# Chapter 1. Nonparametric Curve Estimation Part 3

January 24, 2007

## 1 Applications to data analysis

Suppose we have two (or more) variables $X$ and $Y$. Our goal is to find the possible relation between $X$ and $Y$. In other words, we need to find whether $X$ can provide some information for (the prediction of) $Y$. Therefore, we need to check whether

$$m(x) = E(Y|X = x) \text{ is a nonconstant function of } x$$

i.e. we need to check whether there are two points, $x_0$ and $x_0'$ such that

$$m(x_0) \neq m(x_0')$$

Suppose we have observations $(X_1, Y_1), \cdots, (X_n, Y_n)$. Recall that we have

$$P(L_n(x) \leq m(x) \leq U_n(x)) \approx 0.95$$

where

$$L_n(x) = \hat{m}(x) - 1.96\Big\{\frac{d_0\hat{\sigma}^2}{n\hat{h}\hat{f}(x)}\Big\}^{1/2} \qquad \text{(lower Limit)}$$

and

$$U_n(x) = \hat{m}(x) + 1.96\Big\{\frac{d_0\hat{\sigma}^2}{n\hat{h}\hat{f}(x)}\Big\}^{1/2} \qquad \text{(upper Limit)}$$

Therefore, if

$$L_n(x_0) > U_n(x_0') \quad \text{or} \quad L_n(x_0') > U_n(x_0) \tag{1.1}$$

then, we conclude that $m(x_0) \neq m(x_0')$. If we cannot find two points such that (1.1) holds, then we conclude $m(x)$ does not depend on $x$, i.e. the expected value of $Y$ does not depend on $X$.

**Example 1.1 (Simulations)** *Consider $X$ and $Y$ are from the following model*

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

*where $X \sim Uniform(0,1)$ and $\varepsilon \sim N(0,1)$.*

1. *If $a = 0$ then the expected value of $Y$ does not depend on $X$*

2. *If $a \neq$ then the expected value of $Y$ does not depend on $X$*

*Suppose 50 observations are draw from the model with different $a$. The estimation results are shown in Fig. 1*

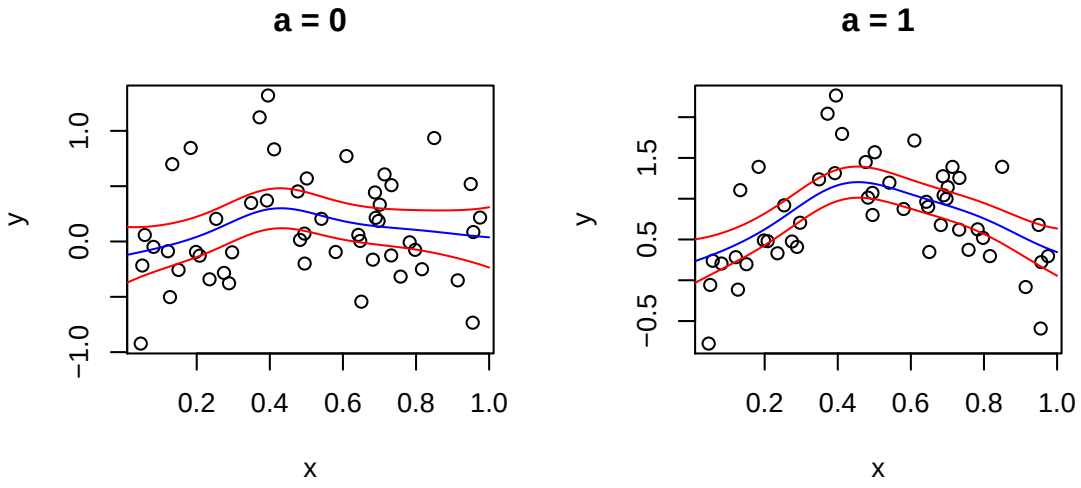


Figure 1: Calculation results for Example 1.1. The line in the central is the estimated regression function, the upper and lower lines are the 95% point-wise confidence band. The left panel is for $a = 0$ the right panel $a = 1$ **(code)**, **(code1)**, **(ks.R)**

*Therefore, in the first panel, the expected value of $Y$ does not depend on $X$, while the right panel does*

**Example 1.2 (ozone)** **(data)** *The level of ozone might be affected by radiation, temperature and wind. consider models*

$$\begin{aligned} ozone &= m_1(radiation) + \varepsilon \\ ozone &= m_2(temperature) + \xi \\ ozone &= m_3(wind) + \eta \end{aligned}$$

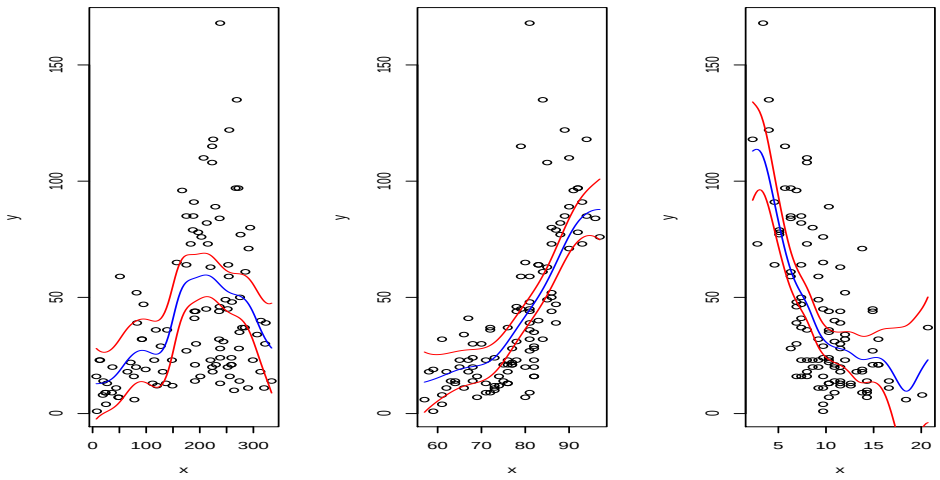


Figure 2: calculation results for example 1.2. The line in the central is the estimated regression function, the upper and lower lines are the 95% point-wise confidence band. the bandwidths are respectively 20, 4, and 1 for the estimattions **(code)**

*there are 111 observations. The estimation results are shown in Fig 2.*

*all the covariates (radiation, temperature, wind) have effects on the expected level of ozone*

## 2 bandwidth selection

**Example 2.1 (simulation)** *the plots below show the kernel smoothing estimator of the regression functions. We can see that the bandwidth in the first panel is too small generating under-smoothed curve estimator, the bandwidth in the last panel is too large generating over-smoothed estimator. The estimation results are shown in Fig 3.*

It is now clear that the selection of bandwidth plays an important role in the estimation of the regression function.

Theoretically, the optimal bandwidth should be chosen to minimize the mean squared error

$$\begin{aligned} MSE(x) = E|\hat{m}(x) - m(x)|^2 &= bias^2 + variance \\ &= c_2^2\{\frac{1}{2}m''(x) + f^{-1}(x)m'(x)f'(x)\}^2h^4 + \frac{d_0\sigma^2}{nhf(x)} \end{aligned}$$

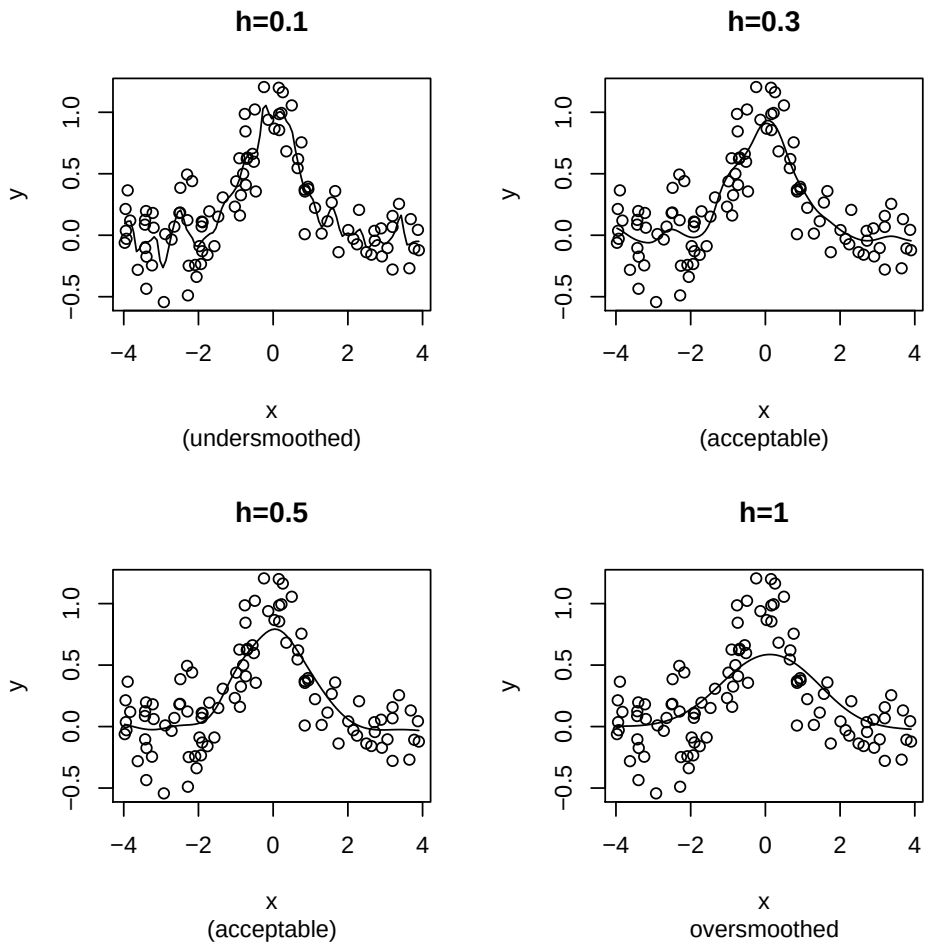


Figure 3: calculation results for example 2.1. **(code)**

The optimal bandwidth is then

$$h_{opt}(x) = \Big\{ \frac{d_0 \sigma^2}{4 f(x) c_2^2 \{\frac{1}{2} m''(x) + f^{-1}(x) m'(x) f'(x)\}^2} \Big\}^{1/5} n^{-1/5}.$$

The bandwidth differs from point to point. This bandwidth is called variable bandwidth.

To get a constant bandwidth for all the points, we need to consider an integration of the MSE, i.e. integrated mean squared error

$$IMSE(or\ MISE) = \int MSE(x) w(x) dx$$

where $w(x)$ is a weight function, which can be taken as 1. we call this bandwidth constant bandwidth.

**Example 2.2 (motorcycle)** **(data)** *the plots below show the kernel smoothing estimator of the regression functions using constant bandwidth for all points. We can see that a constant bandwidth is not suitable for the data.*

*The estimation results are shown in Fig 4.*

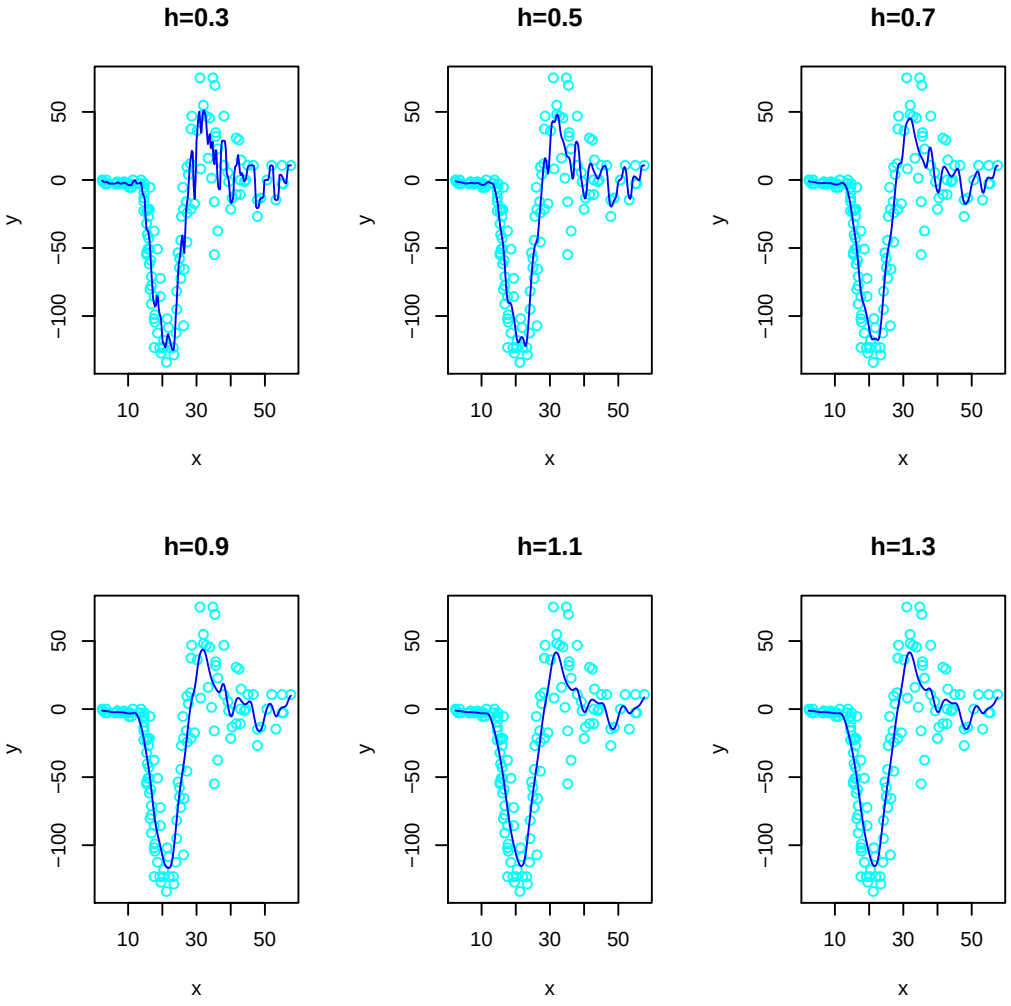


Figure 4: Calculation results for example 2.2. **(code)**

## 3 Leave-one-out cross-validation method for bandwidth selection

Note that for different bandwidth $h$ we may have different estimator of the regression function, denoted it by $\hat{m}_h(x)$. Therefore the estimated model is

$$\hat{Y} = \hat{m}_h(x)$$

The "model" is actually a function of $h$. the best $h$ should be such that the estiamted model has the best prediction capability.

Suppose we have $n$ observations $(X_1, Y_1), \cdots, (X_n, Y_n)$. Note that they have all been observed and are known to us. Pretending that one of the observation $(X_j, Y_j)$ is unknown. Thus, there are $n-1$ observations can be used. we use the n-1 observation to estimate the model with bandwidth $h$.

$$\hat{m}_{h,j}(x) = \sum_{\substack{i=1 \\ i\neq j}}^{n} K_h(X_i - x)Y_i / \sum_{\substack{i=1 \\ i\neq j}}^{n} K_h(X_i - x).$$

Now, we use the estiamted regression function $\hat{m}_{h,j}(x)$ to predict $Y_j$ as

$$\hat{Y_j} = \hat{m}_{h,j}(X_j) = \sum_{\substack{i=1 \\ i\neq j}}^{n} K_h(X_i - X_j)Y_i / \sum_{\substack{i=1 \\ i\neq j}}^{n} K_h(X_i - X_j).$$

The prediction error is then

$$\hat{Y_j} - Y_j$$

the overall prediction errors can be calculated by

$$n^{-1} \sum_{j=1}^{n} (\hat{Y_j} - Y_j)^2$$

we call it cross-validation value, i.e.

$$CV(h) = n^{-1} \sum_{j=1}^{n} (\hat{m}_{h,j}(X_j) - Y_j)^2$$

The best bandwidth in the sense of the leave-one-out cross-validation is

$$\hat{h} = \arg\min_h CV(h)$$

**Example 3.1 (simulation)** *100 observations from*

$$Y = sin(3\pi X) + 0.2\varepsilon$$

*where* $X \sim uniform(0,1)$ *and* $\varepsilon \sim N(0,1)$

*The estimation results are shown in Fig 5.*

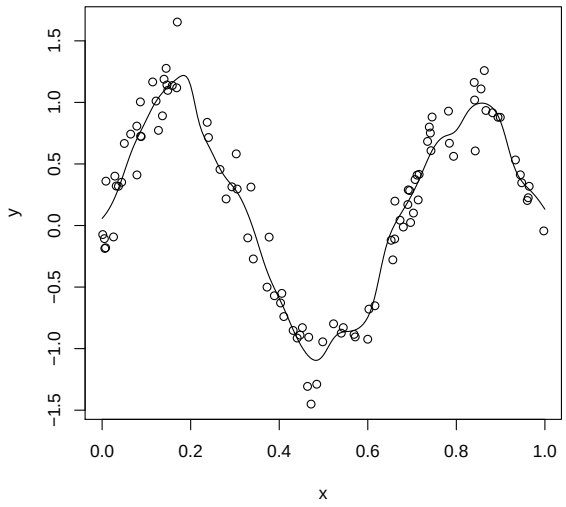


Figure 5: Calculation results for example 3.1. **(code) (cvh)**