HW5 - Lucas Fellmeth, Sven Bergmann

2023-11-16

1

The goal of this problem is to estimate the regression function of acceleration vs time for the `mcycle` data in the package `MASS`.

A

Show that the Nadaraya-Watson estimator can be expressed as $\hat{Y} = HY$. Find the "hat matrix" H explicitly.

We know that the Nadaraya-Watson estimator of $\hat{m}(x_i)$ is defined by

$$\hat{m}(x_i) = \frac{\sum_{j=1}^{n} K_h(x_j - x_i) Y_j}{\sum_{k=1}^{n} K_h(x_k - x_i)},$$

where

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$$

with h as associated bandwidth.

So

$$\begin{aligned}
\hat{m}(x_i) &= \frac{\sum_{j=1}^{n} K_h(x_j - x_i) Y_j}{\sum_{k=1}^{n} K_h(x_k - x_i)} \\
&= \frac{\sum_{j=1}^{n} \frac{1}{h} K(\frac{x_j - x_i}{h}) Y_j}{\sum_{k=1}^{n} \frac{1}{h} K(\frac{x_k - x_i}{h})} \\
&= \frac{\frac{1}{h} \sum_{j=1}^{n} K(\frac{x_j - x_i}{h}) Y_j}{\frac{1}{h} \sum_{k=1}^{n} K(\frac{x_k - x_i}{h})} \\
&= \frac{\sum_{j=1}^{n} K(\frac{x_j - x_i}{h}) Y_j}{\sum_{k=1}^{n} K(\frac{x_k - x_i}{h})} \\
&= \frac{\sum_{j=1}^{n} K(\frac{x_j - x_i}{h})}{\sum_{k=1}^{n} K(\frac{x_k - x_i}{h})} Y_j \\
&= \sum_{j=1}^{n} \underbrace{\frac{K(\frac{x_j - x_i}{h})}{\sum_{k=1}^{n} K(\frac{x_k - x_i}{h})}}_{H_{ij}} Y_j
\end{aligned}$$

$$\implies \underbrace{\begin{pmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{pmatrix}}_{\hat{Y}} = \underbrace{\begin{pmatrix} H(x_{11}) & \dots & H(x_{1n}) \\ \vdots & \ddots & \vdots \\ H(x_{n1}) & \dots & H(x_{nn}) \end{pmatrix}}_{H} \cdot \underbrace{\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}}_{Y}$$

B

For a reasonable range of bandwidths h, compute and plot the generalized cross validation measure $GCV(h)$ and find the optimal bandwidth.

```
library(MASS)
library(splines)
```

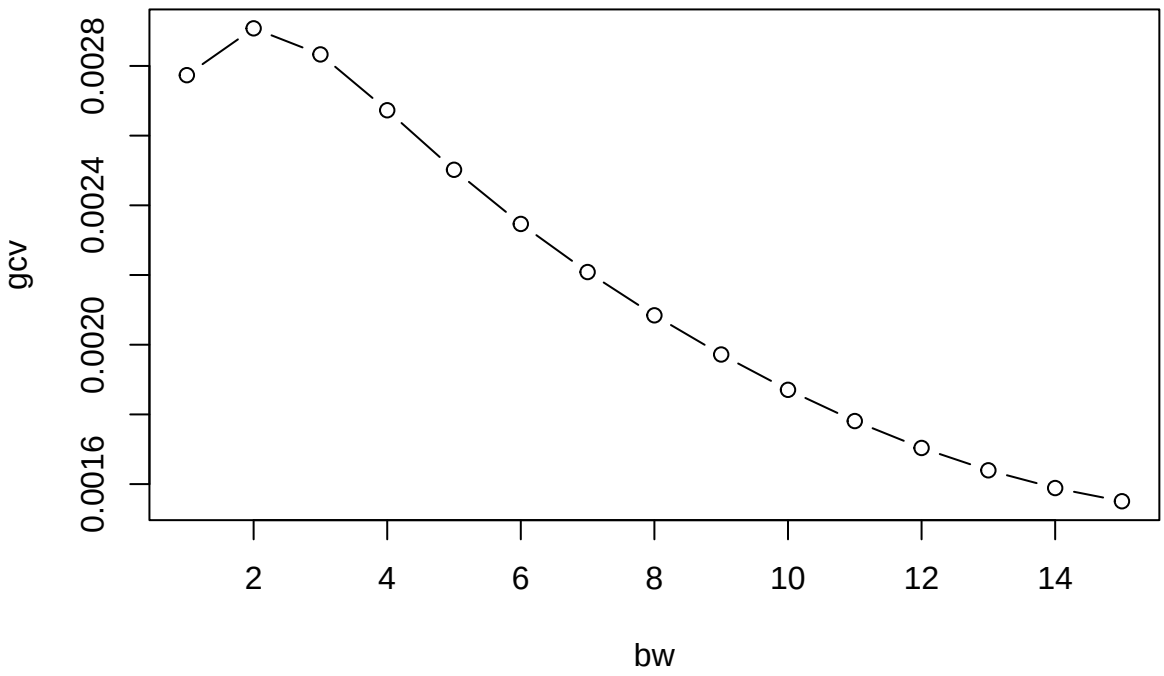
First, we implement the generalized cross validation measure $GCV(h)$ which is defined by

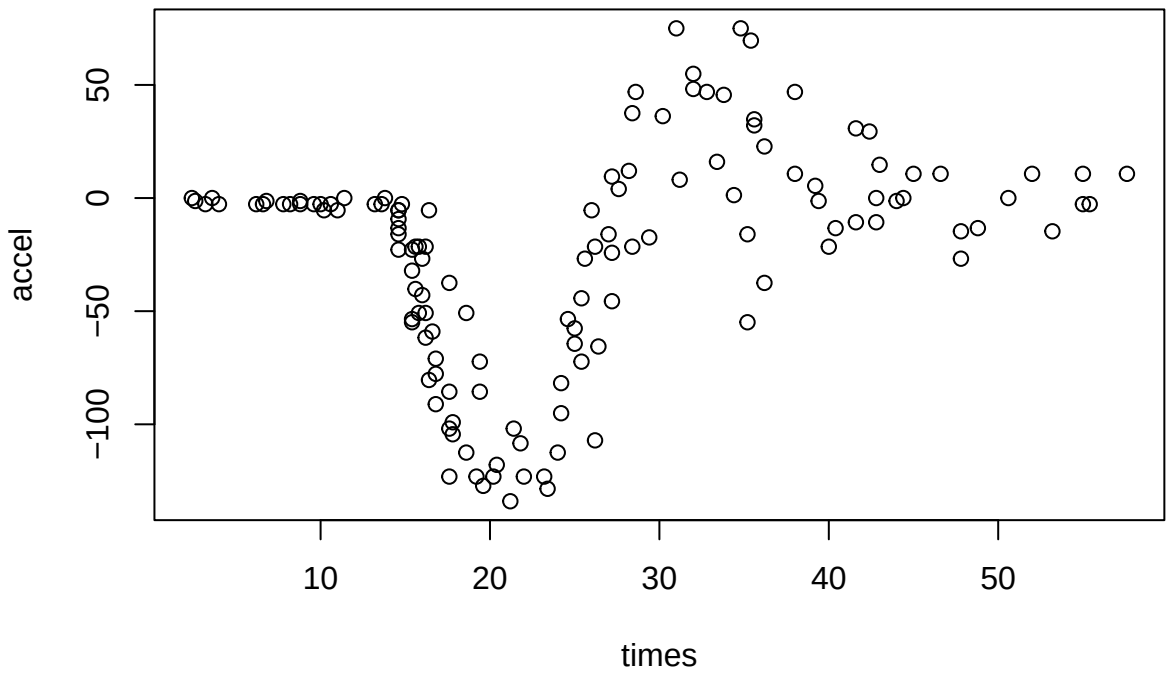
$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}_h(x)}{1 - \frac{trH(h)}{n}}\right]^2$$

where

```
GCV <- function(x, y, h, kernel = "normal") {
    y_hat <- ksmooth(x, y, kernel, bandwidth = h)$y
    H <- y_hat %*% t(y)
    H_trace <- sum(diag(H))
    n <- length(y)
    gcv <- mean((y - y_hat)^2)/(1 - (H_trace/n))^2
    return(gcv)
}
```

```
with(mcycle, {
    bw <- 1:15
    gcv <- sapply(bw, GCV, x = times, y = accel)
    plot(bw, gcv, type = "b")
    plot.new()
    plot(times, accel)
    lines(times, ksmooth(times, accel, kernel = "normal", bandwidth = min(gcv))$y,
        col = "red", type = "l")
})
```


gcv
0.0016
0.0020
0.0024
0.0028
2
4
6
8
10
12
14
bw


accel
50
0
−50
−100
10
20
30
40
50
times