Module 2: Splines and Kernel Methods

Kernel Density Estimation

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 17th, 2014

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

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

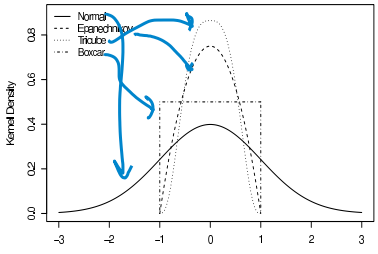
$\int K(u)\,du = 1$

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

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

Example Kernels

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



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:

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

$\hat{f} = L_\lambda y$

$\nu_\lambda = \text{tr}(L_\lambda)$

Local Polynomial Regression

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

$$P_{x_0}(x;\beta_{x_0}) = \beta_{0x_0} + \beta_{1x_0}(x-x_0) + \frac{\beta_{2x_0}}{2!}(x-x_0)^2 + \ldots + \frac{\beta_{dx_0}}{d!}(x-x_0)^d$$

- 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:

$$\min (y - X_{x_0}\beta_{x_0})^T W_{x_0} (y - X_{x_0}\beta_{x_0})$$

$$X_{x_0} = \begin{bmatrix} 1 & x_1 - x_0 & \cdots & \frac{(x_1-x_0)^d}{d!} \\ \vdots & & & \vdots \\ 1 & x_n - x_0 & & \frac{(x_n-x_0)^d}{d!} \end{bmatrix}$$

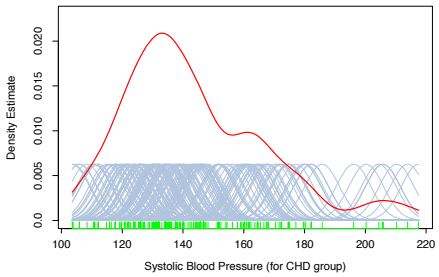
- Return: $\hat{f}(x_0) = \hat{\beta}_{0x_0}$
- Bias only has components of degree *d+1* and higher

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:

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

- 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

Reading

- Hastie, Tibshirani, Friedman: 6.1-6.2, 6.6
- Wakefield: 11.3

What you should know…

- Definition of a kernel and examples
- Nearest neighbors vs. local averages
- Nadarya-Watson estimation
  - Interpretation as local linear regression
- Local polynomial regression
  - Definition
  - Properties/ rules of thumb
- Kernel density estimation
  - Definition
  - Properties
  - Relationship to Nadarya-Watson estimation

Module 2: Splines and Kernel Methods

Inference for Linear Smoothers

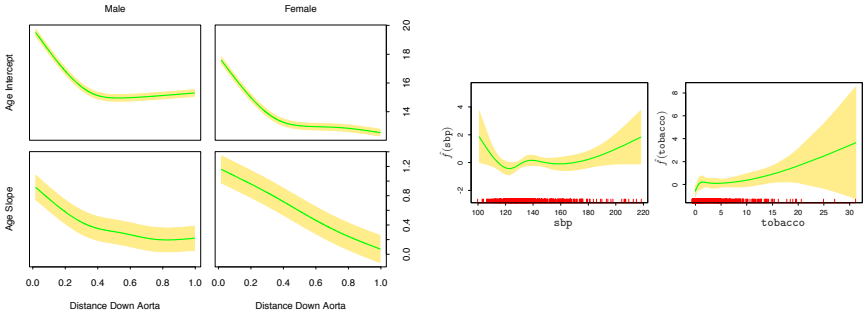
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 17th, 2014

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

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

- 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

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

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

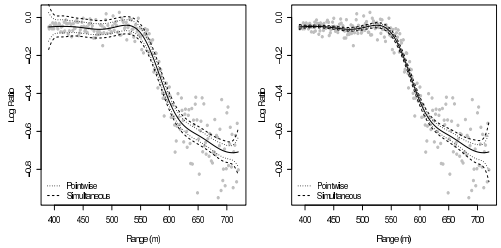
$$\text{var}(\hat{f}(x)) =$$

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

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

Reading

- Wasserman: 5.6-5.7
- Wakefield: 11.2.7, 11.4

What you should know…

- Concept of confidence band for nonparametric inference
  - Confidence band for *mean* of estimator of f(x): $\bar{f}(x) = E[\hat{f}(x)]$
- Confidence bands for linear smoothers under assumption of
  - Homoscedasticity
    - Treating variance as known
    - Treating variance as unknown
  - Heteroscedasticity
- Variance estimators for linear smoothers
  - Homoscedastic: 2 estimators
  - Heteroscedastic: via transformations