Chapter 1. Nonparametric Curve Estimation

January 18, 2007

1 Introduction

Suppose that we have two variables: predictor (independent variable) X and response (dependent variable) Y. Because of random factor, given $X = x$, what we can predict is

$$E(Y|X = x),$$

denoted by $m(x)$. In other words, we have a general model

$$Y = m(X) + \varepsilon. \tag{1.1}$$

Suppose (X, Y) has a joint distribution density function $f(x, y)$. Then

$$m(x) = E(Y|X = x) = \int y f_{Y|X}(y|x)dy = \int y \frac{f(x, y)}{f(x)} dy$$

where $f(x)$ is the marginal density function, and $f_{Y|X}(y|x)$ is the conditional density function.

Example 1.1 *Suppose*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix})$$

we have

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - c^2}} \exp\{-\frac{1}{2(1 - c^2)}[(x - a)^2 + (y - b)^2 - 2c(x - a)(y - b)]\}$$

and

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x - a)^2\}$$

Thus

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1 - c^2)}} \exp\{-\frac{1}{2(1 - c^2)}[y - b - c(x - a)]^2\}$$

The conditional mean is

$$E(Y|X = x) = b + c(x - a).$$

In other words, model (1.1) is a linear regression model

$$Y = b + c(X - a) + \varepsilon$$

Example 1.2 *Suppose* (X, Y) *is defined as above, and* $W = \exp(X)$. *Then the model between* Y *and* W *is*

$$Y = b + c(\ln(W) - a) + \varepsilon.$$

with $m(x) = b + c(\ln(x) - a)$.

In practice, the joint function is unknown, and thus $m(x)$.

Remark 1.3 *In parametric models, we need to estimate parameters in the models. This is actually to estimate the regression curve. The estimation of the parameters is a "by-product" in some sense.*

2 Basic idea of estimating $m(x)$

Suppose $(x_1, y_1), \cdots, (x_n, y_n)$ are n observations.

If for each $X = x$, we have a number of observations, say $(x, y_1), \cdots, (x, y_k)$, then $m(x)$ can be estimated as

$$\hat{m}(x) = \frac{1}{k}(y_1 + \cdots + y_k).$$

Otherwise, we consider a "neighbor" of x, say (1) $D_x = [x - b, x + b]$ for some $b > 0$ or (2) $D_x = \{x_i : x_i \text{ is one of the } k \text{ nearest observation to } x\}$, and estimate $m(x)$ by

$$\hat{m}(x) = \frac{\sum_{x_i \in D_x} y_i}{\#\{x_i \in D_x\}},$$

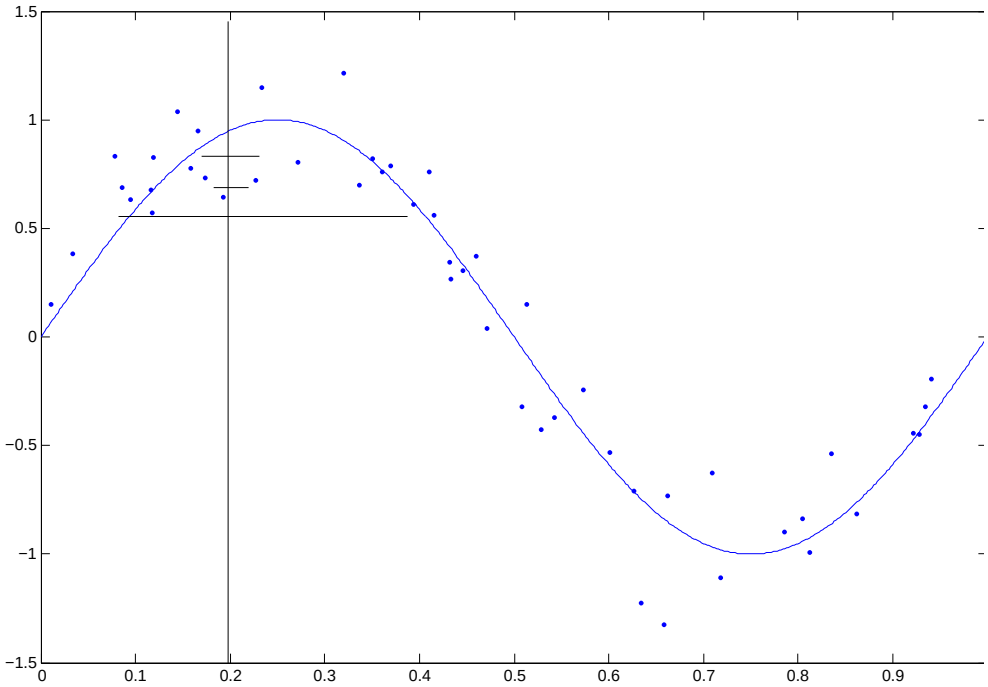
where $\#\{x_i \in D_x\}$ is the number of elements in the set.

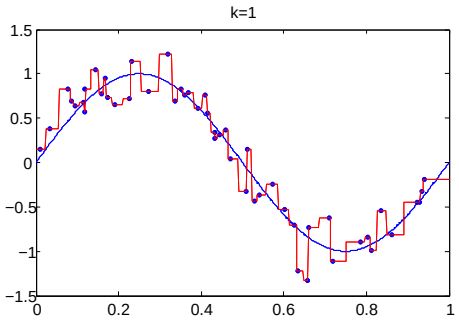
Note that the above estimator can be written as

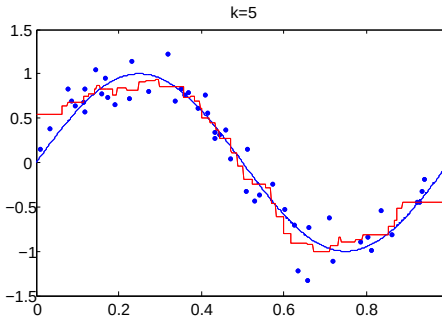
$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w_{ix} y_i}{\sum_{i=1}^{n} w_{ix}},$$

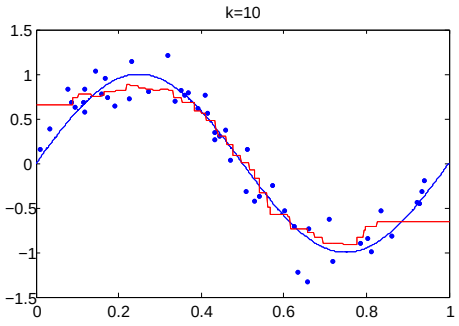
where $w_{ix} = 0$ or 1.

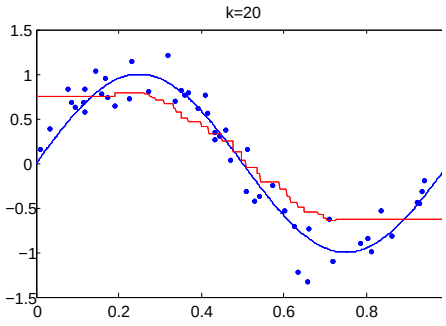
If we use (2), we call the method k-nearest neighbor estimation.


1.5
1
0.5
0
−0.5
−1
−1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


k=1
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.2
0.4
0.6
0.8
1


k=5
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.2
0.4
0.6
0.8


k=10
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.2
0.4
0.6
0.8
1


k=20
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.2
0.4
0.6
0.8

Example 2.1 *Suppose the model is*

$$Y = \sin(2\pi X) + 0.2\varepsilon$$

where $X \in [0,1]$ *and* $\varepsilon \sim N(0,1)$ *are independent. in this model*

$$m(x) = \sin(2\pi x).$$

50 observations are sampled and plotted below. with different k*, k-NN can get different estimator of the*

The main problem for k-NN is how to choose k. The role of k: *too small, the estimator is unstable (big variation); too large, the estimator is biased (cannot detect the variation of the curve)* Theoretically, the "best k" in the sense of asymptotic expansion is $k \sim n^{4/5}$.

The disadvantage is the estimation curve $\hat{m}(x)$ is not "smooth".

3 kernel smoothing

To make the estimated curve "smooth", we can replace the weight function w_{ix} by smooth function. Nadaraya (1964) and Watson (1964) proposed to use

$$w_{ix} = h^{-1} K\left(\frac{x - x_i}{h}\right)$$

The shape of the kernel is determined by function $K(x)$, called kernel function. It is convenient to write

$$K_h(x - X_i) = h^{-1} K\left(\frac{x - x_i}{h}\right).$$

The estimator is then

$$\hat{m}(x) = \sum_{i=1}^{n} K_h(x - X_i) y_i / \sum_{i=1}^{n} K_h(x - X_i),$$

which is called the N-W estimator. Here are some of the kernel functions

1. Epanechnikov kernel

$$K(x) = 0.75(1 - x^2) I(|x| \leq 1)$$

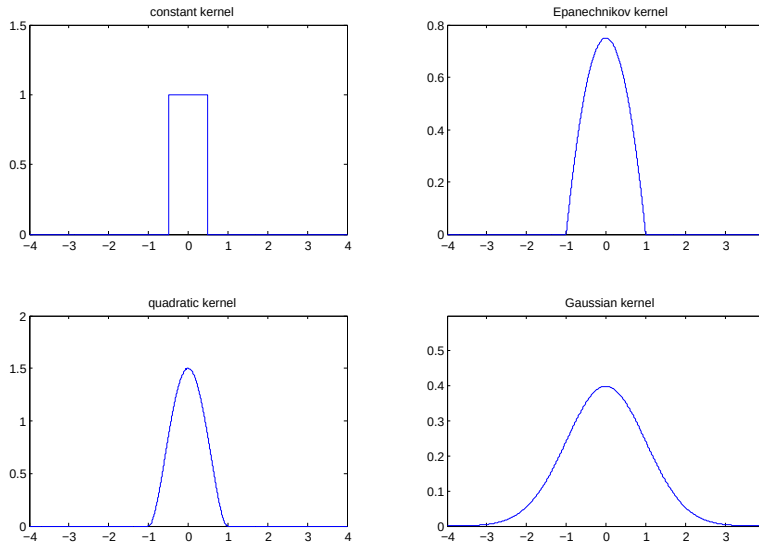
where $I(|x| \leq 1)$ is the indicator function

2. quadratic kernel

$$K(x) = 9/6(1-x^2)^2 I(|x| \le 1)$$

3. Gaussian kernel

$$K(x) = \exp(-x^2/2)/\sqrt{2\pi}.$$



The size of the "neighbor" is controlled by h, which is called *bandwidth* or "window width": *the larger h is, more observations are used to estimate the curve.* Therefore, *if h too small, the estimator is unstable (big variation); if h too large, the estimator is biased (cannot detect the variation of the curve)*

Example 3.1 *(continued) for the 50 observations above, if we use Gaussian kernel and $h = 0.05$, then the estimation kernel is*

$$\begin{aligned} \hat{m}(x) &= \sum_{i=1}^{50} \frac{1}{\sqrt{2\pi}h} \exp\{-\frac{(x-x_i)^2}{2h^2}\} y_i / \sum_{i=1}^{50} \frac{1}{\sqrt{2\pi}h} \exp\{-\frac{(x-x_i)^2}{2h^2}\} \\ &= \frac{\sum_{i=1}^{50} \exp\{-200(x-x_i)^2\} y_i}{\sum_{i=1}^{50} \exp\{-200(x-x_i)^2\}}. \end{aligned}$$

R code for the calculation **(code)**

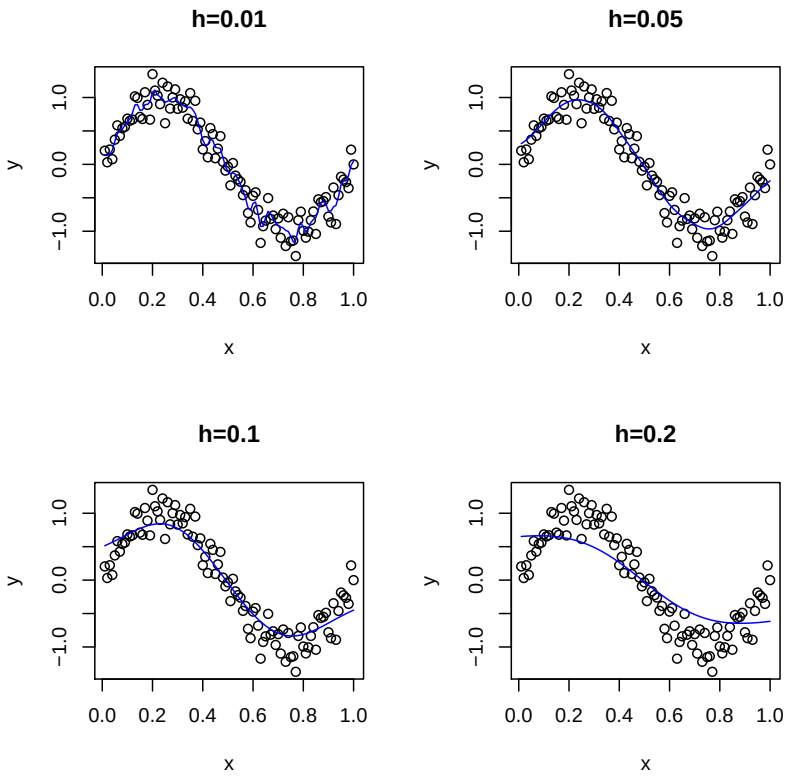


Figure 1: figure for Example 3.1

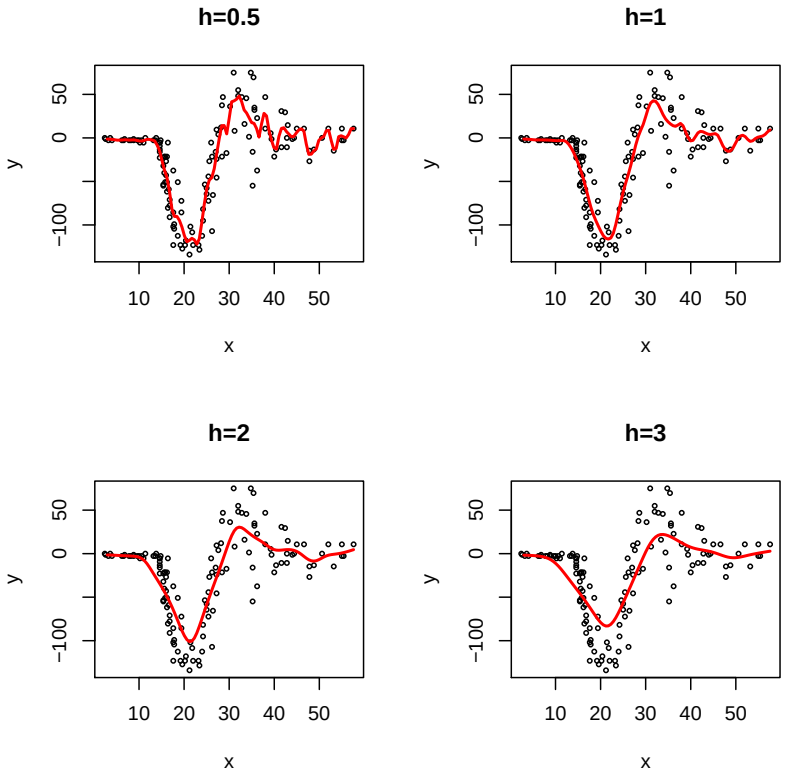


Figure 2: figure for Example 3.2

Example 3.2 (Motorcycle data) *For the motorcycle data, if we use Gaussian kernel and h, then the estimation kernel is*

$$\hat{m}(x) = \sum_{i=1}^{133} \frac{1}{\sqrt{2\pi}h} \exp\{-\frac{(x-x_i)^2}{2h^2}\} y_i / \sum_{i=1}^{133} \frac{1}{\sqrt{2\pi}h} \exp\{-\frac{(x-x_i)^2}{2h^2}\}.$$

R code for the calculation **(code)**

R package: KernSmooth

4 Another look at kernel smoothing

Suppose we are interested in $m(x)$ for a given x. The estimator $\hat{m}(x)$ should be such that

$$\sum_{i=1}^{n} w_{ix}\{y_i - m(x)\}^2.$$

where w_{ix} is the weight: the closer x_i is to x, the larger weight we assign to the difference $y_i - m(x)$.

The solution is

$$\hat{m}(x) = \sum_{i=1}^{n} w_{ix} y_i / \sum_{i=1}^{n} w_{ix}.$$

5 Statistical properties of $K_N N$ estimator

Suppose the true model is $Y = m(X) + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$. We are going to estimate $m(x)$. There are n observations. We take k observations around x, denoted by $(x_{(i)}, Y_{(i)}), i = 1, ..., k$. The estimator of $m(x)$ is

$$\hat{m}(x) = \sum_{i=1}^{k} Y_{(i)}/k = \sum_{i=1}^{k} m(x_{(i)})/k + \sum_{i=1}^{k} \varepsilon_{(i)})/k$$

We have the bias is

$$bias(\hat{m}(x)) = E\hat{m}(x) - m(x) = \sum_{i=1}^{k} m(x_{(i)})/k - m(x)$$

variance

$$var(\hat{m}(x)) = \sigma^2/k$$

When the n observation is fixed, with bigger k (bigger neighbor), we have bigger bias but smaller variance. With smaller k (bigger neighbor), we have smaller bias but bigger variance.