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

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

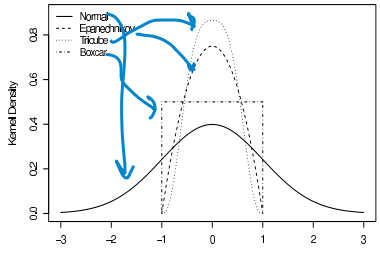
$\int K(u)\,du = 1$

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

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

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:

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

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

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

equates to fitting locally constant models with weights given by kernel

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:

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

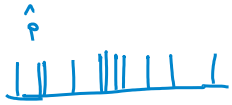
$$X_{x_0} = \begin{bmatrix} 1 & x_1-x_0 & \cdots & \frac{(x_1-x_0)^d}{d!} \\ \vdots & & & \vdots \\ 1 & x_n-x_0 & & \frac{(x_n-x_0)^d}{d!} \end{bmatrix}$$

weighted least squares for each x_0

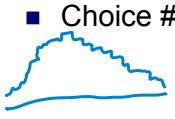
- Return: $\hat{f}(x_0) = \hat{\beta}_{0x_0}$
- Bias only has components of degree *d+1* and higher

Kernel Density Estimation

- Kernel methods are often used for density estimation (actually, classical origin)
- Assume random sample $x_1, \ldots, x_n \overset{iid}{\sim} p$
- Choice #1: empirical estimate? $\hat{p} = \frac{1}{n}\sum \delta_{x_i}$



- Choice #2: as before, maybe we should use an estimator



$$\hat{p}(x_0) = \frac{\#x_i \in \text{Nbhd}(x_0)}{n\lambda}$$

width nbhd

- Choice #3: again, consider kernel weightings instead

$$\hat{p}(x_0) = \frac{1}{n\lambda}\sum K_\lambda(x_0, x_i)$$

Parzen est.

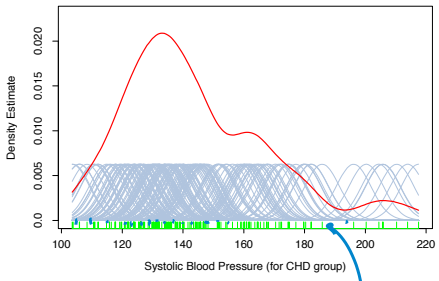
Kernel Density Estimation

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

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

$$= (\hat{p} * \phi_\lambda)(x)$$

empirical dist.



From Hastie, Tibshirani, Friedman book

$\hat{p}$: green = empirical est.

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

$x_1, \ldots, x_n \sim p$

$$E[\hat{p}^\lambda(x)] = \frac{1}{n\lambda}\sum_{i=1}^{n} E\left[K\left(\frac{x - x_i}{\lambda}\right)\right] = \frac{1}{n\lambda}\sum_{i=1}^{n}\int K\left(\frac{x-t}{\lambda}\right)p(t)dt$$

$$= \frac{1}{\lambda}\int K\left(\frac{x-t}{\lambda}\right)p(t)dt = (\lambda^{-1}K_\lambda * p)(x)$$

true density

- 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

$$\lambda \to 0 \quad \text{as} \quad n \to \infty$$

"λ_n"

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

$$E[\hat{p}^{\lambda}(x)] = p(x) + \frac{1}{2}\lambda_n^2 p''(x)\sigma_K^2 + O(\lambda_n^2)$$

$p''(x)$ abs. cont.

as $n \to \infty$ if $\lambda_n \to 0$, then this $\to 0$

asy. unbiased

  - 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 $n\lambda_n \to \infty$
- More details, including IMSE, in Wakefield book

$O(n^{-4/5})$

- Fun fact: There does not exist an estimator that converges faster than KDE assuming only existence of p'' (smoothness of target density)

Connecting KDE and N-W Est.

- Recall task:

$$f(x) = E[Y \mid x] = \int y p(y \mid x) dy = \frac{1}{p(x)} \int y\, p(x, y)\, dy$$

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

$$\hat{p}^{\lambda_x, \lambda_y}(x, y) = \frac{1}{n\lambda_x\lambda_y} \sum_{i=1}^{n} K_x\left(\frac{x - x_i}{\lambda_x}\right) K_y\left(\frac{y - y_i}{\lambda_y}\right)$$

- Estimate margin *p(y)* by

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

Connecting KDE and N-W Est.

- Then,

$$\hat{f}(x) = \frac{\frac{1}{n\lambda_x\lambda_y}\sum \int y\, K_x\left(\frac{x-x_i}{\lambda_x}\right) K_y\left(\frac{y-y_i}{\lambda_y}\right) dy}{\frac{1}{n\lambda_x}\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{\lambda_x}\right)}$$

$$= \frac{\sum K_x(\cdot) \int (y_i + u\lambda_y) K_y(u)\, du}{\sum K_x(\cdot)}$$

use $\int u K(u)\, du = 0$, $\int K(u)\, du = 1$

$$= \frac{\sum K_x\left(\frac{x-x_i}{\lambda_x}\right) y_i}{\sum K_x\left(\frac{x-x_i}{\lambda_x}\right)}$$

- 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~~ const. regression
- Local polynomial regression
  - Definition
  - Properties/ rules of thumb
- Kernel density estimation
  - Definition
  - Properties
  - Relationship to Nadarya-Watson estimation

focus this lecture

Module 2: Splines and Kernel Methods

Inference for Linear Smoothers

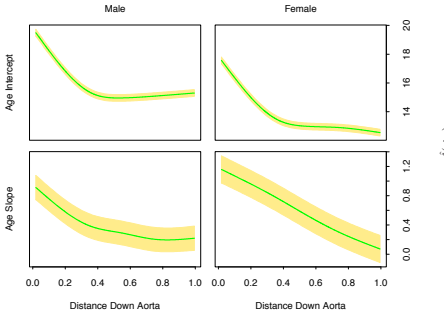
STAT/BIOSTAT 527, University of Washington
Emily Fox
April 17th, 2014

Confidence Bands

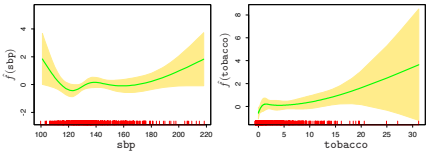
$y = f(x) + \epsilon$ $\quad var(\epsilon) = \sigma^2$ $\quad \sigma(x)^2$?

- 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***

homoscedastic $\sigma(x) = \sigma$



heteroscedastic $\sigma(x)$



From Hastie, Tibshirani, Friedman book

Bias Problem

- Typically, these are of the form $\hat{f}(x) \pm c\ \widehat{\text{se}}(x)$ — est. of st. dev. of $\hat{f}(x)$
- 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,

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

$s(x)$: st. dev. of $\hat{f}(x)$ $\quad = Z_n(x) + \dfrac{bias(\hat{f}(x))}{\sqrt{var(\hat{f}(x))}}$

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*

est. this is harder than est. f!

- So, what should we do?
  - Option #1: Estimate the bias

Hard. Lead term is $f''(x)$

- 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 $\sigma(x) = \sigma$

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

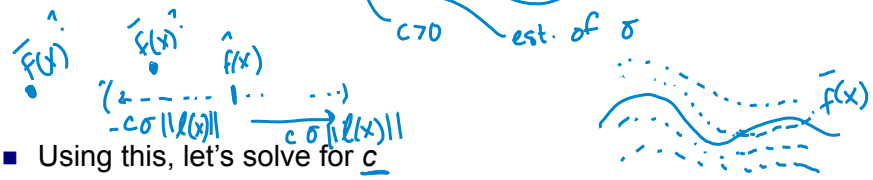
$$\bar{f}(x) = \sum_{i=1}^{n} \ell_i(x) f(x_i)$$

$$\text{var}(\hat{f}(x)) = \sigma^2 \|\ell(x)\|^2$$

- Consider confidence band of the form

$$CI(x) = \hat{f}(x) \pm c\,\hat{\sigma}\,\|\ell(x)\| \qquad a \le x \le b$$

$c > 0$; est. of σ



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

CIs for Linear Smoothers

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

$$P(\bar{f}(x) \notin \mathrm{CI}(x) \text{ for some } x \in [a,b]) = P\left(\max_{x\in[a,b]} \frac{|\hat{f}(x)-\bar{f}(x)|}{\sigma\|\ell(x)\|} > c\right)$$

$$= P\left(\max_{x\in[a,b]} \frac{\sum \epsilon_i \ell_i(x)}{\sigma\|\ell(x)\|} > c\right) = P\left(\max_x |W(x)| > c\right)$$

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

- Good news: max of GP is well studied!

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

"Tube formula"

- 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

$$\mathrm{var}(\hat{f}(x)) = \sum_i \sigma^2(x_i)\ell_i^2(x)$$

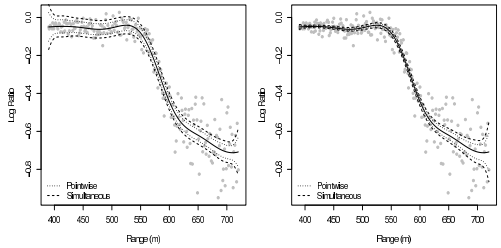
$$\mathrm{CI}(x) = \hat{f}(x) \pm c\sqrt{\sum_i \sigma^2(x_i)\ell_i^2(x)}$$

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

$$\Rightarrow \mathrm{CI}(x) = \hat{f}(x) + c\,\hat{\sigma}(x)\|\ell(x)\|$$

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

$\alpha \to \frac{\alpha}{m}$, $m = \#$ of λ

Variance Estimation

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

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

$\nu = \mathrm{tr}(L)$ $\quad \tilde{\nu} = \mathrm{tr}(L^T L) = \sum_i \|l(x_i)\|^2$

  - 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

$$y - \hat{f} = y - Ly = (I - L)y \triangleq \Lambda^{1/2} y$$

    and

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

  - Then,

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

$$E[\hat{\sigma}^2] = \frac{\mathrm{tr}(\Lambda\sigma^2) + f^T \Lambda f}{\mathrm{tr}(\Lambda)} = \sigma^2 + \frac{f^T \Lambda f}{n - 2\nu + \tilde{\nu}}$$

  - 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:

$$y_{i+1} - y_i = [f(x_{i+1}) - f(x_i)] + [\epsilon_{i+1} - \epsilon_i]$$

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

$$\Rightarrow E[\hat{\sigma}^2] \approx \sigma^2$$

- 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 $y_i = f(x_i) + \sigma(x_i)\epsilon_i$ $E[\epsilon]=0$ $var(\epsilon)=1$ var $\sigma^2(x_i)$
  - Define redefine obs. $Z_i = \log(y_i - f(x_i))^2 \quad \delta_i = \log \epsilon_i^2$
  - Then, $Z_i = \log \sigma^2(x_i) + \delta_i$
- Algorithm: est. w/ log sq. residuals
  1. Estimate $f(x)$ using a nonparametric method w/ constant var to get $\hat{f}(x)$
  2. Define $Z_i = \log(y_i - \hat{f}(x_i))^2$ est. using $\hat{f}(x)$ to get log. sq. res.
  3. Regress Z_i's on x_i's to get estimate $\hat{g}(x)$ of $\log \sigma^2(x)$

$\hat{\sigma}^2(x) = e^{\hat{g}(x)}$ $Z_i = g(x_i) + \delta_i$, $\log(\sigma^2(x_i))$

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*

prev. alg. is a 2-stage approach

- 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