institute of Statistical Mathematics*, 49(1), 1997.

JQ Fan, I Gijbels, TC Hu, and LS Huang. A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6(1):113–127, Jan 1996.

Theo Gasser and Hans-Georg Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11(3):171–185, 1984.

P Hall and MP Wand. On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis*, 56(2):165–184, Feb 1996.

Trevor Hastie and Clive Loader. Local regression: Automatic kernel carpentry. *Statistical Science*, 8(2):120–129, 1993.

Bo Kai, Runze Li, and Hui Zou. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 72(Part 1):49–69, 2010.

Q Li and J Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2): 485–512, Apr 2004.

Q Li, XW Lu, and A Ullah. Multivariate local polynomial regression for estimating average derivatives. *Journal of Nonparametric Statistics*, 15(4-5):607–624, Aug-Oct 2003.

Catherine Loader. *locfit: Local Regression, Likelihood and Density Estimation.*, 2007. URL `http://locfit.herine.net/`. R package version 1.5-4.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

JD Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25(1):186–211, Feb 1997.

S original by Matt Wand. R port by Brian Ripley. *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*, 2009. URL `http://CRAN.R-project.org/package=KernSmooth`. R package version 2.23-3.

Kathryn Prewitt and Sharon Lohr. Bandwidth selection in local polynomial regression using eigenvalues. *Journal of the Royal Statistical Society.Series B, Statistical Methodology*, 68(1):135–154, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`.

D. Ruppert and MP Wand. Multivariate locally weighted least-squares regression. *Annals of Statistics*, 22(3):1346–1370, Sep 1994.

D. Ruppert, SJ Sheather, and MP Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270, Dec 1995.

WR Schucany. Adaptive bandwidth choice for kernel regression. *Jounral of the American Statistical Association*, 90(430):535–540, Jun 1995.

B. Seifert and T. Gasser. Data adaptive ridging in local polynomial regression. *Journal of Computation and Graphical Statistics*, 9(2), Jun 2000.

Burkhardt Seifert, Michael Brockmann, Joachim Engel, and Theo Gasser. Fast algorithms for nonparametric curve estimation. *Journal of Computational and Graphical Statistics*, 3(2):192–213, 1994.

V. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Annals of Statistics*, 26(4):1356–1378, 1998.

Charles Stone. Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–645, 1977.

MP Wand and MC Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

Geoffrey S. Watson. Smooth regression analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Yingcun Xia and W. K. Li. Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of multivariate analysis*, 83(2):265–287, 2002.