Lecture 14 Nonparametric Regression II

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

March 11, 2021

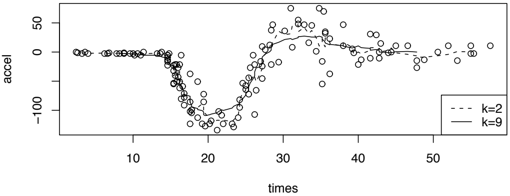
Outline

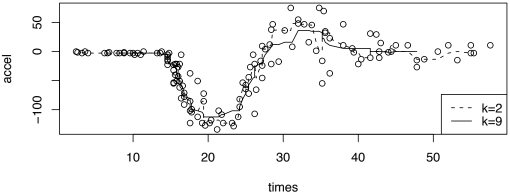
Kernel Smoothing

- When there is no parametric model relating the fitted values at one point to those at other points, it is reasonable to let the fit at x be determined by those points (x_i, y_i) with x_i close to x.
- A first attempt might be running means, in which $\hat{y}_j$ is the average of the y_i with $|i-j| \leq k$ (assuming that $\cdots x_i \leq x_{i+1} \cdots$).
- Alternatively, running medians.
- Then `plot(...,type="l")` for linear interpolation between the $(x_j, \hat{y}_j)$.

Motorcycle Data

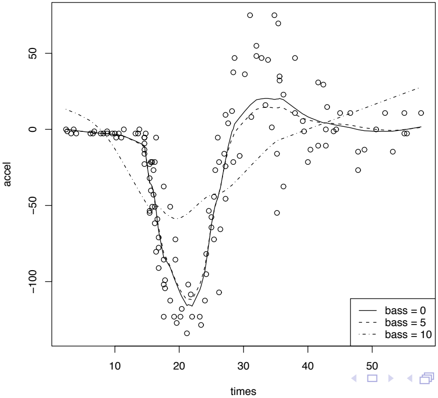
- Running means (top) and medians (bottom)





Motorcycle Data

- The super smoother function `supsmu(...)` on R will replace running means with running linear regressions - at each point (x_i, y_i), $\hat{y}_i$ is obtained by doing a linear regression using only k nearby points as data.



Kernel Smoothing

- More flexible is kernel smoothing, in which the fitted value at x is a weighted average of those values of y observed at points x_j near x:

$$\hat{y}(x) = \sum_{i=1}^{n} w\left(x - x_i\right) y_i,$$

- where $w(x - x_i)$ is typically a symmetric function, decreasing in $|x - x_i|$ and satisfying $\sum_{i=1}^{n} w\left(x - x_i\right) = 1$.
- The Nadaraya-Watson kernel uses

$$w(x - x_i) = \frac{K_\lambda\left(x - x_i\right)}{\sum_{i=1}^{n} K_\lambda\left(x - x_i\right)},$$

where $K(t)$ is a unimodal probability density, symmetric about 0, and $K_\lambda(t) = \frac{1}{\lambda} K\left(\frac{t}{\lambda}\right)$. (So $\lambda \to 0 \Rightarrow \hat{y}(x) \to ?$ (interpolation); $\lambda \to \infty \Rightarrow \hat{y}(x) \to ?$ (Mean))

Kernel Functions

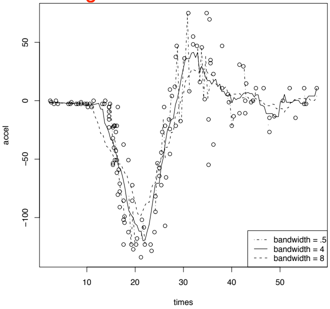
- Common choices of kernel functions:
  1. Epanechnikov kernel: $K(t) = \frac{3}{4}\left(1 - t^2\right) I(|t| \leq 1)$.
  2. Tri-cube function: $K(t) \propto \left(1 - |t|^3\right)^3 I(|t| \leq 1)$.
  3. Uniform (box in R): $K(t) = .5I(|t| \leq 1)$. - running mean?
  4. Gaussian: $K(t) = \phi(t)$.
- In R one can choose a bandwidth, which is a monotonic function of λ defined by

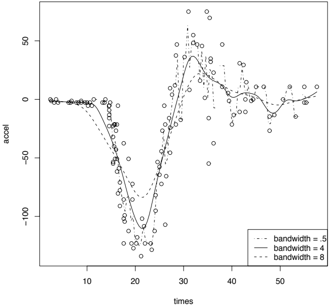
$$\text{span} = 0.75 = \int_{-\infty}^{\text{bandwidth/4}} K_\lambda(x)dx.$$

Motorcycle Data

- Kernel smooths to motorcycle data; box kernel (left) and normal kernel (right). Bandwidth = .5 is the default.

larger bandwidth —> more smooth





Kernel Smoothing

- Kernel smooths can be badly biased near the edges of the region containing the xs (since there are too few x_is on one side of x).
- Without special conditions on the design (the choice of the x_i) or on the kernel, they can be badly biased elsewhere.
- In recent years attention seems to have shifted away from kernel smoothing and towards local regression methods.

Local Regression

- Suppose we have data $(x_i, y_i = f(x_i) + \varepsilon_i)$.
- Consider estimating $f(x_0)$ by a constant $\hat{\theta}(x_0)$ defined by

new point

$$\hat{\theta}(x_0) = \arg\min_{\theta} \sum_{i=1}^{n} K_\lambda(x_0 - x_i)(y_i - \theta)^2 .$$

kernel weights

- Then

$$\hat{\theta}(x_0) = \sum_{i=1}^{n} \frac{K_\lambda(x_0 - x_i)}{\sum_{i=1}^{n} K_\lambda(x_0 - x_i)} y_i,$$

the kernel smoother.

- This is then a special case of local regression - Locally Constant.

Local Regression

- A locally linear fit is

$$\hat{f}(x_0) = \hat{\theta}_0(x_0) + \hat{\theta}_1(x_0)\,x_0 \text{ for}$$
$$\hat{\boldsymbol{\theta}}(x_0) = \arg\min_{\theta} \sum_{i=1}^{n} K_\lambda(x_0 - x_i)\left(y_i - \theta_0 - \theta_1 x_i\right)^2;$$

a locally quadratic fit includes $\theta_2 x_i^2$, etc.

- For general multiple regression with regressors $\mathbf{x}$ one solves

$$\hat{\boldsymbol{\theta}}(\mathbf{x}_0) = \arg\min_{\theta} \sum_{i=1}^{n} K_\lambda(\mathbf{x}_0, \mathbf{x}_i)\left(y_i - \left(1, \mathbf{x}_i^T\right)\theta\right)^2$$

multiple regression locally linear fit

and sets $\hat{f}(\mathbf{x}_0) = \left(1, \mathbf{x}_0^T\right)\hat{\boldsymbol{\theta}}(\mathbf{x}_0)$;

- $K_\lambda(\mathbf{x}_0, \mathbf{x}_i)$ is typically radially symmetric, i.e. a function of $\|\mathbf{x}_0 - \mathbf{x}_i\|$ such as $\frac{1}{\lambda}\phi\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_i\|}{\lambda}\right)$.

Local Regression

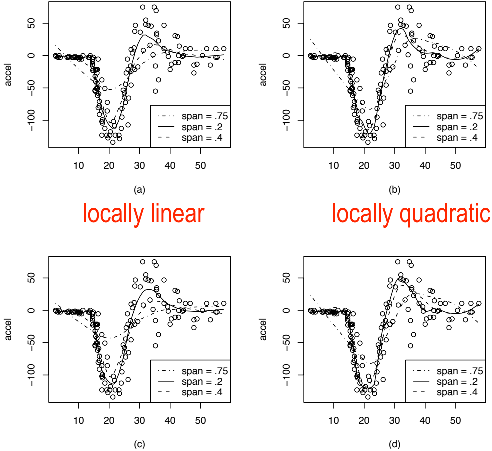
- An advantage of local regression estimators over kernel smoothing is in bias reduction.
- The variance ($\text{VAR}[\hat{f}(x_0)]$) increases as more terms are added; there is a trade-off between bias and variance.
- A (possibly) robust version of local polynomial regression (for $r = 0, 1, 2$) is incorporated in R, as the function `loess(...)`. Important options are

  how wide the kernel is

  1. `span` - related to λ; the default of .75 often gives too much smoothness. A bigger span leads to smoothness.
  2. `family` - gaussian for least squares fitting, symmetric for fitting using a redescending M-estimate.

Loess fits

- Locally linear; gaussian family (a) and symmetric family (c). Locally quadratic; gaussian family (b) and symmetric family (d).



Summary and Remark

- Kernel Smoothing
- Local Regression
- Read textbook Chapter 6 and `R` code
- Do `R` lab

For Lecture 13, 14, it's important to understand it qualitatively.