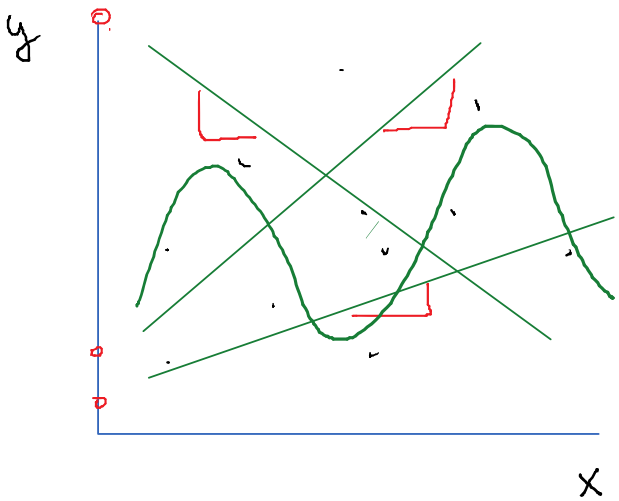
What makes a nonparametric model "nonparametric"?

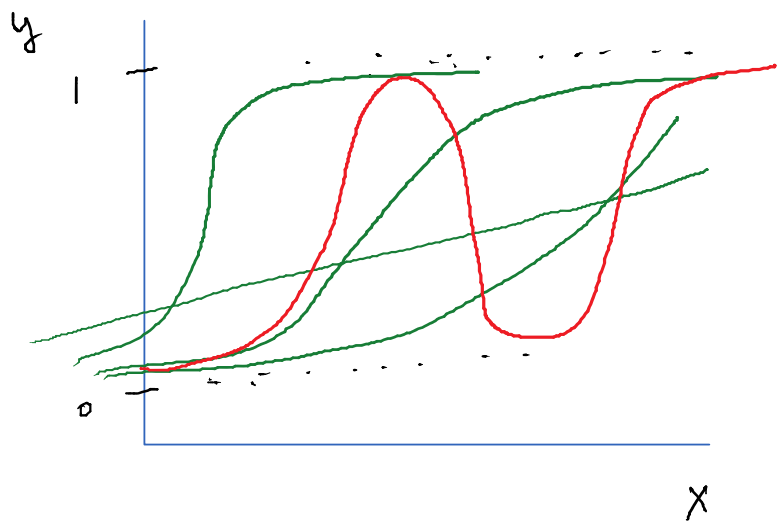
Tuesday, January 21, 2014 5:29 PM

- distinction: parametric models presume a strong functional linkage (x and y)

data are used to tune/adjust parameters inside the function to provide good fit.



$$y = \beta_0 + \beta_1 x + \varepsilon$$



$$pr(y=1) = \Phi(\beta_0 + \beta_1 x)$$

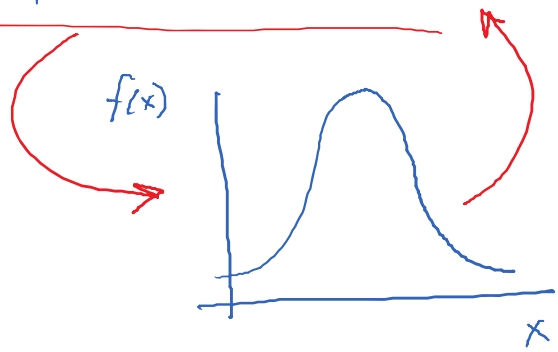
Non-parametric models allow greater freedom in fit.

Histograms and Kernel Density Estimation

Monday, January 20, 2014 3:16 PM

$X = \{1, 5, 4, 3, \ldots\}$

$f(x)$

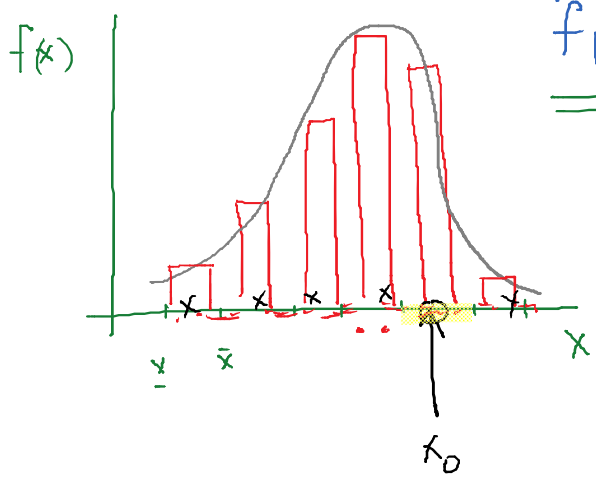


<u>parametric</u>: assume that $X \sim \Phi(\mu, \sigma^2)$

use data to estimate $\hat{\mu}$, $\hat{\sigma}^2$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

<u>Histogram</u>



$$\hat{f}_h(x) = \frac{1}{N \cdot h} \cdot \sum_{i=1}^{N} \sum_{j} I(x_i \in B_j)\, I(x \in B_j)$$

is obs. x_i in B_j? $= 1$ yes, $= 0$ no

is x in B_j? $= 1$ yes, $= 0$ no

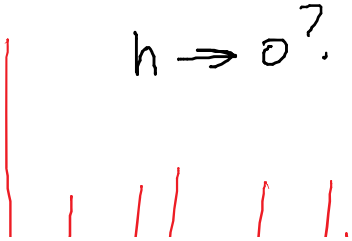
h = bin width

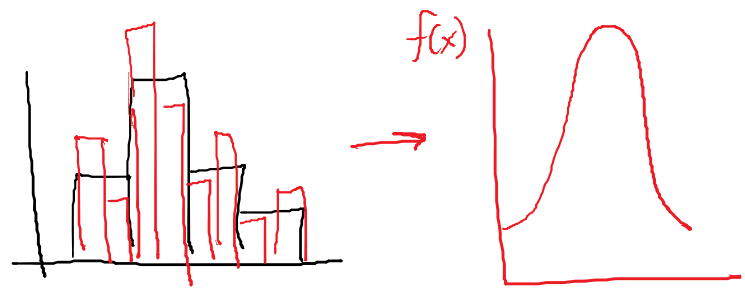
N = sample size

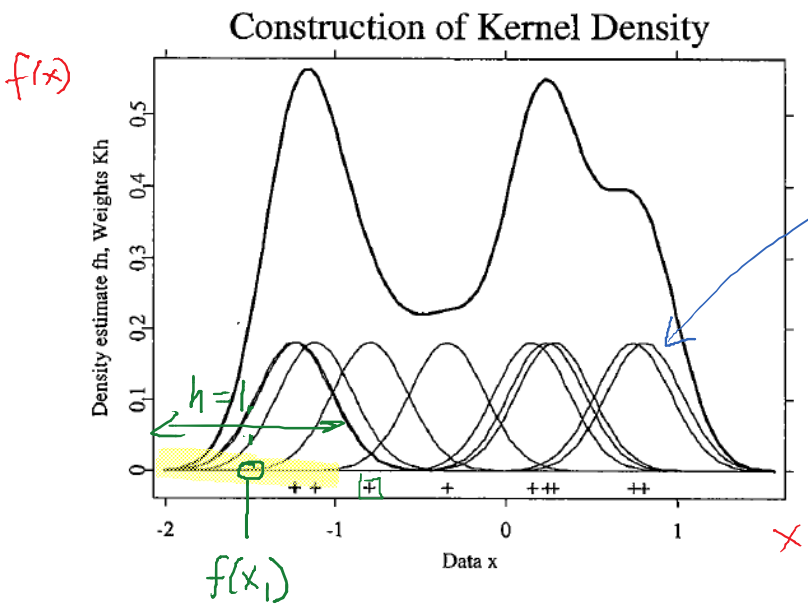
$$\hat{f}_h(x) = \frac{1}{N \cdot h} * \# B_j$$

$f(x)$

$h \to 0$?







Kernel = K

$$\frac{1}{Nh} \sum_{i=1}^{N} K(x_i | x_0)$$

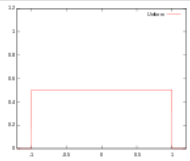
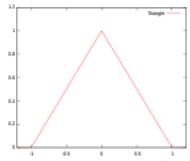
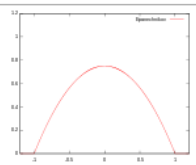
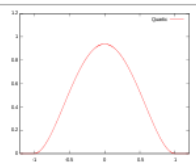
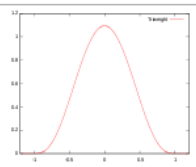
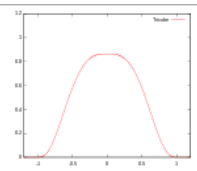
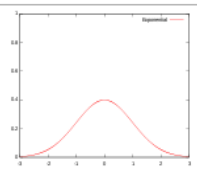
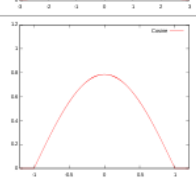
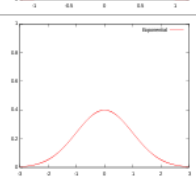
$$\int \sum_{i=1}^{N} K(x_i | x)\, dx = 1$$

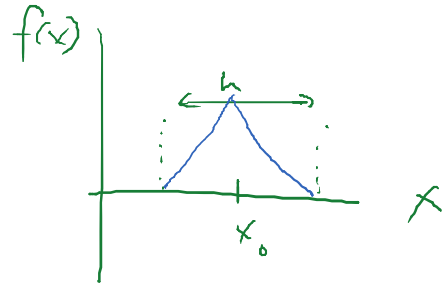
Figure 3.5. Kernel density estimate as a sum of bumps SPMkdeconstruct

h = bandwidth

Kernel Choices

Tuesday, January 21, 2014 5:34 PM

Kernel Functions, *K*(*u*)			$\int u^2 K(u)du$	$\int K^2(u)du$
Uniform	$K(u) = \frac{1}{2}\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{1}{3}$	$\frac{1}{2}$
Triangular	$K(u) = (1-\lvert u\rvert)\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{1}{6}$	$\frac{2}{3}$
Epanechnikov	$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{1}{5}$	$\frac{3}{5}$
Quartic (biweight)	$K(u) = \frac{15}{16}(1-u^2)^2\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{1}{7}$	$\frac{5}{7}$
Triweight	$K(u) = \frac{35}{32}(1-u^2)^3\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{1}{9}$	$\frac{350}{429}$
Tricube	$K(u) = \frac{70}{81}(1-\lvert u\rvert^3)^3\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$\frac{35}{243}$	$\frac{175}{247}$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$		1	$\frac{1}{2\sqrt{\pi}}$
Cosine	$K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\mathbf{1}_{\{\lvert u\rvert\le 1\}}$		$1-\frac{8}{\pi^2}$	$\frac{\pi^2}{16}$
Logistic	$K(u) = \frac{1}{e^u+2+e^{-u}}$		$\frac{\pi^2}{3}$	$\frac{1}{6}$



h choose

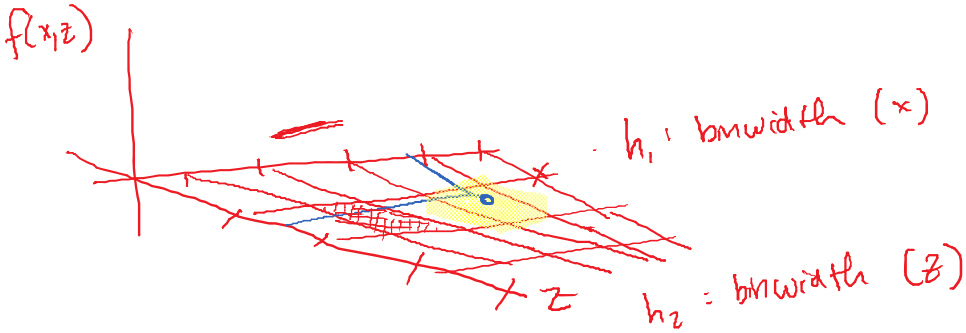
Multivariate Kernel Densities

Tuesday, January 21, 2014 5:39 PM

$$\hat{f}(x, z)$$

Histogram

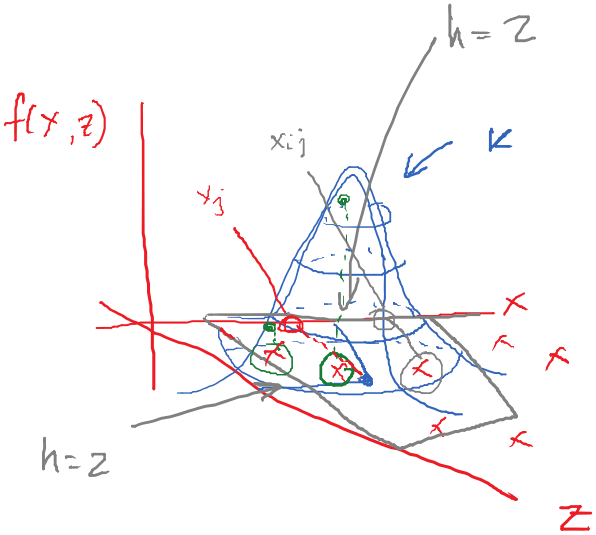
$$\hat{f}_h(x, z) = \frac{1}{N h_1 h_2} \sum_{i=1}^{N} \sum_j \underline{I\left((x_i, z_i) \in B_j\right)} \, \underline{\underline{I\left((x, z) \in B_j\right)}}$$



$$\hat{f}_h(x_1, x_2, \ldots, x_m) = \frac{1}{N} \cdot \sum_{i=1}^{N} \underbrace{\frac{1}{h^m}} * \underbrace{K\left(\frac{x_1 - x_{1i}}{h}, \frac{x_2 - x_{2i}}{h}, \ldots, \frac{x_m - x_{mi}}{h}\right)}_{\text{multivariate kernel}}$$

m: # of variables

h: bandwidth



$$K \sim \Phi\left(\begin{matrix} u_1 \\ \vdots \\ u_m \end{matrix} \,\middle|\, \mu = 0, \Sigma = I\right)$$

$$u_j = \left(\frac{\boxed{x_j} - \boxed{x_{ij}}}{\boxed{h}}\right)$$

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ & & & & 1 \end{bmatrix}$$

Z

$$I = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

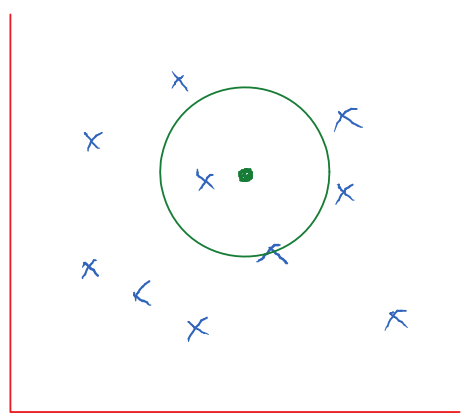
The Curse of Dimensionality

Tuesday, January 21, 2014 5:39 PM

Suppose: $N = 1000$ and I want to estimate $\hat{f}(\cdot)$

	# variable	bin cutoffs/dim	total bins	data / bin
$\hat{f}(x)$	1	10	10	$1000/10 = 100$
$\hat{f}(x,z)$	2	10	$10 \cdot 10 = 100$	$1000/100 = 10$
$\hat{f}(x,y,z)$	3	10	$10^3 = 1000$	$1000/1000 = 1$

"curse of dimensionality"



The Bias-Variance Tradeoff

Tuesday, January 21, 2014 5:40 PM

results = data + assumptions

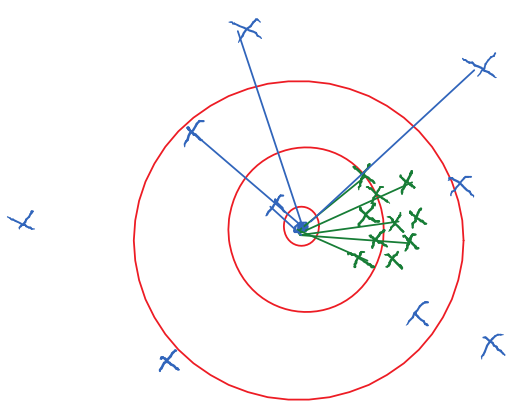
parametrics make more/stronger.

if parametric assumptions are correct : parametric model will be more precise/efficient (smaller CIs) compared to N-P model.

if parametric assumptions are NOT correct : parametric models will generally be biased, but will remain lower-variance.

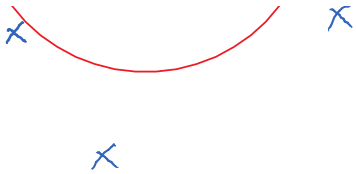
by moving from parametric to N-P models:

generally decrease bias, but increase variance, in your estimates.



there is a real tradeoff.

Sometimes a little bias

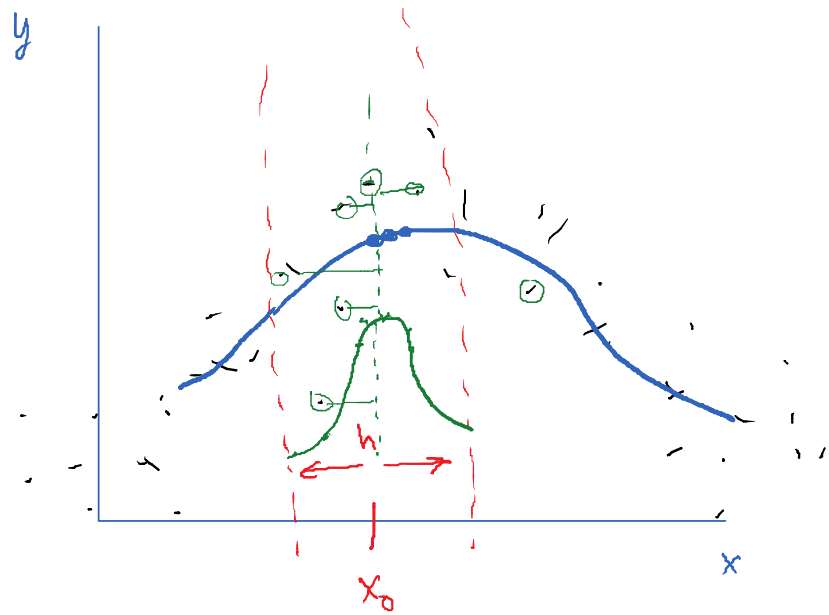


Sometimes a little bias is a good exchange for low variance.

Kernel Regression

Tuesday, January 21, 2014 5:40 PM

$E[y|x]$



Nadaraya - Watson estimator :

$$\hat{f}_n(x) = \frac{1/N \sum_{i=1}^{N} K_h(x - x_i)\, y_i}{1/N \sum_{j=1}^{N} K_h(x - x_j)} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{K_h(x - x_i)}{\frac{1}{N} \sum_{j=1}^{N} K_h(x - x_j)} \right) y_i$$

$$\hat{f}_n(x) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{W_{hi}(x)}_{\text{weight}}\, \underbrace{y_i}_{\text{dv obs}}$$

The N-W model is an optimal nonparametric estimator of $E[y|x]$.

$$y(x) = E[y|x] = \int^{\infty} y \cdot \underline{p(y|x)}\, dy$$

$p(y|x)$: pdf of $y|x$.

$$= \frac{\int y\, p(y,x)\, dy}{\int p(y,x)\, dy}$$

$$= \frac{\sum_{i=1}^{N} \int y \cdot f(x-x_i, y-y_i)\, dy}{\sum_{j=1}^{N} \int f(x-x_j, y-y_j)\, dy}$$

estimating $p(y,x) = \frac{1}{N} \cdot \sum_{i=1}^{N} f(x-x_i, y-y_i)$

kernel density estimator!

let $Y = y - y_i$.

$$= \frac{\sum_{i=1}^{N} \int (Y + y_i) \cdot f(x-x_i, Y)\, dY}{\sum_{i=1}^{N} \int f(x-x_i, Y)\, dY \rightarrow \text{den}}$$

$$= \frac{\sum_{i=1}^{N} \left[\int Y \cdot f(x-x_i, Y)\, dY + \int y_i f(x-x_i, Y)\, dY \right]}{\text{den}}$$

$= 0$

den $\sum_{i=1}^{N} y_i \int f(x-x_i, y-y_i)\, dy$

$$= \frac{\sum_{i=1}^{N} \int y_i\, f(x-x_i\ \ Y)\, dY}{\sum_{j=1}^{N} \int f(x-x_i, Y)\, dY} = \boxed{\frac{\sum_{i=1}^{N} y_i \cdot g(x-x_i)}{\sum_{j=1}^{N} g(x-x_i)}}$$

$$\underline{g(x-x_i)} = \int f(x-x_i, y)\,dy = \underline{\int f(x-x_i, y-y_i)\,dy}$$

$$E[y|x] = \frac{\sum_{i=1}^{N} y_i \cdot g(x-x_i)}{\sum_{j=1}^{N} g(x-x_i)} \rightarrow \text{N-W estimator.}$$

95% Pointwise Confidence Intervals for N-W Kernel Regression

Monday, January 20, 2014 10:00 PM

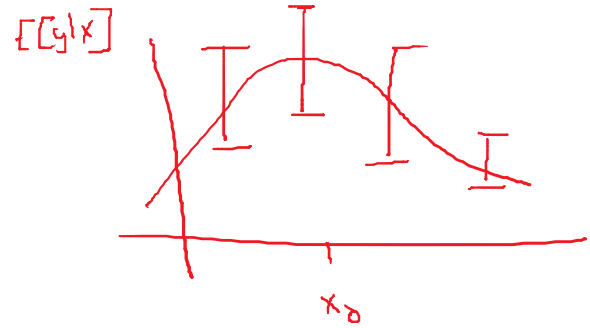
$n = N =$ sample size.

$h = \lambda =$ bandwidth

Table 3.1. *Asymptotic confidence intervals for kernel estimators – implementation.*

1. Select λ so that $n^{1/5}\lambda \to 0$, for example, $\lambda = O(n^{-1/4})$. This ensures that the bias term does not appear in (3.2.5).
2. Select a kernel K and obtain $b_K = \int K^2(u)du$. For the uniform kernel on $[-1,1]$ $b_K = 1/2$.[a]
3. Estimate f using the Nadaraya–Watson estimator (3.2.3).
4. Calculate $\hat{\sigma}_\varepsilon^2 = 1/n \sum (y_i - \hat{f}(x_i))^2$.
5. Estimate $p(x_o)$ using (3.2.9). If the uniform kernel is used, $\hat{p}(x_o)$ equals the proportion of x_i in the interval $x_o \pm \lambda$ divided by the width of the interval 2λ.
6. Calculate the confidence interval at $f(x_o)$ using $\hat{f}(x_o) \pm 1.96\sqrt{b_K \hat{\sigma}_\varepsilon^2 / \hat{p}(x_o)\lambda n}$
7. Repeat at other points if desired.

[a] For other kernels, the values of b_K are as follows: triangular, 2/3; quartic or biweight, 5/7; Epanechnikov, 3/5; triweight, 350/429; normal, $1/(2\pi^{1/2})$.



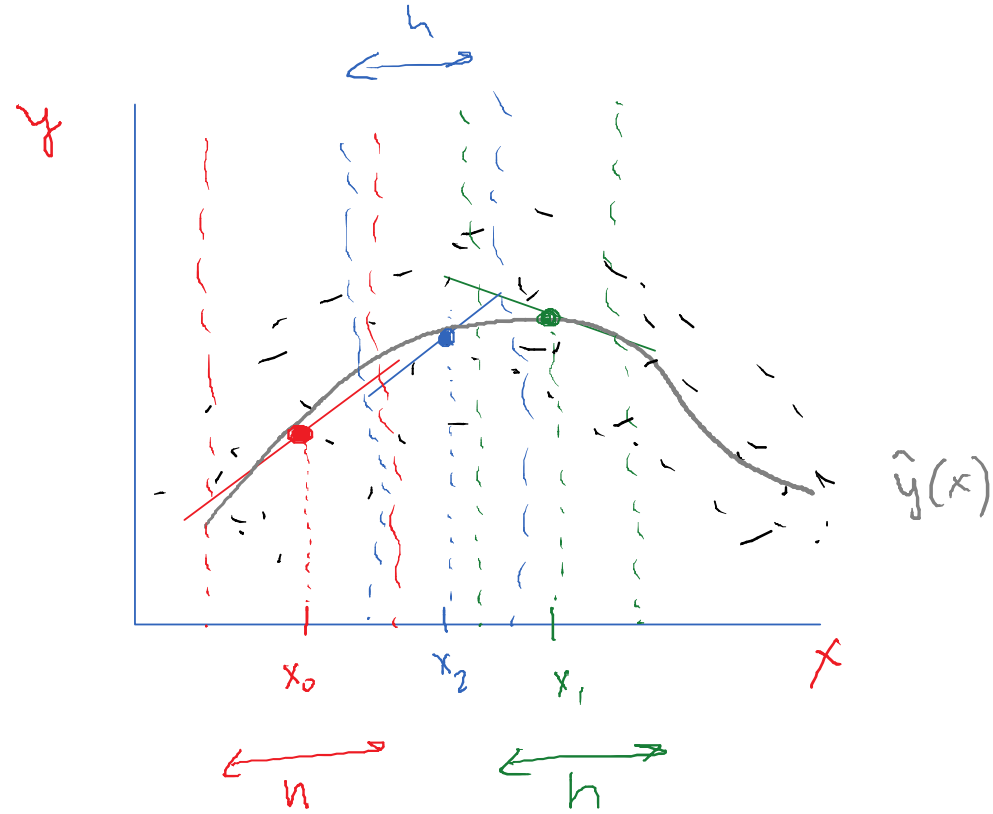
$$\frac{1}{Nh} \sum_{i=1}^{N} K_h\left[(x - x_i)/h\right]$$

st. error of $\widehat{E[y|x]} = \hat{y}(x)$.

Local Linear and Quadratic Regression

Tuesday, January 21, 2014 5:41 PM

higher-order polynomial KR.



$\Rightarrow$ solving this minimization problem: (at every point, i)

$$\min_{\beta} \sum_{i=1}^{N} \left(\beta_0 + \beta_1 (x - x_i) - y_i \right)^2 \cdot K_h (x - x_i)$$

$$= \min_{\beta} \sum_{i=1}^{N} \left(\hat{y}_i(x) - y_i \right)^2 \cdot K_h (x - x_i)$$

this is = minimizing a sum of weighted sq. errors.

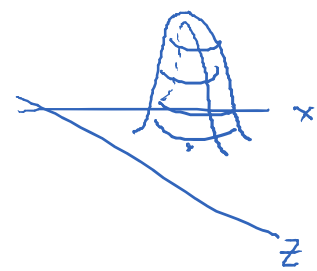
NW estimator: $\min_{\beta} \sum_{i=1}^{N} \left(\beta_0 - y_i \right)^2 \cdot K (x - x_i)$

NW estimator

$$\min_{\beta} \sum_{i=1}^{N} \left[\hat{y}_i(x) - y_i\right]^2 \cdot K_h(x - x_i)$$

local quadratic estimator:

$$\min_{\beta} \sum_{i=1}^{N} \left(\beta_0 + \beta_1 (x - x_i) + \beta_2 (x - x_i)^2 - y_i\right) \cdot K(x - x_i)$$



multivariate KR is possible.

for the N-W estimator:

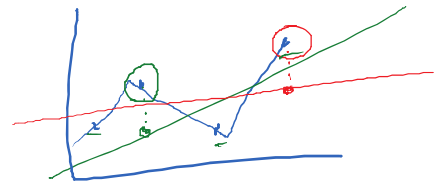
$$\hat{y}_H(x) = \frac{\sum_{i=1}^{N} K_H(\vec{x}_i - \vec{x}) y_i}{\sum_{j=1}^{N} K_H(\vec{x}_j - \vec{x})}$$

K_H is a multivariate kernel, such as MV Normal PDF.

curse of dimensionality applies.

Bandwidth Selection

Tuesday, January 21, 2014 5:42 PM



① "eyeball test"

② cross-validation

for every obs i:

$\boxed{1}$ drop i from data

$\boxed{2}$ fit model w/ data that deletes i

$\boxed{3}$ calculate $\left(\hat{y}_i^{(-i)} - y_i\right)^2$ from the model.

CV statistic: $\sum_{i=1}^{N} \left[\hat{y}_i^{(-i)} - y_i\right]^2$.

CV stat. for many values of bw (h).

③ complexity-penalized SSE.

$$\underbrace{\Xi(h)}_{\text{penalty function.}} \cdot \underbrace{\sum_{i=1}^{N} \left(\hat{y}_i^{(h)} - y_i\right)^2}_{SSE}$$

AIC / BIC smaller $h \longrightarrow$ bigger $\Xi(h)$.

$\underline{\text{AIC}}$: $\exp(2h) = \Xi(h)$.

GCV : $(1-h)^{-2} = \dfrac{1}{\sqrt{1-h}} = \Xi(h)$.