Module 2: Splines and Kernel Methods

Local Polynomial Reg., Kernel Density Estimation

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 18th, 2013

Motivating Kernel Methods

- Recall original goal from Lecture 1:
  - We don't actually know the data-generating mechanism
  - Need an estimator $\hat{f}_n(\cdot)$ based on a random sample $Y_1, \ldots, Y_n$, also known as ***training data***
- Proposed a simple model as estimator of $E[\,Y \mid X\,]$

$$\hat{f}(x) = \text{Avg}(y_i \mid x_i \in \text{Nbhd}(x))$$

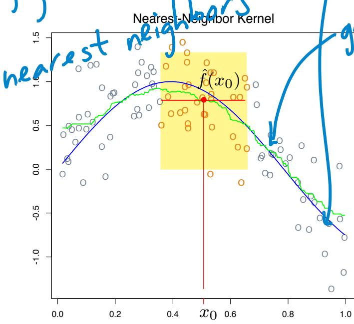
↑ use all obs. y_i in a neighborhood of target x

Choice #1: k Nearest Neighbors

- Define nbhd of each data point x_i by the k nearest neighbors
  - Search for k closest observations and average these

$\hat{f}(x) = \text{Avg}(y_i \mid x_i \in N_k(x))$ ↑ k-nearest neighbors

blue = true
green = est.



From Hastie, Tibshirani, Friedman book

- Discontinuity is unappealing

neighbors are either in or out → disc.

Choice #2: Local Averages

- A simpler choice examines a fixed distance h around each x_i
  - Define set: $B_x = \{i : |x_i - x| \leq h\}$
  - # of x_i in set: n_x

$\hat{f}(x) = \frac{1}{n_x} \sum_{i \in B_x} y_i$ avg. obs. within distance h

- Results in a linear smoother

$\hat{f}(x) = \sum_{i=1}^{n} l_i(x) y_i$ $l_i(x) = \begin{cases} \frac{1}{n_x} & \text{if } |x_i - x| \leq h \\ 0 & \text{ow} \end{cases}$

- For example, with $x_i = \frac{i}{9}$ and $h = \frac{1}{9}$

$$L = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & \cdots \\ 1/3 & 1/3 & 1/3 & 0 & \cdots \\ 0 & 1/3 & 1/3 & 1/3 & \\ & \ddots & & & \end{bmatrix}$$

More General Forms

- Instead of weighting all points equally, slowly add some in and let others gradually die off
- ***Nadaraya-Watson kernel weighted average***

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

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

kernel (K); bandwidth (λ)

- But what is a ***kernel*** ???

©Emily Fox 2013

5

Kernels

- Could spend an entire quarter (or more!) just on kernels
- Will see them again in the Bayesian nonparametrics portion
- For now, the following definition suffices

$K(\cdot)$ is a kernel if

$K(x) \geq 0 \quad \forall x$

$\int K(u)\,du = 1$

$\int u K(u)\,du = 0$

$\sigma_K^2 = \int u^2 K(u)\,du < \infty$

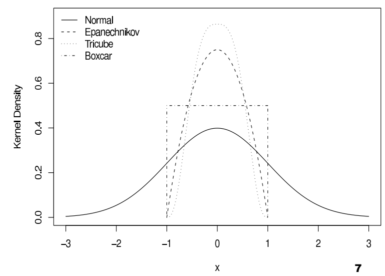
©Emily Fox 2013

6

Example Kernels

- *Gaussian* $$K(x) = \frac{1}{2\pi} e^{-\frac{x}{2}}$$
- *Epanechnikov* $$K(x) = \frac{3}{4}(1-x)^2 I(x)$$
- *Tricube* $$K(x) = \frac{70}{81}(1-|x|^3)^3 I(x)$$
- *Boxcar* $$K(x) = \frac{1}{2} I(x)$$

ind. on -1,1



Nadaraya-Watson Estimator

- Return to Nadaraya-Watson kernel weighted average

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

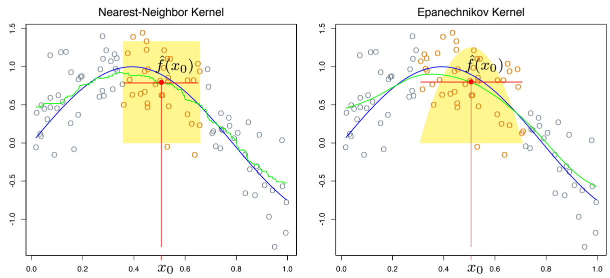
- Linear smoother:

Nadaraya-Watson Estimator

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

- Example:
  - Boxcar kernel →
  - Epanechnikov
  - Gaussian

- Often, choice of kernel matters much less than choice of λ

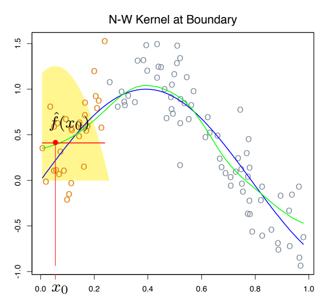


From Hastie, Tibshirani, Friedman book

Local Linear Regression

- Locally weighted averages can be badly biased at the boundaries because of asymmetries in the kernel

- Reinterpretation:



From Hastie, Tibshirani, Friedman book

- Equivalent to the Nadaraya-Watson estimator
- Locally constant estimator obtained from weighted least squares

Local Linear Regression

- Consider locally weighted linear regression instead
- Local linear model around fixed target x_0 :
- Minimize:
- Return:
- Fit a new local polynomial for *every* target x_0

Local Linear Regression

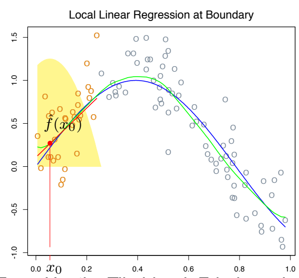
$$\min_{\beta_{x_0}} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \beta_{0x_0} - \beta_{1x_0}(x_i - x_0))^2$$

- Equivalently, minimize
- Solution:

Local Linear Regression

- Bias calculation:

$$E[\hat{f}(x_0)] = \sum_i \ell_i(x_0) f(x_i)$$



From Hastie, Tibshirani, Friedman book

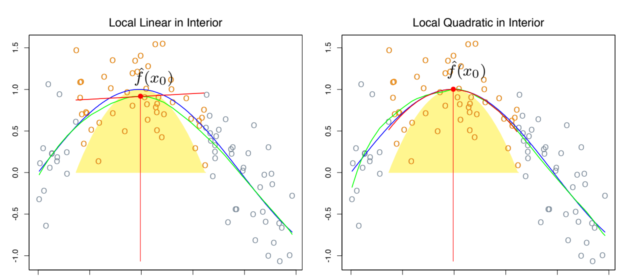
- Bias $E[\hat{f}(x_0)] - f(x_0)$ only depends on quadratic and higher order terms

- Local linear regression corrects bias exactly to 1st order

Local Polynomial Regression

- Local linear regression is biased in regions of curvature
  - "Trimming the hills" and "filling the valleys"

- Local quadratics tend to eliminate this bias, but at the cost of increased variance



From Hastie, Tibshirani, Friedman book

Local Polynomial Regression

- Consider local polynomial of degree *d* centered about x_0

  $P_{x_0}(x;\beta_{x_0}) =$

- Minimize: $\min_{\beta_{x_0}} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - P_{x_0}(x;\beta_{x_0}))^2$
- Equivalently:
- Return:
- Bias only has components of degree *d+1* and higher

Local Polynomial Regression

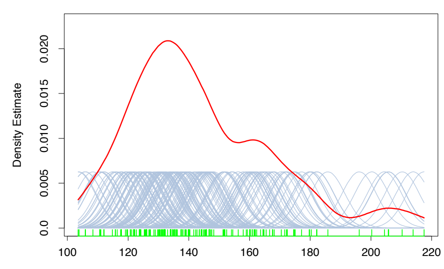
- Rules of thumb:
  - Local linear fit helps at boundaries with minimum increase in variance
  - Local quadratic fit doesn't help at boundaries and increases variance
  - Local quadratic fit helps most for capturing curvature in the interior
  - Asymptotic analysis →
    local polynomials of odd degree dominate those of even degree
    (MSE dominated by boundary effects)
  - Recommended default choice: **local linear regression**

Kernel Density Estimation

- Kernel methods are often used for density estimation (actually, classical origin)
- Assume random sample
- Choice #1: empirical estimate?
- Choice #2: as before, maybe we should use an estimator
- Choice #3: again, consider kernel weightings instead

Kernel Density Estimation

- Popular choice = Gaussian kernel → ***Gaussian KDE***



From Hastie, Tibshirani, Friedman book

KDE Properties

$$\hat{p}^{\lambda}(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

- Let's examine the bias of the KDE

$$E[\hat{p}^{\lambda}(x)] =$$

- Smoothing leads to biased estimator with mean a smoother version of the true density
- For kernel estimate to concentrate about *x* and bias→0, want

KDE Properties

$$\hat{p}^{\lambda}(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

- Assuming smoothness properties of the target distribution, it's straightforward to show that

$$E[\hat{p}^{\lambda}(x)] =$$

  - In peaks, negative bias and KDE underestimates *p*
  - In troughs, positive bias and KDE over estimates *p*
  - Again, "trimming the hills" and "filling the valleys"
- For var→0, require
- More details, including IMSE, in Wakefield book
- Fun fact: There does not exist an estimator that converges faster than KDE assuming only existence of p''

Connecting KDE and N-W Est.

- Recall task:

$$f(x) = E[Y \mid x] = \int yp(y \mid x)dy$$

- Estimate joint density *p(x,y)* with product kernel

$$\hat{p}^{\lambda_x,\lambda_y}(x,y) =$$

- Estimate margin *p(y)* by

$$\hat{p}^{\lambda_x}(x) =$$

Connecting KDE and N-W Est.

- Then,

$$\hat{f}(x) =$$

- Equivalent to Naradaya-Watson weighted average estimator

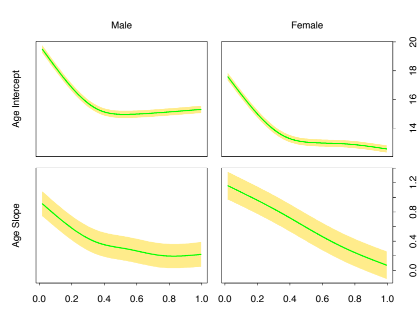
Module 2: Splines and Kernel Methods

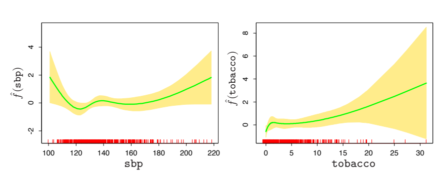
Inference for Linear Smoothers

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 18th, 2013

Confidence Bands

- So far we have focused on point estimation: $\hat{f}(x)$
- Often, we want to define a ***confidence interval*** for which $f(x)$ is in this interval with some pre-specified probability
- Looking over all *x*, we refer to these as ***confidence bands***





From Hastie, Tibshirani, Friedman book

Bias Problem

- Typically, these are of the form

$$\hat{f}(x) \pm c\ \mathrm{se}(x)$$

- This is really not a confidence band for $f(x)$, but for

$$\bar{f}(x) = E[\hat{f}(x)]$$

- In parametric inference, these are normally equivalent
- More generally,

$$\frac{\hat{f}(x) - f(x)}{s(x)} =$$

Bias Problem

$$\frac{\hat{f}(x) - f(x)}{s(x)} = Z_n(x) + \frac{\mathrm{bias}(\hat{f}(x))}{\sqrt{\mathrm{var}(\hat{f}(x))}}$$

- Typically, $Z_n(x)$ → standard normal
- In parametric inference, 2nd term normally → 0 as *n* increases
- In nonparametric settings,
  - optimal smoothing = balance between bias and variance
  - 2nd term does *not* vanish, even with large *n*
- So, what should we do?
  - Option #1: Estimate the bias
  - Option #2: Live with it and just be clear that the CI's are for $\bar{f}(x)$ not $f(x)$

CIs for Linear Smoothers

- For linear smoothers, and assuming constant variance

$$\hat{f}(x) = \sum_{i=1}^{n} \ell_i(x) y_i$$

- Consider confidence band of the form

- Using this, let's solve for *c*

CIs for Linear Smoothers

- Based on approach of Sun and Loader (1994)
  - Case #1: Assume σ known

$$P(\bar{f}(x) \notin \text{CI}(x) \text{ for some } x \in [a,b]) =$$

$$W(x) = \sum_i Z_i T_i(x) \quad Z_i = \frac{\epsilon_i}{\sigma} \sim N(0,1) \quad T_i(x) = \frac{\ell_i(x)}{||\ell(x)||}$$

- Good news: max of GP is well studied!

$$P(\max_x |\sum_i Z_i T_i(x)| > c) \approx 2(1 - \phi(c)) + \frac{\kappa_0}{\pi} e^{\frac{-c^2}{2}}$$

- Assuming confidence level α, set equal to α and solve for *c*

CIs for Linear Smoothers

- Based on approach of Sun and Loader (1994)
  - Case #2: Assume σ unknown
  - Case #3: Assume $\sigma(x)$ non-constant

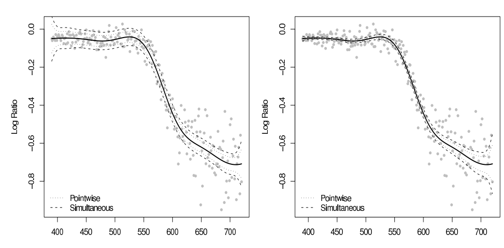
$$\mathrm{var}(\hat{f}(x)) =$$

$$\mathrm{CI}(x) =$$

  - If $\hat{\sigma}(x)$ varies slowly with *x*, then (Faraway and Sun 1995)

CIs for Linear Smoothers

- Example from Wakefield textbook
  - Fit penalized cubic regression spline (penalty on trunc. power basis coef.)
  - For $\alpha = 0.05$, we calculate $c \approx 3.11$
  - Estimate both constant and non-constant variance



- Notes: Ignored uncertainty introduced by choice of λ
  - Restrict search to finite set and do Bonferroni correction
  - Sophisticated bootstrap techniques
  - Bayesian approach treats λ as a parameter with a prior and averages over uncertainty in λ for subsequent inferences

Variance Estimation

- In most cases σ is unknown and must be estimated
- For linear smoothers, consider the following estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{f}(x_i))^2}{n - 2\nu + \tilde{\nu}}$$

  - If target function is sufficiently smooth, $\nu = o(n), \tilde{\nu} = o(n)$
  - Then $\hat{\sigma}^2$ is a consistent estimator of σ^2

Variance Estimation

- Proof outline:
  - Recall that

$$Y - \hat{f} =$$

and

$$E[Y^T Q Y] = \text{tr}(QV) + \mu^T Q \mu$$

  - Then,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{f}(x_i))^2}{n - 2\nu + \tilde{\nu}}$$

$$E[\hat{\sigma}^2] =$$

  - Therefore, bias→0 for large *n* if *f* is smooth.
  - Likewise for variance.

Alternative Estimator

- Estimator:

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i-1}^{n-1} (y_{i+1} - y_i)^2$$

- Motivation:

$$y_{i+1} - y_i =$$

$$E[(y_{i+1} - y_i)^2] \approx$$

- Estimator will be inflated
- Other estimators exist, too. See Wakefield or Wasserman.

Heteroscedasticity

- The point estimate $\hat{f}(x)$ is relatively insensitive to heterosced., but confidence bands need to account for non-constant variance
- Re-examine model $y_i = f(x_i) + \sigma(x_i)\epsilon_i$
  - Define
    $$Z_i = \log(y_i - f(x_i))^2 \quad \delta_i = \log \epsilon_i^2$$
  - Then,
- Algorithm:
  1. Estimate $f(x)$ using a nonparametric method w/ constant var to get $\hat{f}(x)$
  2. Define $Z_i = \log(y_i - \hat{f}(x_i))^2$
  3. Regress Z_i's on x_i's to get estimate $\hat{g}(x)$ of $\log \sigma^2(x)$

Heteroscedasticity

- Drawbacks:
  - Taking log of a very small residual leads to a large outlier
  - A more statistically rigorous approach is to jointly estimate *f, g*

- Alternative = Generalized linear models