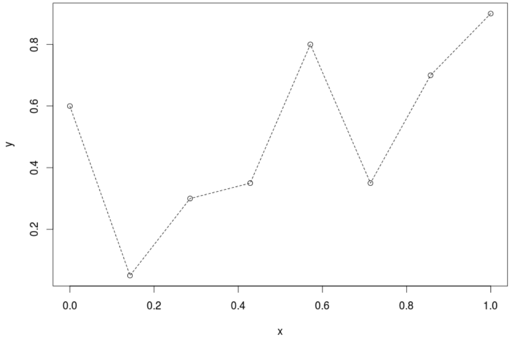
The Elements of Statistical Learning

Ch5: Basis expansion and Local regression methods

Philip Lin

Data Science in Hsinchu

2015.10.07



A quick question

" How to connect these 8 points smoothly ? "

Mathemetician said...

" How do you define smoothness ? "

Thm (*Natural cubic spline is the smoothest interpolators*)

Of all function that are continuous on $[x_1, x_m]$, have absolutely continuous first derivatives and interpolate $\{x_i, y_i\}$, **natural cubic spline** g is the one that is smoothest in the sense of minimizing

$$\int_{x_1}^{x_m} f''(x)^2 \mathrm{d}x$$

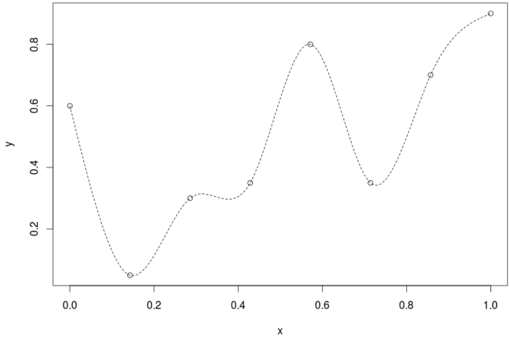
where **natural cubic spline** is defined as a function that satisfies

1. g interpolate all data points, $\{x_i, y_i\}$
2. on each interval $[x_i, x_{i+1}]$, g is a function made up of sections of cubic polynomial.
3. Except for boundaries, the function g is continuous to second derivative.
4. $g''(x_1) = g''(x_m) = 0$, i.e., the function is linear beyond the boundary.

Note:

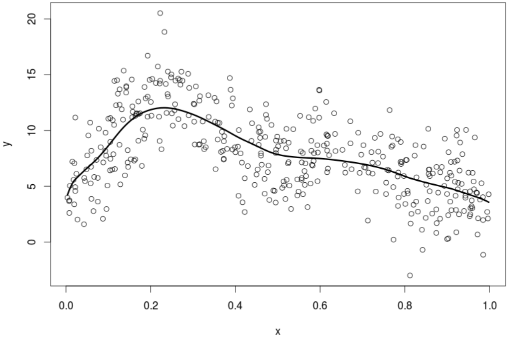
- When we remove boundary restriction (4.), we got **cubic spline**.
- It is claimed that cubic splines are the lowerst-order spline for which the knot-discontinuity is not visible to the human eye.

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.



Statistican said

" We don't connect points, we seek for the trend of data."


y
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
x

Framework

Today, we discuss the fundamental methods about one-dimensional nonparametric curve fitting

- Basis expansion
  - Natural cubic spline
  - Regression spline
  - Smoothing spline and basis expansion
  - an example: wavelet series expansion
- Local regression method
  - Local weighted average
  - Local linear regression
  - Local polynomial regression

Keywords:

infinite-dimensional function, basis, kernel function, Nearest-Neighbor, equivalent kernel, optimal bandwith

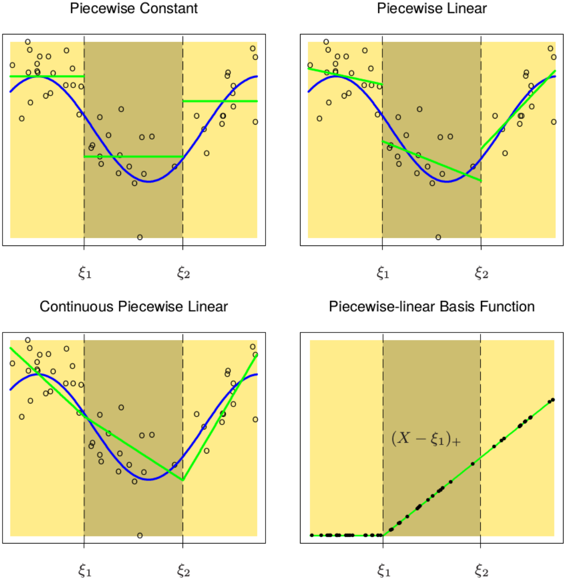
Let's go back to statistical problem (ASSUME X has one dimension today.)

$$y = f(X) + \epsilon$$

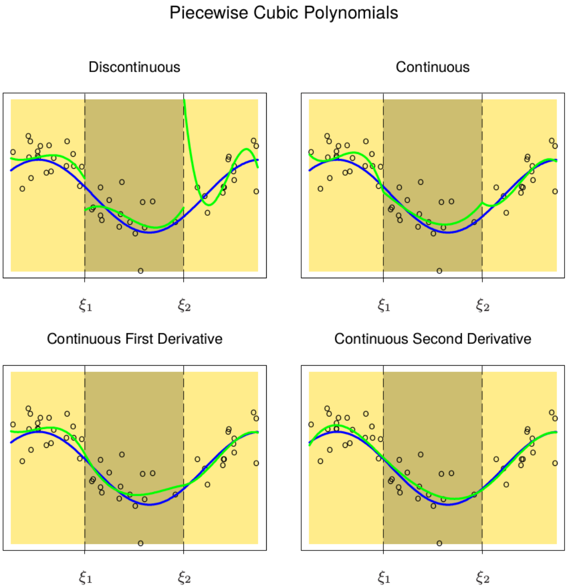
- ▷ In the linear assumption, we apply the first-order Taylor approximation to $f(X)$

$$\begin{aligned} f(X) =& E(Y|X) \\ \approx& E(Y|X=x) + \frac{\partial E(Y|X=x)}{\partial X}(X-x) \\ =& \beta_0 + \beta_1 X \end{aligned}$$

- ▷ What if $f(X)$ is nonlinear ?
  - In linear regression, we have two parameters (intercept, and slope). When nonlinear, how many parameters shall we estimate ? Actually, in this case, we are estimating an **infinitely dimensional** function.
  - Borrowed the idea from spline function, we assume $f(X)$ is **smooth** and estimate $f(X)$ **locally**.



Piecewise Constant
Piecewise Linear
Continuous Piecewise Linear
Piecewise-linear Basis Function
$(X - \xi_1)_+$
ξ_1
ξ_2


Piecewise Cubic Polynomials
Discontinuous
Continuous
Continuous First Derivative
Continuous Second Derivative
ξ_1
ξ_2

Remarks

- We ask our local cubic polynomial satisfy
  1. continuous second derivative at all knots
  2. cubic polynomial fitting in each subsection

  the fitted line (green) get closed to the true function $f(X)$ (blue).
- This procedure is quite similar to that in natural cube spline function. Isn't it ?
- Actually, the fitted model can be represented as

$$\hat{f}(X) = \sum_{i=1}^{6} \hat{\beta}_i h_i(X)$$

  where

$$\begin{aligned} h_1(X) = 1, \;\; h_3(X) = X^2, \;\; h_5(X) = (X - \xi_1)_+^3 \\ h_2(X) = X, \;\; h_4(X) = X^3, \;\; h_6(X) = (X - \xi_2)_+^3 \end{aligned}$$

- Once we decide **knots** (ξ_1, ξ_2) and **basis functions** (h_i's), we approximate $E(Y|X)$ by usual linear regression with transformed X instead. We call this method Regression Spline.
- What's are the characteristics of functions h_j ? they are unrelated to our data and well formulated.

Questions

1. In regression spline, how do we decide knots ? More for better or less for better ?

   What if I choose all data points as knots, i.e., $(\xi_1, \ldots, \xi_N)$
2. In this example, is cubic polynomial the unique choice for basis ? Generally, what kind of function is a valid basis that can reconstruct $f(X)$?

There is another way out

$$RSS(f, \lambda) = \sum_{i=1}^{N} \left\{ y_i - f(x_i) \right\}^2 + \lambda \int \left\{ f''(t) \right\}^2 \mathrm{d}t$$

Remarks

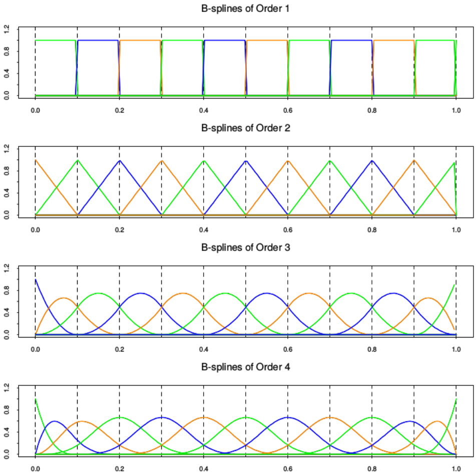
- This penalized approach get closer to the idea of natural cubic spline.
- When $\lambda = 0$, f can be any function that interpolates the data.

  When $\lambda = \infty$, simple least square line fit.
- The criterion is defined on an infinite-dimensional function space. However, it can be shown that the unique minimizer is N-dimensional natural cubic spline with knots at each of $x_i, i = 1, \dots, N$

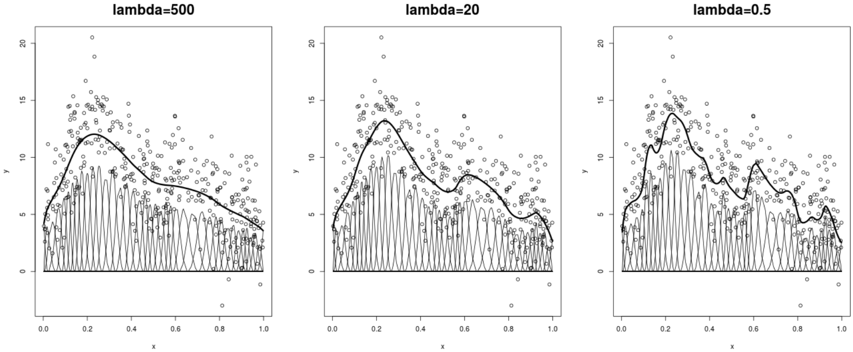
$$f(x) = \sum_{j=1}^{N} N_j(x)\theta_j$$

  where $N_j(x)$ are an $N-$dimensional set of basis function for representing the family of natural splines.
- Compared to regression spline, it transfer the complexity of knots position to the tuning parameter λ. This is called Smoothing Spline.

an example of basis in natural cubic family: B-spline basis



a simulated data fitted by B-spline basis: $N = 400, m = 40$



How to choose λ ?

eye-balling selection (Ni-Shuang-Ghiu-Hao)

The criterion can be reduced to

$$RSS(\theta, \lambda) = (\boldsymbol{y} - \boldsymbol{N}\theta)^T(\boldsymbol{y} - \boldsymbol{N}\theta) + \lambda\theta^T\boldsymbol{\Omega}_N\theta$$

where $\{\boldsymbol{N}\}_{ij} = N_j(x_i)$ and $\{\boldsymbol{\Omega}_N\}_{jk} = \int N_j''(t)N_k''(t)\mathrm{d}t$

the solution is β can be easily obtaind (ridge solution)

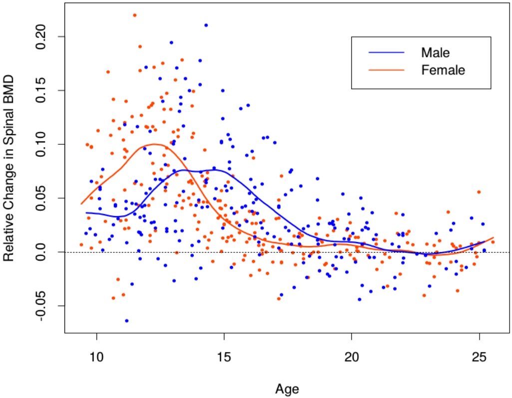
$$\hat{\theta} = (\boldsymbol{N}^T\boldsymbol{N} + \lambda\boldsymbol{\Omega}_N)^{-1}\boldsymbol{N}^T\boldsymbol{y}$$

The fitted smoothing spline is given by

$$\hat{f}(x) = \sum_{j=1}^{N} N_j(x)\hat{\theta}_j$$

Remarks

- In fact, we can set m ($m < n$) basis to reconstruct $f(x)$, i.e., dimension of matrix $\boldsymbol{N}$ is $N \times m$ and we only need to solve m dimensional parameters, θ.
- In practice, m is usually much less then N so that smoothing spline is a kind of diemsnion reduction.



The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with $\lambda \approx 0.00022$.

Till now, we focus on the family of cubic spline family.

Is it the only choice ?

A formal discussion (Grace Wahba, 1990)

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} L\big(y_i, f(x_i)\big) + \lambda \|f\|^2_{\mathcal{H}_K} \right]$$

where $\mathcal{H}_K$ is called a reproducing kernel Hilbert space (RKHS)

- The unique solution of this infinite-dimensional problem is finite-dimensional

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$$

  where $h_i(x) = K(x, x_i)$ is the basis function and the RKHS is generated by a positive definite kernel function, $K(x, y)$.

- The problem thus can be reduced to

$$\min_{\boldsymbol{\alpha}} L(\boldsymbol{y}, \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}$$

  where $\boldsymbol{K}$ is the $N \times N$ matrix with ijth entry $K(x_i, x_j)$.

- A necessary and sufficient condition for $\boldsymbol{K}$ to be a valid kernel is that $\boldsymbol{K}$ should be positive semidefinite for all possible choice of set $\{x_n\}$
- Once we can construct kernel $\boldsymbol{K}$, the solution $\boldsymbol{\alpha}$ can be easily obtained. It is always possible to define a kernel by choosing a linearization function ϕ and an inner product. (ref: http://crsouza.com/2010/03/kernel-functions-for-machine-learning-applications/)

Wahba, G. (1990), Spline Models for Observational Data, Philadelphia: SIAM

Examples of commonly used basis function

- Truncated power basis
- B-spline basis
- $\surd$ Wavelet basis (very interesting)
- Eigen-basis

Wavelet series expansion

Given $f \in L^2[0,1]$, the Wavelet series expansion is

$$f(x) = \sum_{k=0}^{2^{j_0}-1} c_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{jk}\psi_{jk}(x)$$

$$\begin{aligned}\phi_{j_0,k}(x) =& 2^{j_0/2}\phi(2^{j_0}x-k)\\ \psi_{jk}(x) =& 2^{j/2}\psi(2^j x-k), j = j_0, j_0+1, \cdots\end{aligned}$$

where

- ϕ is father wavelet
  ψ is mother wavelet
- $c_{j_0 k} = \int f(t)\phi_{j_0 k}(t)\mathrm{d}t$
  $d_{jk} = \int f(t)\psi_{jk}(t)\mathrm{d}t$
- support of $\psi_{jk} = [k2^{-j}, (k+1)2^{-j})$
- $\{c_{j_0 k}\}$ and $\{d_{jk}\}$ can be estimated empirically

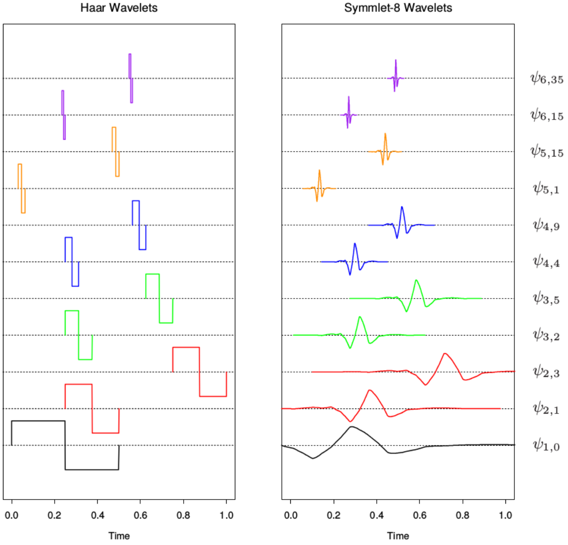
$$\hat{c}_{j_0 k} = \frac{1}{n}\sum_{i=1}^{n}\phi_{j_0 k}(t_i)y_i, \quad \hat{d}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(t_i)y_i$$

It can be solved by Discrete Wavelet Transform. (a super fast algorithm)

Haar wavelet basis (one of the most simplest wavelet basis)

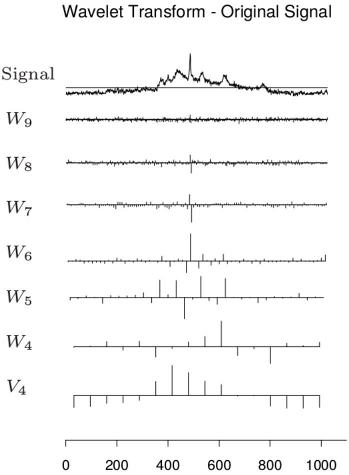
$$\phi(x) = \mathbf{1}(0 \leq x < 1)$$

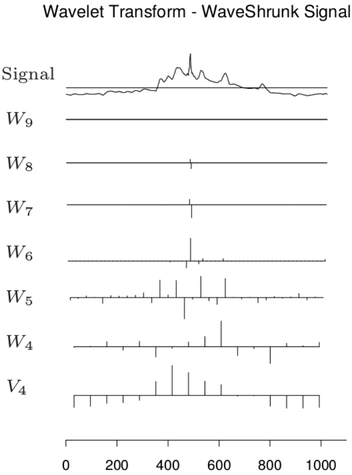
$$\psi(x) = \begin{cases} 1, & 0 \leq x < 1/2 \\ -1, & 1/2 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

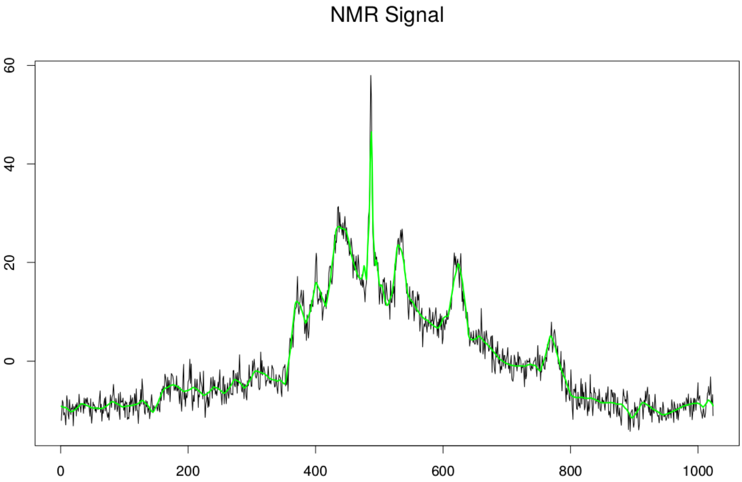


Haar Wavelets
Symmlet-8 Wavelets
$\psi_{6,35}$
$\psi_{6,15}$
$\psi_{5,15}$
$\psi_{5,1}$
$\psi_{4,9}$
$\psi_{4,4}$
$\psi_{3,5}$
$\psi_{3,2}$
$\psi_{2,3}$
$\psi_{2,1}$
$\psi_{1,0}$
0.0
0.2
0.4
0.6
0.8
1.0
Time

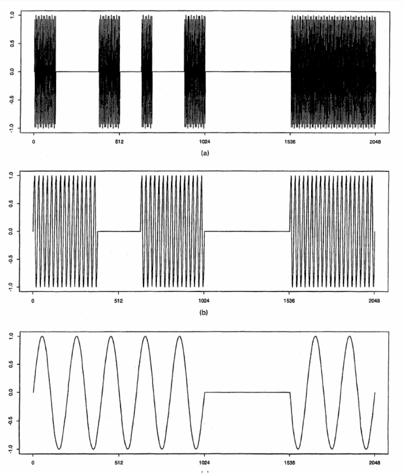
Example 1. (Signal preprocessing, or denosing)

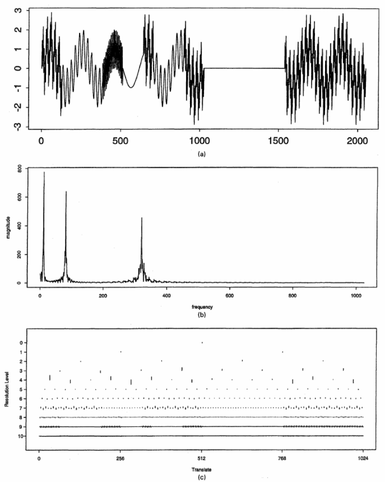





NMR Signal
60
40
20
0
0
200
400
600
800
1000

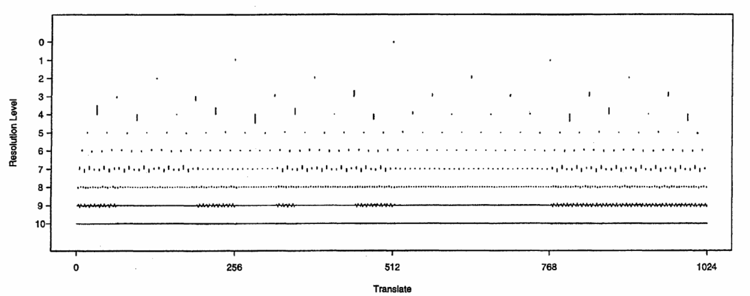
Example 2. (violin, cello, base)





For Frourier Transform, we can capture three peaks of frequency $\{10, 80, 320\}$, but we cannot understand the playing-time for each instrument.

Abramovich, F., Bailey, T.C. & Sapatinas, T., 2000. Wavelet analysis and its statistical applications. The Statistician, 49, pp.1-29



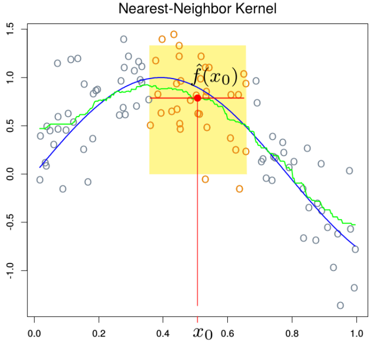
From Wavelet, Resolution level 4 indicates the coefficients for base; 7 for cello and 9 for violin.

Basis expansion approach for nonparametric model fitting is like Lugo.

(ref: www.baconbrix.com)

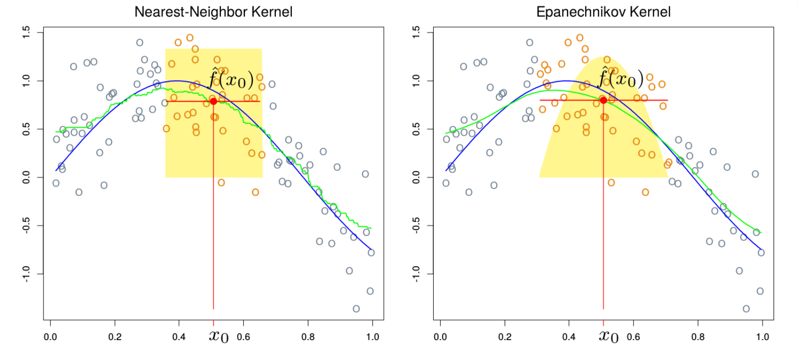
We can also fit the nonparametric curve by nearest neighbor

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



where $N_k(x)$ is the set of k points nearest to x is squared distance.

- The baisc idea to to relax the definition of conditional expection $E(Y|X = x)$ and compute an average in a neighborhood of the target point.
- Why the green lines is so ugly ? or why the discontinuity ?



improve discontinuity

- ▷ Rather than give all the points in the neighborhood equal, we can assign weights that die off smoothly with distance from the target point.
- ▷ As we move the target from left to right, points enter the neighborhood initially with weight zero, and then their contribution slowly increases.

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

with Epanechnikov quadratic kernel

$$K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{\lambda}\right), \quad D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

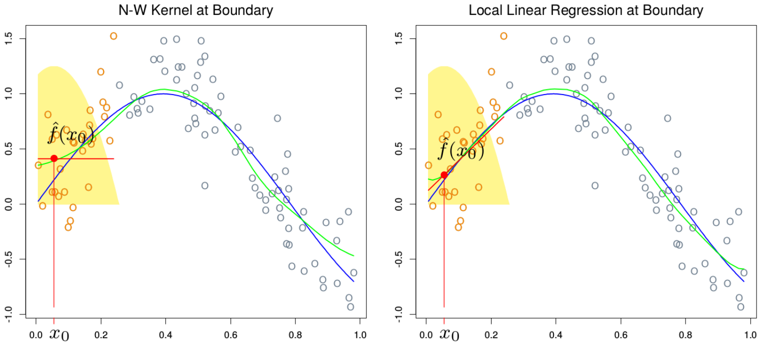
Remarks

- Two things should be decided in advance
  1. What kernel should we use ? (https://en.wikipedia.org/wiki/Kernel_(statistics))
  2. **band width** λ, how much neighbors should we include ?
- Larger λ implies lower variance (averages over more observations) but higher bias (we essentially assume the true function is constant within the window).
- Local weighted averages, or **Nadaraya-Watson**, can be badly biased on the boundaries of the domain because of the asymmetry of the kernel in that region.

Local linear regression

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

By fitting straight lines rather than constants locally, we can remove this boundary bias. In other words, we assume linearity out of boundary.



we can formulate the local fitted solution

$$\begin{aligned}\hat{f}(x_0) =& \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0 \\ =& b(x_0)^T \left(\boldsymbol{B}^T \boldsymbol{W}(x_0)^{-1} \boldsymbol{B}\right)^{-1} \boldsymbol{B}^T \boldsymbol{W}(x_0) \boldsymbol{y} \\ =& \sum_{i=1}^{N} l_i(x_0) y_i\end{aligned}$$

where

- $b(x)^T = (1, x)$, and $\boldsymbol{B}$ be the $N \times 2$ design matrix with ith row $b(x_i)^T$
- $\boldsymbol{W}(x_0)$ is the $N \times N$ diagonal matrix with ith diagonal element $K_\lambda(x_0, x_i)$.
- $l_i(x_0)$ is refered as **equivalent kernel** and can be shown that $\sum_{i=1}^{N} l_i(x_0) = 1$ in local linear case.

under the structure of

$$y = f(X) + \epsilon, \quad \epsilon \sim N(n, \sigma^2)$$

consider the expansion of $E\hat{f}(x_0)$

$$\begin{aligned} E\hat{f}(x_0) =& \sum_{i=1}^{N} l_i(x_0) f(x_i) \\ =& f(x_0) \sum_{i=1}^{N} l_i(x_0) + f'(x_0) \sum_{i=1}^{N} (x_i - x_0) l_i(x_0) \\ &+ \frac{f''(x_0)}{2} \sum_{i=1}^{N} (x_i - x_0)^2 l_i(x_0) + R \end{aligned}$$

we can see that

- for local linear regression,

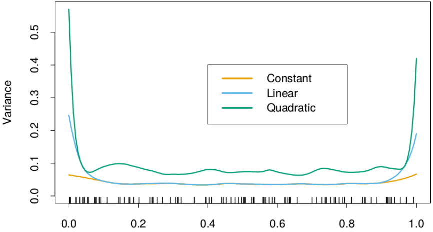
$$\sum_{i=1}^{N} l_i(x_0) = 1 \text{ and } \sum_{i=1}^{N} (x_i - x_0) l_i(x_0) = 0$$

  the bias $E\hat{f}(x_0) - f(x_0)$ depends only on the quadratic and higher-order terms in the expansion of f.

We can extend local linear to higher order: **Local polynomial regression**

$$\min_{\alpha(x_0),\beta_j(x_0),j=1,\dots,d} \sum_{i=1}^{N} K_\lambda(x_0,x_i)\left[y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta_j(x_0) x_i^j\right]^2$$

Giving higher order in local expansion will reduce bias, but the variance will be dramatically increased. (variance function: $\|l(x)\|^2$)



Remarks

- Local polynomial method selects bandwidth to control the complexity of local fitting. The controller of complexity corresponds to the tuning parameter in smoothing spline approach.
- It is easier to evaluate (pointwise) asymptotic bias and variance in local polynomial. However, the fitting performance can only evaluated by your eyes in spline smoothing.
- In local linear case, the optimal bandwith can be evaluated

$$h_{opt} = C_0 n^{-\frac{1}{5}}$$