Kernel Smoothing Methods (Part 1)

Henry Tan

Georgetown University

April 13, 2015

Introduction - Kernel Smoothing

Previously

- Basis expansions and splines.
- Use all the data to minimise least squares of a piecewise defined function with smoothness constraints.

Kernel Smoothing

A different way to do regression.

Not the same inner product kernel we've seen previously

Kernel Smoothing

In Brief

For any query point x_0, the value of the function at that point $f(x_0)$ is some combination of the (nearby) observations, s.t., $f(x)$ is smooth.

The contribution of each observation $x_i, f(x_i)$ to $f(x_0)$ is calculated using a weighting function *or* Kernel $K_\lambda(x_0, x_i)$.

λ - the width of the neighborhood

Kernel Introduction - Question

Question

Sicong 1) Comparing Equa. (6.2) and Equa. (6.1), it is using the Kernel values as weights on y_i to calculate the average. What could be the underlying reason for using Kernel values as weights?

Answer

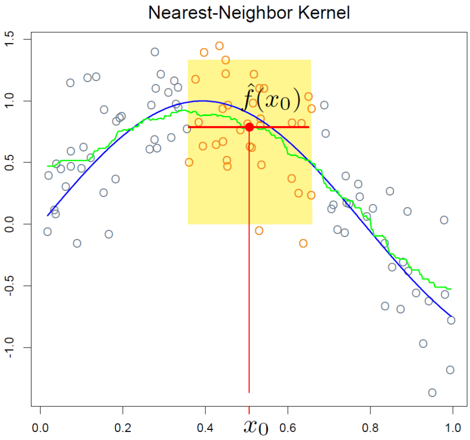
By definition, the kernel is the weighting function.
The goal is to give more importance to closer observations without ignoring observations that are further away.

K-Nearest-Neighbor Average

Consider a problem in 1 dimension x-
A simple estimate of $f(x_0)$ at any point x_0 is the mean of the k points closest to x_0.

$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x)) \tag{6.1}$$

KNN Average Example



True function

KNN average

Observations contributing to $\hat{f}(x_0)$

Problem with KNN Average

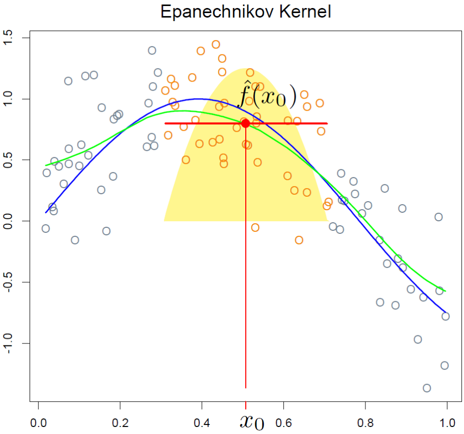
Problem

Regression function $\hat{f}(x)$ is discontinuous - “bumpy”.
Neighborhood set changes discontinuously.

Solution

Weigh all points such that their contribution drop off smoothly with distance.

Epanechnikov Quadratic Kernel Example



Estimated function is smooth

Yellow area indicates the weight assigned to observations in that region.

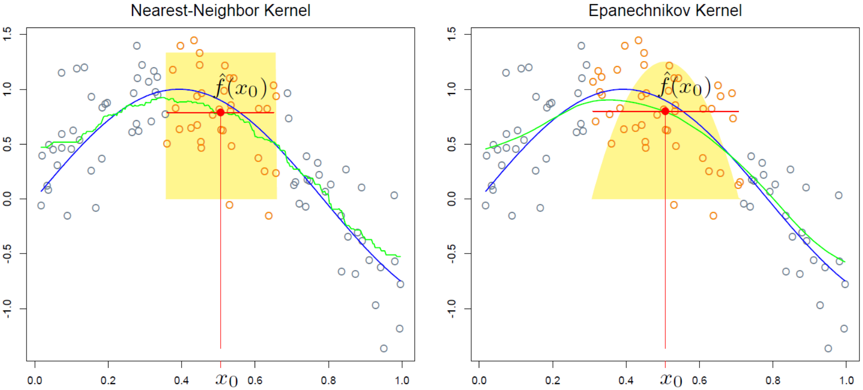
Epanechnikov Quadratic Kernel Equations

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)} \tag{6.2}$$

$$K_\lambda(x_0, x) = D(\frac{|x - x_0|}{\lambda}) \tag{6.3}$$

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if} |t| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

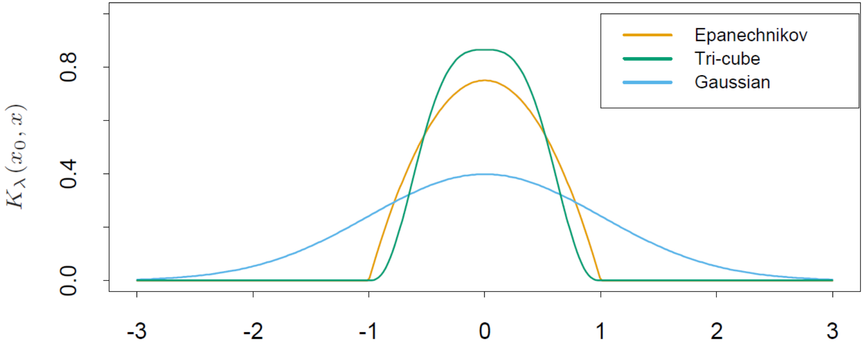
KNN vs Smooth Kernel Comparison



Other Details

- Selection of λ - covered later
- Metric window widths vs KNN widths - bias vs variance
- Nearest Neigbors - multiple observations with same x_i - replace with single observation with average y_i and increase weight of that observation
- Boundary problems - less data at the boundaries (covered soon)

Popular Kernels



Epanechnikov *Compact* (only local observations have non-zero weight)

Tri-cube Compact and differentiable at boundary

Gaussian density Non-compact (all observations have non-zero weight)

Popular Kernels - Question

Question

Sicong

2) The presentation in Figure. 6.2 is pretty interesting, it mentions that "The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none." Can you explain this more in detail in class?

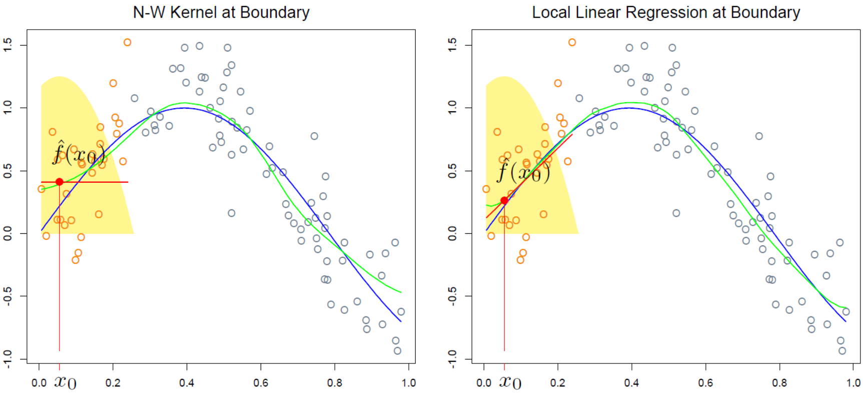
Answer

Tricube Kernel - $D(t) = \begin{cases} (1-|t|^3)^3 & \text{if} |t| \leq 1; \\ 0 & \text{otherwise} \end{cases}$

$D'(t) = 3 * (-3t^2)(1-|t|^3)^2$

Epanechnikov Kernel - $D(t) = \begin{cases} \frac{3}{4}(1-t^2) & \text{if} |t| \leq 1; \\ 0 & \text{otherwise} \end{cases}$

Problems with the Smooth Weighted Average



Boundary Bias

At some x_0 at a boundary, more of the observations are on one side of the x_0 - The estimated value becomes biased (by those observations).

Local Linear Regression

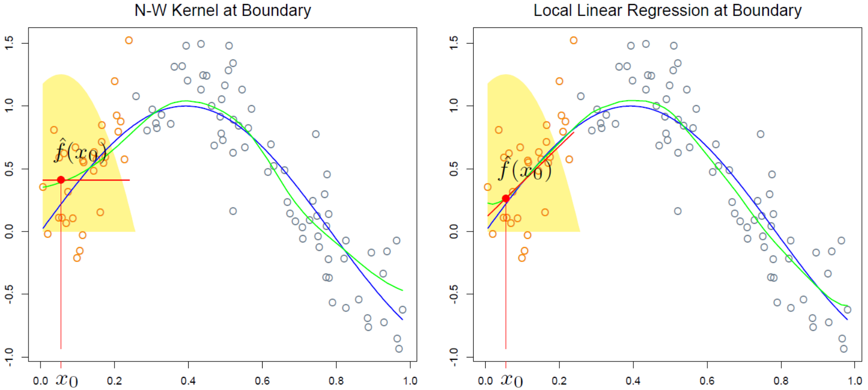
Constant vs Linear Regression

Technique described previously : equivalent to local constant regression at each query point.
Local Linear Regression : Fit a line at each query point instead.

Note

The bias problem can exist at an internal query point x_0 as well if the observations local to x_0 are not well distributed.

Local Linear Regression



N-W Kernel at Boundary

Local Linear Regression at Boundary

Local Linear Regression Equations

$$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_1)[y_i - \alpha(x_0) - \beta(x_0)x_i]^2 \tag{6.7}$$

Solve a separate weighted least squares problem at each target point (i.e., solve the linear regression on a subset of weighted points).

Obtain $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0$
where $\hat{\alpha}, \hat{\beta}$ are the constants of the solution above for the query point x_0

Local Linear Regression Equations 2

$$\hat{f}(x_0) = b(x_0)^T(\mathbf{B}^T\mathbf{W}(x_0)\mathbf{B})^{-1}\mathbf{B}^T\mathbf{W}(x_0)\mathbf{y} \tag{6.8}$$

$$= \sum_{i=1}^{N} l_i(x_0)y_i \tag{6.9}$$

6.8 : General solution to weighted local linear regression
6.9 : Just to highlight that this is a linear model (linear contribution from each observation).

Question - Local Linear Regression Matrix

Question

Yifang
1. What is the regression matrix in Equation 6.8? How does not Equation 6.9 derive from 6.8?

Answer

Apparently, for a linear model (i.e., the solution is comprised of a linear sum of observations), the least squares minimization problem has the solution as given by Equation 6.8.
Equation 6.9 can be obtained from 6.8 by expansion, but it is straightforward since **y** only shows up once.

Historical (worse) Way of Correcting Kernel Bias

Modifying the kernel based on "theoretical asymptotic mean-square-error considerations" (don't know what this means, probably not important).

Linear local regression : Kernel correction to first order (*automatic kernel carpentry*)

Locally Weighted Regression vs Linear Regression - Question

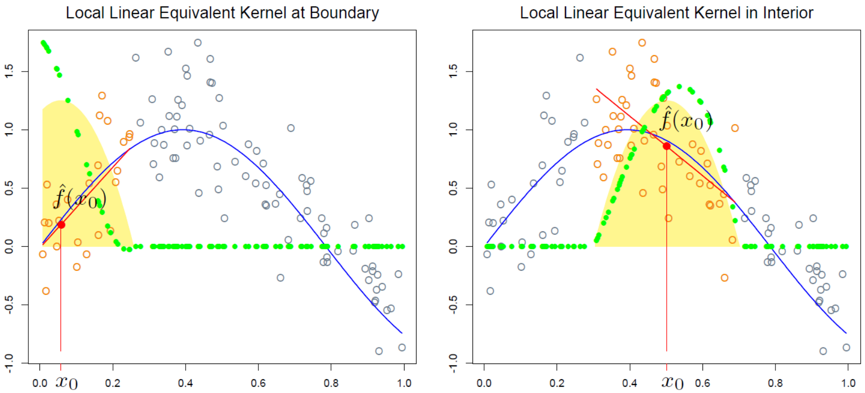
Question

Grace 2. Compare locally weighted regression and linear regression that we learned last time. How does the former automatically correct the model bias?

Answer

Interestingly, simply by solving a linear regression using local weights, the bias is accounted for (since most functions are approximately linear at the boundaries).

Local Linear Equivalent Kernel



Dots are the equivalent kernel weight $l_i(x_0)$ from 6.9
Much more weight are given to boundary points.

Bias Equation

Using a taylor series expansion on $\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0) f(x_i)$,

the bias $\hat{f}(x_0) - f(x_0)$ is dependent only on superlinear terms.

More generally, polynomial-p regression removes the bias of p-order terms.

Local Polynomial Regression

Local Polynomial Regression

Similar technique - solve the least squares problem for a polynomial function.

Trimming the hills and *Filling the valleys*

Local linear regression tends to flatten regions of curvature.

Question - Local Polynomial Regression

Question

Brendan 1) Could you use a polynomial fitting function with an asymptote to fix the boundary variance problem described in 6.1.2?

Answer

Ask for elaboration in class.

Question - Local Polynomial Regression

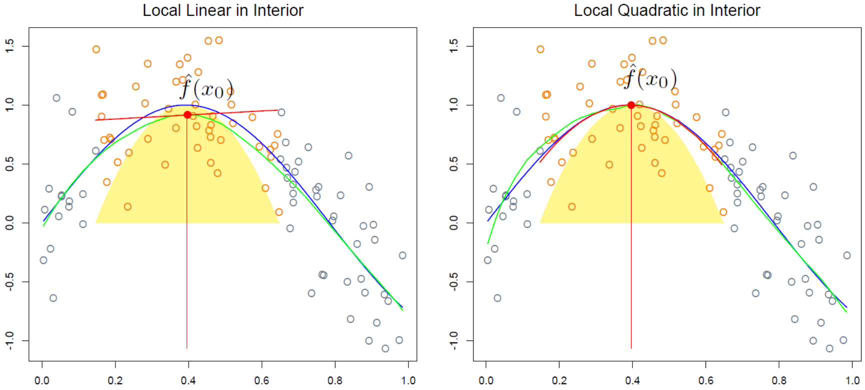
Question

Sicong 3) In local polynomial regression, can the parameter d also be a variable rather than a fixed value? As in Equa. (6.11).

Answer

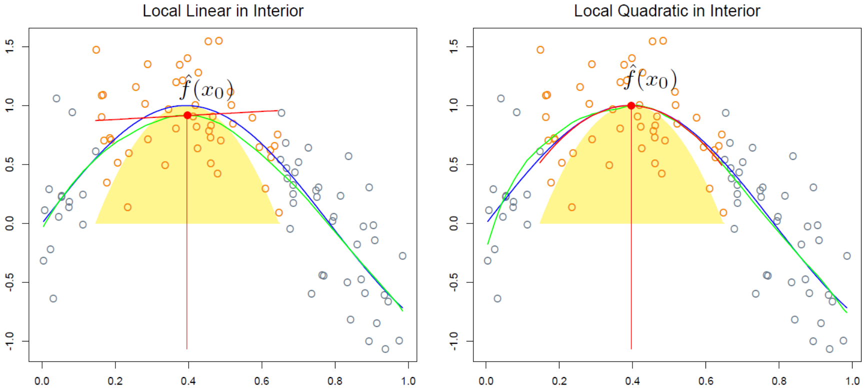
I don't think so. It seems that you have to choose the degree of your polynomial before you can start solving the least squares minimization problem.

Local Polynomial Regression - Interior Curvature Bias


Local Linear in Interior

Local Quadratic in Interior

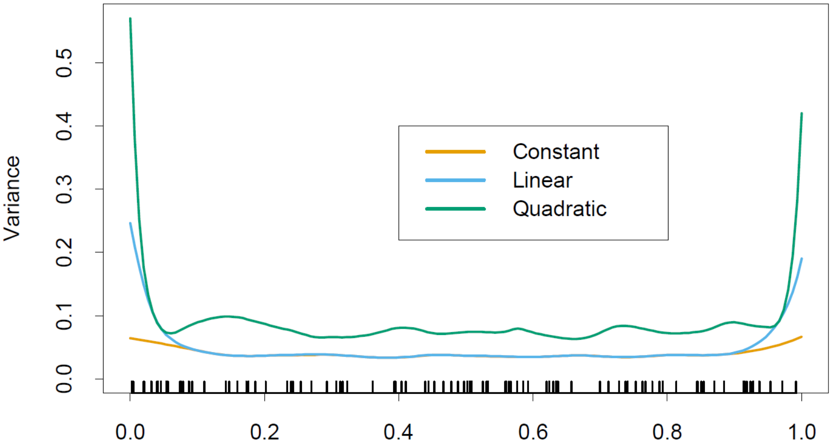
Cost to Polynomial Regression



Variance for Bias

Quadratic regression reduces the bias by allowing for curvature.
Higher order regression also increases variance of the estimated function.

Variance Comparisons



Final Details on Polynomial Regression

- Local linear removes bias dramatically at boundaries
- Local quadratic increases variance at boundaries but doesn't help much with bias.
- Local quadratic removes interior bias at regions of curvature
- Asymptotically, local polynomials of odd degree dominate local polynomials of even degree.

Kernel Width λ

Each kernel function K_λ has a parameter which controls the size of the local neighborhood.

- Epanechnikov/Tri-cube Kernel , λ is the fixed size radius around the target point
- Gaussian kernel, λ is the standard deviation of the gaussian function
- $\lambda = k$ for KNN kernels.

Kernel Width - Bias Variance Tradeoff

Small $\lambda =$ Narrow Window

Fewer observations, each contribution is closer to x_0:
High variance (estimated function will vary a lot.)
Low bias - fewer points to bias function

Large $\lambda =$ Wide Window

More observations over a larger area:
Low variance - averaging makes the function smoother
Higher bias - observations from further away contribute to the value at x_0

Local Regression in $\mathcal{R}^p$

Previously, we considered problems in 1 dimension.
Local linear regression fits a local hyperplane, by weighted least squares, with weights from a p-dimensional kernel.

Example : dimensions $p = 2$, polynomial degree $d = 2$

$b(X) = (1, X_1, X_2, X_1^2, X_2^2, X_1 X_2)$
At each query point $x_0 \in \mathcal{R}^p$, solve

$$\min_{\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_1)(y_i - b(x_i)^T \beta(x_0))^2$$

to obtain fit $\hat{f}(x_0) = b(x_0)^T \hat{\beta}(x_0)$

Kernels in $\mathcal{R}^p$

Radius based Kernels

Convert distance based kernel to radius based: $D(\frac{x-x_0}{\lambda}) \rightarrow D(\frac{||x-x_0||}{\lambda})$

The Euclidean norm depends on the units in each coordinate - each predictor variable has to be normalised somehow, e.g., unit standard deviation, to weigh them properly.

Question - Local Regression in $\mathcal{R}^p$

Question

Yifang

What is the meaning of “local” in local regression? Equation 6.12 uses a kernel mixing polynomial kernel and radial kernel?

Answer

I think “local” still means weighing observations based on distance from the query point.

Problems with Local Regression in $\mathcal{R}^p$

Boundary problem

More and more points can be found on the boundary as p increases. Local polynomial regression still helps automatically deal with boundary issues for any p

Curse of Dimensionality

However, for high dimensions $p > 3$, local regression still isn't very useful - As with the problem with Kernel width,

- difficult to maintain localness of observations (for low bias)
- sufficient samples (for low variance)

Since the number of samples increases exponentially in p.

Non-visualizable

Goal of getting a smooth fitting function is to visualise the data which is difficult in high dimensions.

Structured Local Regression Models in $\mathcal{R}^p$

Structured Kernels

Use a positive semidefinite matrix **A** to weigh the coordinates (instead of normalising all of them) -
$K_{\lambda,A}(x_0, x) = D(\frac{(x-x_0)^T \mathbf{A}(x-x_0)}{\lambda})$

Simplest form is the diagonal matrix which simply weighs the coordinates without considering correlations.
General forms of **A** are cumbersome.

Question - ANOVA Decomposition and Backfitting

Question

Brendan 2. Can you touch a bit on the backfitting described in 6.4.2? I don't understand the equation they give for estimating g_k.

Structured Local Regression Models in $\mathcal{R}^p$ - 2

Structured Regression Functions

Ignore high order correlations, perform iterative backfitting on each sub-function.

In more detail

Analysis-of-variance (ANOVA) decomposition form-

$$f(X_1, X_2, ..., X_p) = \alpha + \sum_j g_j(X_j) + \sum_{k<l} g_{kl}(X_k, X_l) + ...$$

Ignore high order cross terms, e.g., $g_{klm}(X_k, X_l, X_m)$ (to reduce complexity)

Iterative backfitting -
Assume that all terms are known except for some $g_k(X_k)$.
Perform local (polynomial) regression to find $\hat{g}_k(X_k)$.
Repeat for all terms, and repeat until convergence.

Structured Local Regression Models in $\mathcal{R}^p$ - 3

Varying Coefficients Model

Perform regression over some variables while keeping others constant.

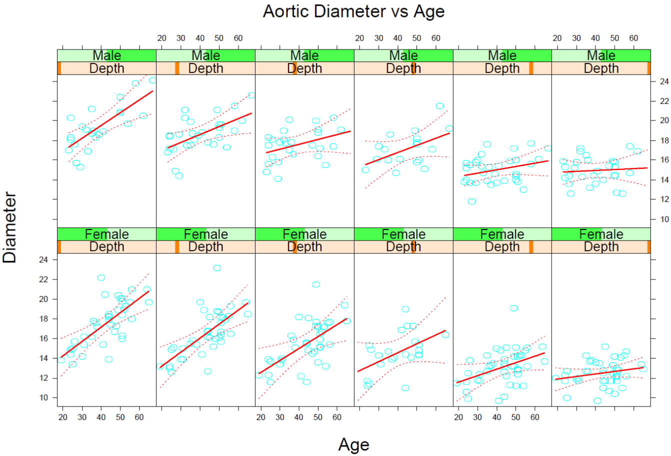
Select the first p out of q predictor variables from $(X_1, X_2, ..., X_q)$ and express the function as
$f(X) = \alpha(Z) + \beta_1(Z)X_1 + ... + \beta_q(Z)X_q$
where for any given z_0, the coefficients α, β_i are fixed.

This can be solved using the same locally weighted least squares problem.

Varying Coefficients Model Example

Human Aorta Example

Predictors - (*age*, *gender*, *depth*) , Response - *diameter*
Let Z be (*gender*, *depth*) and model *diameter* as a function of *depth*



Question - Local Regression vs Structured Local Regression

Question

Tavish

1. What is the difference between Local Regression and Structured Local Regression? And is there any similarity between the “structuring” described in this chapter to the one described in previous one where the input are transformed/structured into a different inputs that are fed to linear models?

Answer

Very similar I think. But the interesting thing is that different methods have different “natural” ways to perform the transformations or simplifications.

Discussion Question

Question

Tavish
3)As a discussion question and also to understand better, this main idea of this chapter to find the best parameters for kernels keeping in mind the variance-bias tradeoff. I would like to know more as to how a good fit/model can be achieved and what are the considerations for trading off variance for bias and vice-versa? It would be great if we can discuss some examples.

Starter Answer

As far as the book goes - if you want to examine the boundaries, use a linear fit for reduced bias. If you care more abount interior points, use a quadratic fit to reduce internal bias (without as much of a cost in internal variance).