Kernels

Patrick Breheny

October 20

Introduction

- The previous approach to nonparametric regression, splines, was based on expanding a continuous variable into a series of basis functions, each with localized effects, and then fitting this expanded model
- We turn our attention now to kernel methods
- Like splines, they are used to estimate $\mathrm{E}(y|x) = f(x)$ in a nonparametric way, but they do so by employing a very different strategy

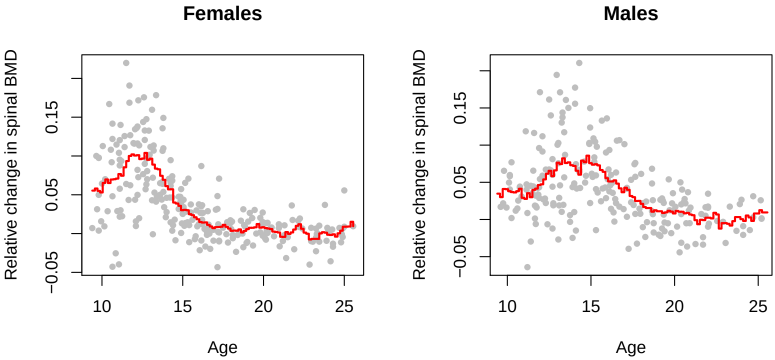
Main idea

- Rather than expand the problem into a large multivariate problem, the idea behind kernels is to fit a very simple model (*e.g.*, simple linear regression), but to fit a different simple model at every point x_0
- This is done by using only those observations close to x_0 to fit the model
- This infinite collection of simple models adds up to allow tremendous flexibility in modeling f

The local average

The simplest local model is the local average:

$$\hat{f}(x_0) = \frac{\sum_i y_i I(|x_i - x_0| < \lambda)}{\sum_i I(|x_i - x_0| < \lambda)}$$



Problems with the local average

- Note that the set of points $\{x_i\}$ within λ of x_0 changes in an abrupt fashion, leading to a discontinuous estimate of f
- This is unrealistic and unnecessary
- Rather than assign all the points in the sliding window equal weight, we can construct a weighting function that makes the transition to zero weight smoothly as x gets further away from x_0

The Nadaraya-Watson estimator

- Specifically, consider estimators of the following form, known as the *Nadaraya-Watson kernel estimator*:

$$\hat{f}(x_0) = \frac{\sum_i K_\lambda(x_0, x_i) y_i}{\sum_i K_\lambda(x_0, x_i)},$$

  where
  - The function K is called the *kernel*, and it controls the weight given to the observations $\{x_i\}$ at each point x_0 based on their proximity
  - λ, which controls the size of the neighborhood around x_0, is the smoothing parameter
- Note that if K is continuous, then so is $\hat{f}$

Kernels

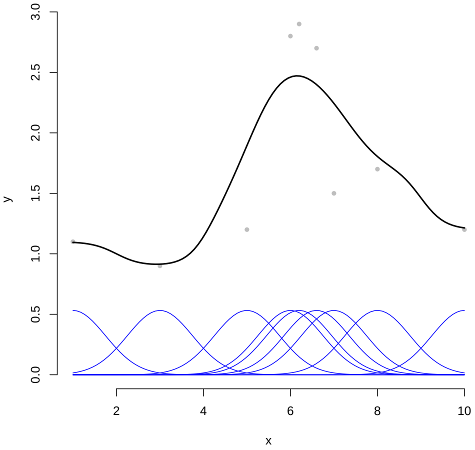
- We will consider kernels of the form

$$K_\lambda(x_i, x_0) = K\left(\frac{x_i - x_0}{\lambda}\right)$$

- To be considered a kernel, a function must also be symmetric about 0

Gaussian kernel: Illustration

An example of a kernel function is the Gaussian density



Other kernels

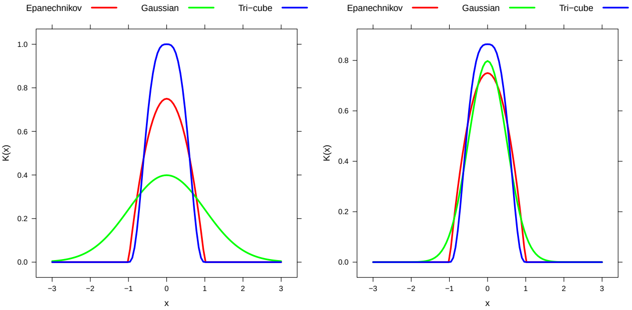
- One drawback of the Gaussian kernel is that its support runs over the entire real line; computationally it is desirable that a kernel have compact support
- Two popular compact kernels are the Epanechnikov kernel:

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & if\ |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and the tri-cube kernel:

$$K(u) = \begin{cases} (1-|u|^3)^3 & if\ |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

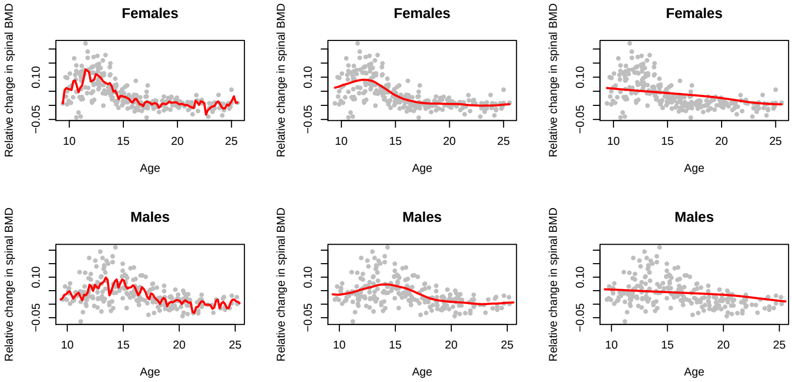
Kernels: illustration



Generally, estimates are usually quite robust to choice of kernel

Effect of changing bandwidth

Growth data, $\lambda = \{0.25, 2, 10\}$, Epanechnikov kernel:



Effect of changing bandwidth

Growth data, $\lambda = \{0.125, 1, 5\}$, Gaussian kernel: