

Nonparametrics

Michael Peress

SUNY-Stony Brook

October 24, 2023

Nonparametrics

Overview of Nonparametric Models

- ▶ Parametric Model: statistical model characterized by finite dimensional unknown parameter
 - $y_n \sim N(\mu, \sigma^2)$

Nonparametrics

Overview of Nonparametric Models

- ▶ Parametric Model: statistical model characterized by finite dimensional unknown parameter
 - $y_n \sim N(\mu, \sigma^2)$
 - $y_n|x_n \sim N(\beta'x_n, \sigma^2)$ (normal-linear model)

Nonparametrics

Overview of Nonparametric Models

- ▶ Parametric Model: statistical model characterized by finite dimensional unknown parameter
 - $y_n \sim N(\mu, \sigma^2)$
 - $y_n|x_n \sim N(\beta'x_n, \sigma^2)$ (normal-linear model)
- ▶ Nonparametric Model: statistical model characterized by infinite dimensional unknown parameter
 - $y_n \sim f$ where f is unknown (density estimation)

Nonparametrics

Overview of Nonparametric Models

- ▶ Parametric Model: statistical model characterized by finite dimensional unknown parameter
 - $y_n \sim N(\mu, \sigma^2)$
 - $y_n|x_n \sim N(\beta'x_n, \sigma^2)$ (normal-linear model)
- ▶ Nonparametric Model: statistical model characterized by infinite dimensional unknown parameter
 - $y_n \sim f$ where f is unknown (density estimation)
 - $y_n = g(x_n) + \varepsilon_n, E[\varepsilon_n|x_n] = 0$ (nonparametric regression)

Nonparametrics

Overview of Nonparametric Models

- ▶ Semiparametric Model: statistical model characterized by finite dimensional parameter of interest and infinite dimensional **nuisance parameter**
 - $y_n = \beta' x_n + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$
(semiparametric linear model)

Nonparametrics

Overview of Nonparametric Models

- ▶ Semiparametric Model: statistical model characterized by finite dimensional parameter of interest and infinite dimensional **nuisance parameter**

- $y_n = \beta' x_n + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$
(semiparametric linear model)

- ▶ β is **parameter of interest**
- ▶ F is a nuisance parameter

Nonparametrics

Overview of Nonparametric Models

- ▶ Semiparametric Model: statistical model characterized by finite dimensional parameter of interest and infinite dimensional **nuisance parameter**

- $y_n = \beta' x_n + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$ (semiparametric linear model)
 - ▶ β is **parameter of interest**
 - ▶ F is a nuisance parameter
- $y_n = g(\beta' x_n) + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$ (linear index model)

Nonparametrics

Overview of Nonparametric Models

- ▶ Semiparametric Model: statistical model characterized by finite dimensional parameter of interest and infinite dimensional **nuisance parameter**

- $y_n = \beta' x_n + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$ (semiparametric linear model)
 - ▶ β is **parameter of interest**
 - ▶ F is a nuisance parameter
- $y_n = g(\beta' x_n) + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$ (linear index model)
- $y_n = g(x_n) + \beta' z_n + \varepsilon_n$, $\varepsilon_n | x_n \sim F(\varepsilon | x)$ with $E[\varepsilon_n | x_n] = 0$ (partially linear model)

Nonparametrics

Overview of Nonparametric Models

▶ Parametric Models:

- MLE is efficient if parametric model is correct

Nonparametrics

Overview of Nonparametric Models

► Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect

Nonparametrics

Overview of Nonparametric Models

► Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

Nonparametrics

Overview of Nonparametric Models

▶ Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

▶ Nonparametric Models:

- More generality, but...

Nonparametrics

Overview of Nonparametric Models

► Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

► Nonparametric Models:

- More generality, but...
- Theory more difficult

Nonparametrics

Overview of Nonparametric Models

▶ Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

▶ Nonparametric Models:

- More generality, but...
- Theory more difficult
- Implementation difficult

Nonparametrics

Overview of Nonparametric Models

► Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

► Nonparametric Models:

- More generality, but...
- Theory more difficult
- Implementation difficult
- Slower convergence (slower than parametric rate of \sqrt{N})

Nonparametrics

Overview of Nonparametric Models

► Parametric Models:

- MLE is efficient if parametric model is correct
- MLE is often inconsistent if parametric model is incorrect
- \sqrt{N} -convergence rate

► Nonparametric Models:

- More generality, but...
- Theory more difficult
- Implementation difficult
- Slower convergence (slower than parametric rate of \sqrt{N})
- Efficiency loss (relative to MLE if parametric model is correct)

Nonparametrics

Overview of Nonparametric Models

- ▶ Semiparametric Models:
 - More generality, and...

Nonparametrics

Overview of Nonparametric Models

► Semiparametric Models:

- More generality, and...
- Often, \sqrt{N} -convergence for parameter of interest

Nonparametrics

Overview of Nonparametric Models

► Semiparametric Models:

- More generality, and...
- Often, \sqrt{N} -convergence for parameter of interest
- Sometimes, easy to implement

Nonparametrics

Overview of Nonparametric Models

► Semiparametric Models:

- More generality, and...
- Often, \sqrt{N} -convergence for parameter of interest
- Sometimes, easy to implement
- Often, little efficiency loss

Nonparametrics

Overview of Nonparametric Models

- ▶ Examples of “Easy” Semiparametric Estimators:
 - OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic

Nonparametrics

Overview of Nonparametric Models

- ▶ Examples of “Easy” Semiparametric Estimators:
 - OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic
 - Poisson regression w/ robust se's - semiparametric because estimator is consistent when dependent variable is not Poisson distributed

Nonparametrics

Overview of Nonparametric Models

- ▶ Examples of “Easy” Semiparametric Estimators:
 - OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic
 - Poisson regression w/ robust se's - semiparametric because estimator is consistent when dependent variable is not Poisson distributed
 - ATE with randomized binary treatment

Nonparametrics

Overview of Nonparametric Models

- ▶ Examples of “Easy” Semiparametric Estimators:
 - OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic
 - Poisson regression w/ robust se's - semiparametric because estimator is consistent when dependent variable is not Poisson distributed
 - ATE with randomized binary treatment
- ▶ Examples of Harder Semiparametric Estimator:
 - Linear index models

Nonparametrics

Overview of Nonparametric Models

- ▶ Examples of “Easy” Semiparametric Estimators:
 - OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic
 - Poisson regression w/ robust se's - semiparametric because estimator is consistent when dependent variable is not Poisson distributed
 - ATE with randomized binary treatment
- ▶ Examples of Harder Semiparametric Estimator:
 - Linear index models
 - Partially linear model

Nonparametrics

Overview of Nonparametric Models

► Examples of “Easy” Semiparametric Estimators:

- OLS w/ robust se's - semiparametric because OLS is consistent even if error terms are non-normal and heteroskedastic
- Poisson regression w/ robust se's - semiparametric because estimator is consistent when dependent variable is not Poisson distributed
- ATE with randomized binary treatment

► Examples of Harder Semiparametric Estimator:

- Linear index models
- Partially linear model
- ATE without randomized treatment

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation
- ▶ Other nonparametric methods:
 - k-nearest neighbor estimators (also called **matching** estimators)

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation
- ▶ Other nonparametric methods:
 - k-nearest neighbor estimators (also called **matching** estimators)
 - Smoothing splines

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation
- ▶ Other nonparametric methods:
 - k-nearest neighbor estimators (also called **matching** estimators)
 - Smoothing splines
 - Sieve estimators (estimation using orthogonal functions such as polynomials or Fourier series), possibly combined with LASSO

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation
- ▶ Other nonparametric methods:
 - k-nearest neighbor estimators (also called **matching** estimators)
 - Smoothing splines
 - Sieve estimators (estimation using orthogonal functions such as polynomials or Fourier series), possibly combined with LASSO
 - Tree-based methods

Nonparametrics

Overview of Nonparametric Models

- ▶ We will start with Kernel techniques applied to density estimation
- ▶ Kernel techniques generalize to problems beyond density estimation
- ▶ Other nonparametric methods:
 - k-nearest neighbor estimators (also called **matching** estimators)
 - Smoothing splines
 - Sieve estimators (estimation using orthogonal functions such as polynomials or Fourier series), possibly combined with LASSO
 - Tree-based methods
 - Support vector machines

Nonparametrics

Kernel Density Estimation

► The Density Estimation Problem:

- We assume that $\{X_n\}_{n=1}^N$ are i.i.d. draws from a common distribution $f_0(x)$

Nonparametrics

Kernel Density Estimation

► The Density Estimation Problem:

- We assume that $\{X_n\}_{n=1}^N$ are i.i.d. draws from a common distribution $f_0(x)$
- The density estimation problem is the problem of estimating $f_0(x)$ while placing only minimal restrictions on f_0

Nonparametrics

Kernel Density Estimation

► The Density Estimation Problem:

- We assume that $\{X_n\}_{n=1}^N$ are i.i.d. draws from a common distribution $f_0(x)$
- The density estimation problem is the problem of estimating $f_0(x)$ while placing only minimal restrictions on f_0
- We would like to develop an estimator \hat{f} of the density f_0

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

- ▶ Here, K denotes the kernel and h denotes the bandwidth

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

- ▶ Here, K denotes the kernel and h denotes the bandwidth
- ▶ The Kernel satisfies:
 - (i) $K(u) \geq 0$

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

- ▶ Here, K denotes the kernel and h denotes the bandwidth
- ▶ The Kernel satisfies:
 - (i) $K(u) \geq 0$
 - (ii) $\int_u K(u) dx = 1$

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

- ▶ Here, K denotes the kernel and h denotes the bandwidth
- ▶ The Kernel satisfies:
 - (i) $K(u) \geq 0$
 - (ii) $\int_{-\infty}^{\infty} K(u) du = 1$
 - (iii) $\int_{-\infty}^{\infty} uK(u) du = 0$

Nonparametrics

Kernel Density Estimation

- ▶ The Kernel Density Estimator (KDE) is defined by,

$$\hat{f}(x; h) = \frac{1}{hN} \sum_{n=1}^N K\left(\frac{X_n - x}{h}\right)$$

- ▶ Here, K denotes the kernel and h denotes the bandwidth
- ▶ The Kernel satisfies:

- $K(u) \geq 0$
- $\int_{-\infty}^{\infty} K(u) du = 1$
- $\int_{-\infty}^{\infty} uK(u) du = 0$
- $\int_{-\infty}^{\infty} u^2 K(u) du > 0$

Nonparametrics

Kernel Density Estimation

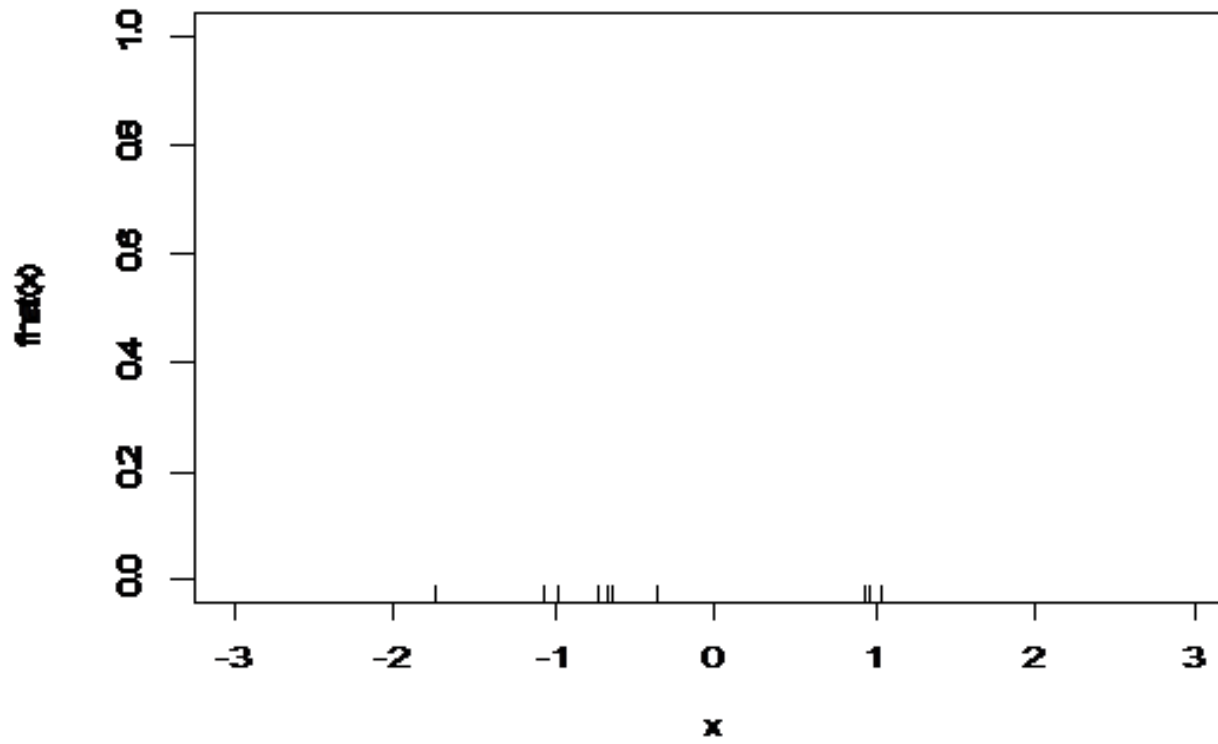
► Examples of Kernels:

Name	Kernel ($K(u)$)
Uniform	$\begin{cases} \frac{1}{2}, & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$
Triangle	$\begin{cases} 1 - u , & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$
Epanechnikov	$\begin{cases} \frac{3}{4}(1 - u^2), & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$

Nonparametrics

Kernel Density Estimation

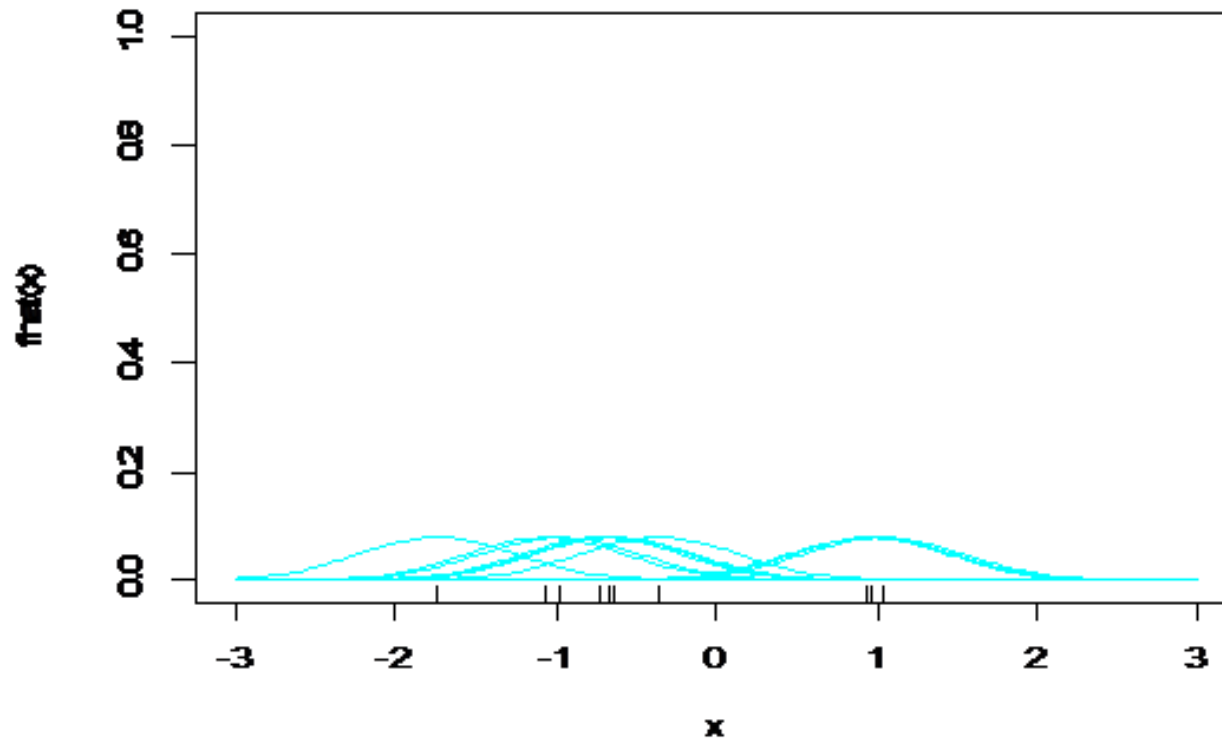
- ▶ Example w/ $N=10$ Data Points – Rug Plot:



Nonparametrics

Kernel Density Estimation

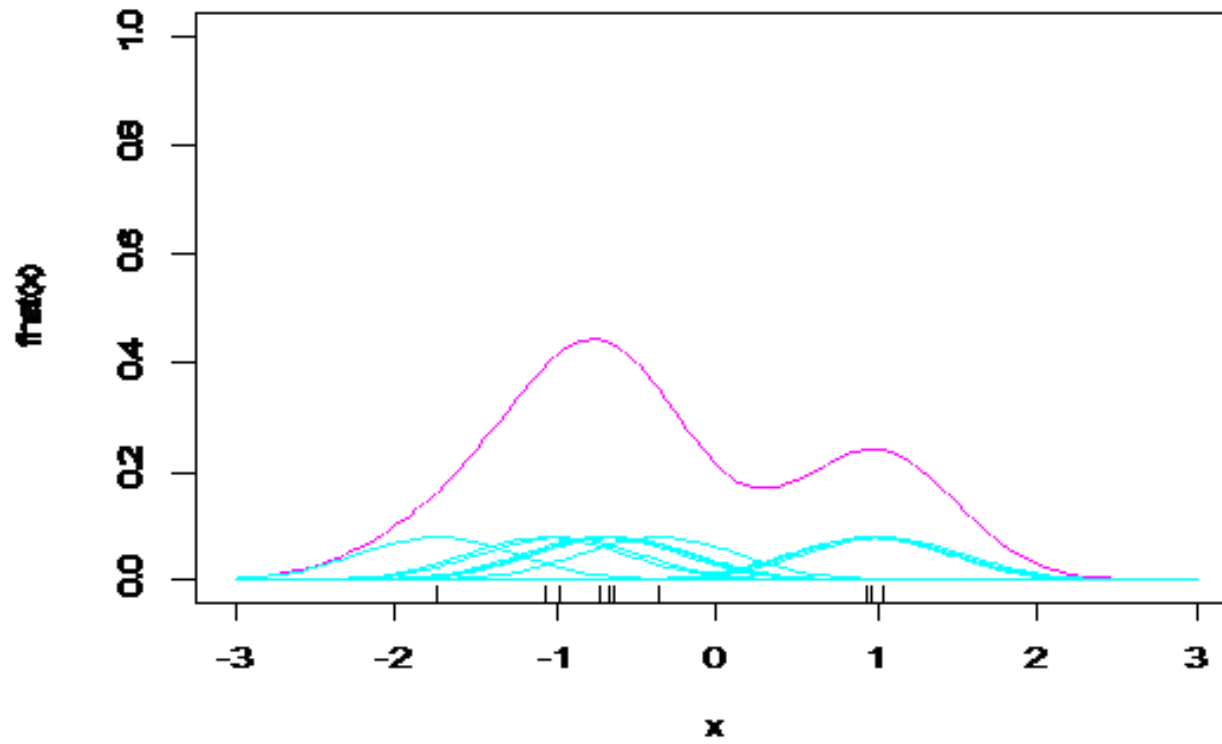
- ▶ Example w/ $N=10$ Data Points – Individual Kernels ($h = .5$):



Nonparametrics

Kernel Density Estimation

- ▶ Example w/ $N=10$ Data Points – Density Estimate ($h = .5$):



Nonparametrics

Kernel Density Estimation

- ▶ The bandwidth h controls the amount of smoothing

Nonparametrics

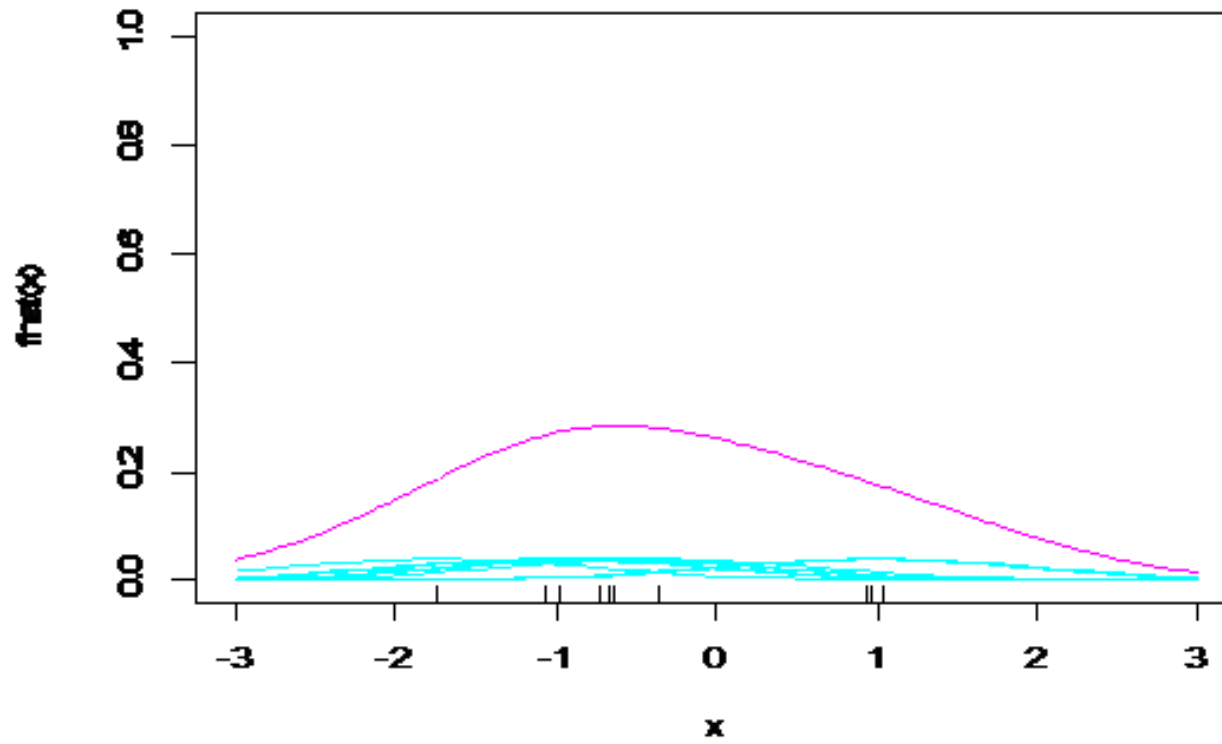
Kernel Density Estimation

- ▶ The bandwidth h controls the amount of smoothing
 - Large values of h denote a large degree of smoothing
 - Small values of h denotes a small degree of smoothing

Nonparametrics

Kernel Density Estimation

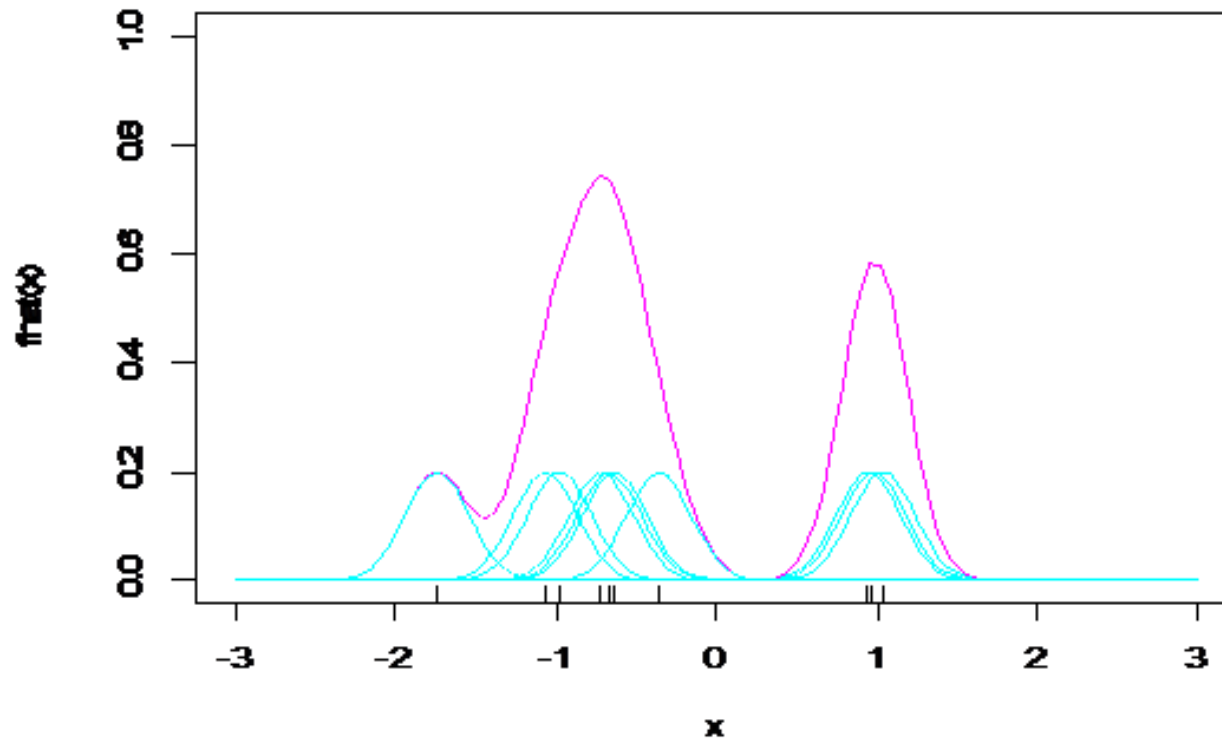
- ▶ Example w/ $N=10$ Data Points – Density Estimate ($h = 1$):



Nonparametrics

Kernel Density Estimation

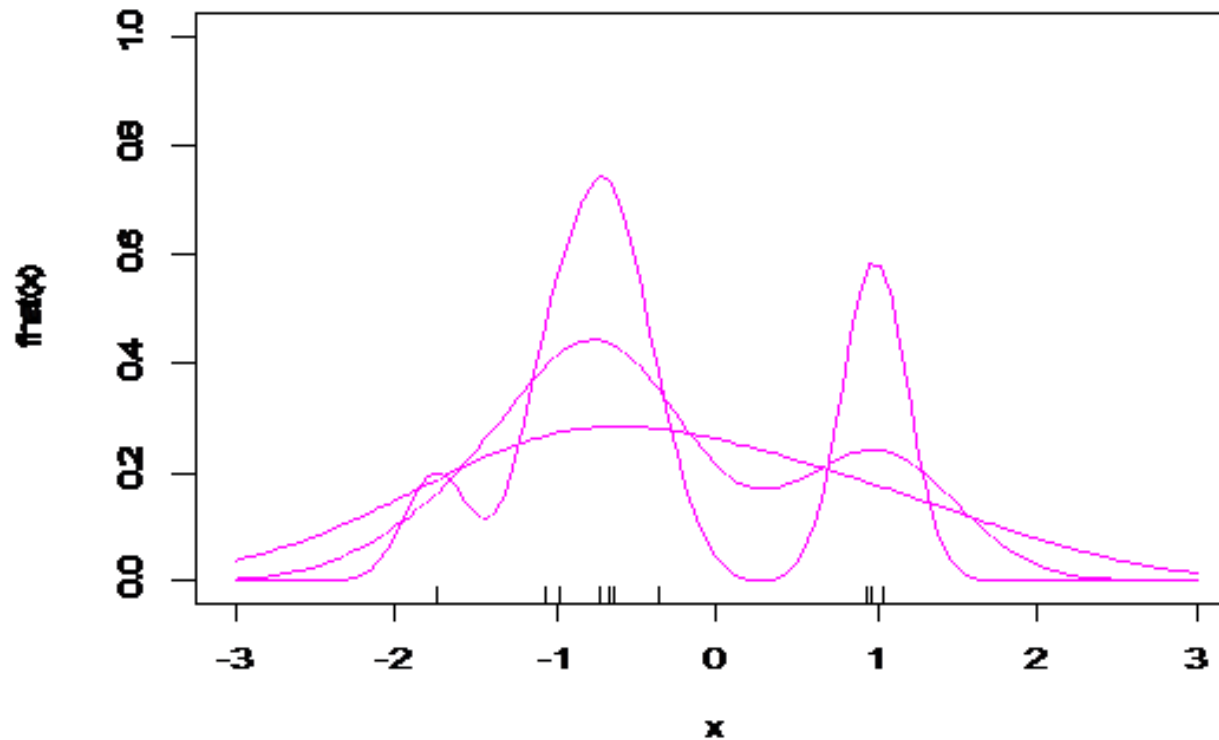
- ▶ Example w/ $N=10$ Data Points – Density Estimate ($h = .2$):



Nonparametrics

Kernel Density Estimation

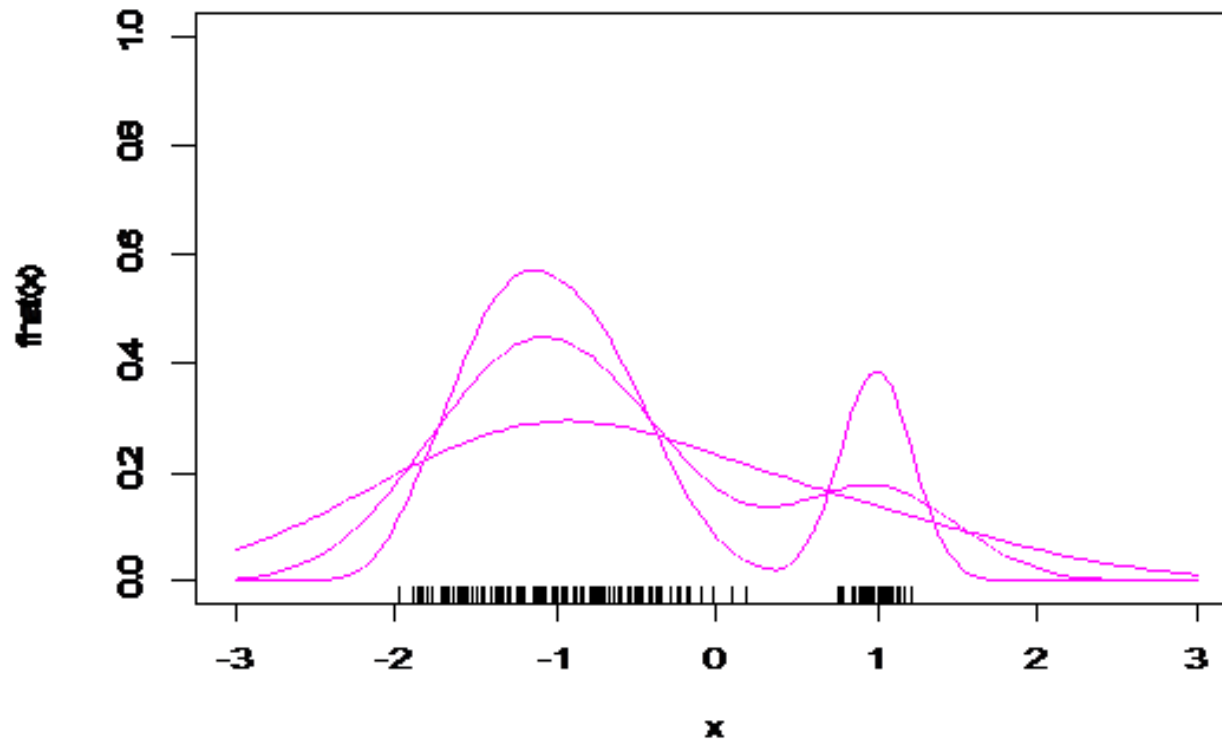
- ▶ Example w/ $N=10$ Data Points – Density Estimate ($h = .2$, $h = .5$, and $h = 1$):



Nonparametrics

Kernel Density Estimation

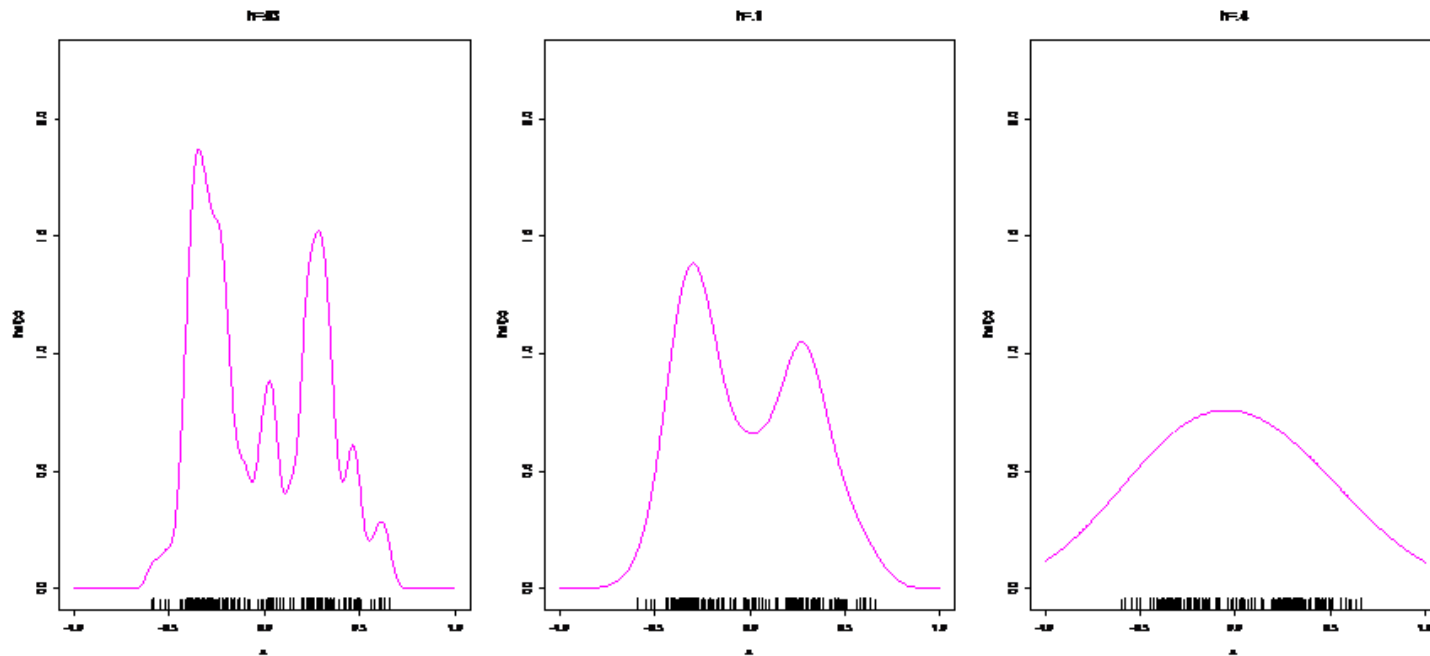
- ▶ Example w/ $N=200$ Data Points – Density Estimate ($h = .2$, $h = .5$, and $h = 1$):



Nonparametrics

Kernel Density Estimation

- ▶ Example: Positions of Senate incumbents ($h = .03$, $h = .1$, and $h = .4$)



Nonparametrics

Kernel Density Estimation

- Bias goes to zero as h goes to zero:

$$\text{Bias}[\hat{f}(x; h)] = \frac{1}{2}\mu_2 h^2 f_0''(x) + o(h^3)$$

Nonparametrics

Kernel Density Estimation

- ▶ Bias goes to zero as h goes to zero:

$$\text{Bias}[\hat{f}(x; h)] = \frac{1}{2}\mu_2 h^2 f_0''(x) + o(h^3)$$

- ▶ Variance goes to 0 as Nh goes to ∞ :

$$\text{Var}(\hat{f}(x; h)) = \frac{1}{Nh} \nu_2 f_0(x) + O(N^{-1})$$

Nonparametrics

Kernel Density Estimation

- ▶ Bias goes to zero as h goes to zero:

$$\text{Bias}[\hat{f}(x; h)] = \frac{1}{2}\mu_2 h^2 f_0''(x) + o(h^3)$$

- ▶ Variance goes to 0 as Nh goes to ∞ :

$$\text{Var}(\hat{f}(x; h)) = \frac{1}{Nh} \nu_2 f_0(x) + O(N^{-1})$$

- ▶ Select h to minimize integrated mean-squared error:

$$\text{IMSE}(\hat{f}; h) = \frac{1}{Nh} \nu_2 + \frac{1}{4} h^4 \mu_2^2 \int_x f_0''(x)^2 dx + O(N^{-1}) + o(h^5)$$

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$
- ▶ h^* will clearly satisfy this

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$
- ▶ h^* will clearly satisfy this
- ▶ This result tells us that rate at which to increase h to obtain optimal results, but we still need a way to determine the constant

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$
- ▶ h^* will clearly satisfy this
- ▶ This result tells us that rate at which to increase h to obtain optimal results, but we still need a way to determine the constant
- ▶ Notice that ν_2 and μ_2 can be computed easily

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$
- ▶ h^* will clearly satisfy this
- ▶ This result tells us that rate at which to increase h to obtain optimal results, but we still need a way to determine the constant
- ▶ Notice that ν_2 and μ_2 can be computed easily
- ▶ Must have estimate of $\int_x f_0''(x)^2 dx$

Nonparametrics

Kernel Density Estimation

- ▶ Theoretical bandwidth that minimizes IMSE:

$$h^* = \left(\frac{\nu_2}{\mu_2^2 \int_x f_0''(x)^2 dx} \right)^{1/5} N^{-1/5}$$

- ▶ Naturally, we would like $IMSE(\hat{f}; h) \rightarrow 0$ (which is a property weaker than consistency)
- ▶ We require $h \rightarrow 0$ and $Nh \rightarrow 0$ as $N \rightarrow \infty$
- ▶ h^* will clearly satisfy this
- ▶ This result tells us that rate at which to increase h to obtain optimal results, but we still need a way to determine the constant
- ▶ Notice that ν_2 and μ_2 can be computed easily
- ▶ Must have estimate of $\int_x f_0''(x)^2 dx$
- ▶ Estimating $f_0(x)$ requires estimating $\int_x f_0''(x)^2 dx!$

Nonparametrics

Kernel Density Estimation

► Constants Characterizing Kernel:

Name	$K(u)$	μ_2	ν_2
Uniform	$\begin{cases} \frac{1}{2}, & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$	$\frac{1}{3}$	$\frac{1}{2}$
Triangle	$\begin{cases} 1 - u , & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$	$\frac{1}{6}$	$\frac{2}{3}$
Epanech.	$\begin{cases} \frac{3}{4}(1 - u^2), & -1 \leq u \leq 1 \\ 0, & \textit{otherwise} \end{cases}$	$\frac{1}{5}$	$\frac{3}{5}$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	1	$\frac{1}{2\sqrt{\pi}}$

Nonparametrics

Selecting the Bandwidth

► Normal Reference Rule:

- Compute $\int_x f_0''(x)^2 dx$ for some special density

Nonparametrics

Selecting the Bandwidth

► Normal Reference Rule:

- Compute $\int_x f_0''(x)^2 dx$ for some special density
- The normal reference rule involves assuming that $f_0(x)$ is the $N(\mu, \sigma^2)$ distribution

Nonparametrics

Selecting the Bandwidth

► Normal Reference Rule:

- Compute $\int_x f_0''(x)^2 dx$ for some special density
- The normal reference rule involves assuming that $f_0(x)$ is the $N(\mu, \sigma^2)$ distribution
- We have that $\int_x f_0''(x)^2 dx = \frac{3}{8\sigma^5\sqrt{\pi}}$

Nonparametrics

Selecting the Bandwidth

► Normal Reference Rule:

- Compute $\int_x f_0''(x)^2 dx$ for some special density
- The normal reference rule involves assuming that $f_0(x)$ is the $N(\mu, \sigma^2)$ distribution
- We have that $\int_x f_0''(x)^2 dx = \frac{3}{8\sigma^5\sqrt{\pi}}$
- Normal reference rule (or rule-of-thumb bandwidth) suggests,

$$h = \left(\frac{\nu_2 8\sqrt{\pi}}{3\mu_2^2} \right)^{1/5} \sigma N^{-1/5} = c\sigma N^{-1/5}$$

Nonparametrics

Selecting the Bandwidth

- ▶ For the normal kernel, we can determine that $c = \left(\frac{4}{3}\right)^{1/5} \approx 1.059$, so that,

$$h = 1.059\sigma N^{-1/5}$$

Nonparametrics

Selecting the Bandwidth

- ▶ For the normal kernel, we can determine that $c = \left(\frac{4}{3}\right)^{1/5} \approx 1.059$, so that,

$$h = 1.059\sigma N^{-1/5}$$

- ▶ Other kernels will yield different constants. To estimate σ , we could use the variance of the data

Nonparametrics

Selecting the Bandwidth

- ▶ For the normal kernel, we can determine that $c = \left(\frac{4}{3}\right)^{1/5} \approx 1.059$, so that,

$$h = 1.059\sigma N^{-1/5}$$

- ▶ Other kernels will yield different constants. To estimate σ , we could use the variance of the data
- ▶ Silverman (1986) suggests employing a robust estimator,

$$\hat{\sigma} = \min\{s, 1.34(q_{0.75} - q_{0.25})\}$$

where $q_{0.25}$ and $q_{0.75}$ represent the 25 and 75% quantiles

Nonparametrics

Selecting the Bandwidth

► Plug-In Method:

- Estimate $\int_x f_0''(x)^2 dx$ rather than guessing it

Nonparametrics

Selecting the Bandwidth

► Plug-In Method:

- Estimate $\int_x f_0''(x)^2 dx$ rather than guessing it
- Use normal reference rule to obtain an initial kernel density estimator, $\hat{f}(x)$

Nonparametrics

Selecting the Bandwidth

► Plug-In Method:

- Estimate $\int_x f_0''(x)^2 dx$ rather than guessing it
- Use normal reference rule to obtain an initial kernel density estimator, $\hat{f}(x)$
- Then, we can use this to approximate $\int_x f_0''(x)^2 dx$ by taking a second numerical derivative of $\hat{f}(x)$ and integrating

Nonparametrics

Selecting the Bandwidth

► Plug-In Method:

- Estimate $\int_x f_0''(x)^2 dx$ rather than guessing it
- Use normal reference rule to obtain an initial kernel density estimator, $\hat{f}(x)$
- Then, we can use this to approximate $\int_x f_0''(x)^2 dx$ by taking a second numerical derivative of $\hat{f}(x)$ and integrating
- We then re-estimate $f_0(x)$ using the new bandwidth

Nonparametrics

Selecting the Bandwidth

► Cross Validation:

- Obtain an estimate of the integrated mean-squared error as a function of h , and minimize it

Nonparametrics

Selecting the Bandwidth

► Cross Validation:

- Obtain an estimate of the integrated mean-squared error as a function of h , and minimize it
- The actual integrated mean squared error is,

$$\begin{aligned}IMSE(h) &= \int_x (\hat{f}(x; h) - f_0(x))^2 dx \\ &= \int_x \hat{f}^2(x; h) dx + \int_x f_0^2(x) dx - 2 \int_x \hat{f}(x; h) f_0(x) dx\end{aligned}$$

Nonparametrics

Selecting the Bandwidth

- ▶ Estimate $IMSE(h)$ using leave-one-out estimator,

$$\frac{1}{N^2 h} \sum_{n=1}^N \sum_{m=1}^N (K \circ K) \left(\frac{X_n - X_m}{h} \right) - 2 \frac{1}{Nh(N-1)} \sum_{n=1}^N \sum_{m \neq n} K \left(\frac{X_n - X_m}{h} \right)$$

where $K \circ K$ denotes the convolution of the kernel with itself and can be (tediously) computed analytically for a given choice of kernel

Nonparametrics

Selecting the Bandwidth

- ▶ The expression $IM\hat{S}E(h)$ can then be numerically minimized to determine the cross validation bandwidth, h_{CV}

Nonparametrics

Selecting the Bandwidth

- ▶ The expression $IM\hat{S}E(h)$ can then be numerically minimized to determine the cross validation bandwidth, h_{CV}
- ▶ One must be careful however, because if there are two data points such that $X_n = X_m$, then the cross validation function will have a minimum at 0

Nonparametrics

Selecting the Bandwidth

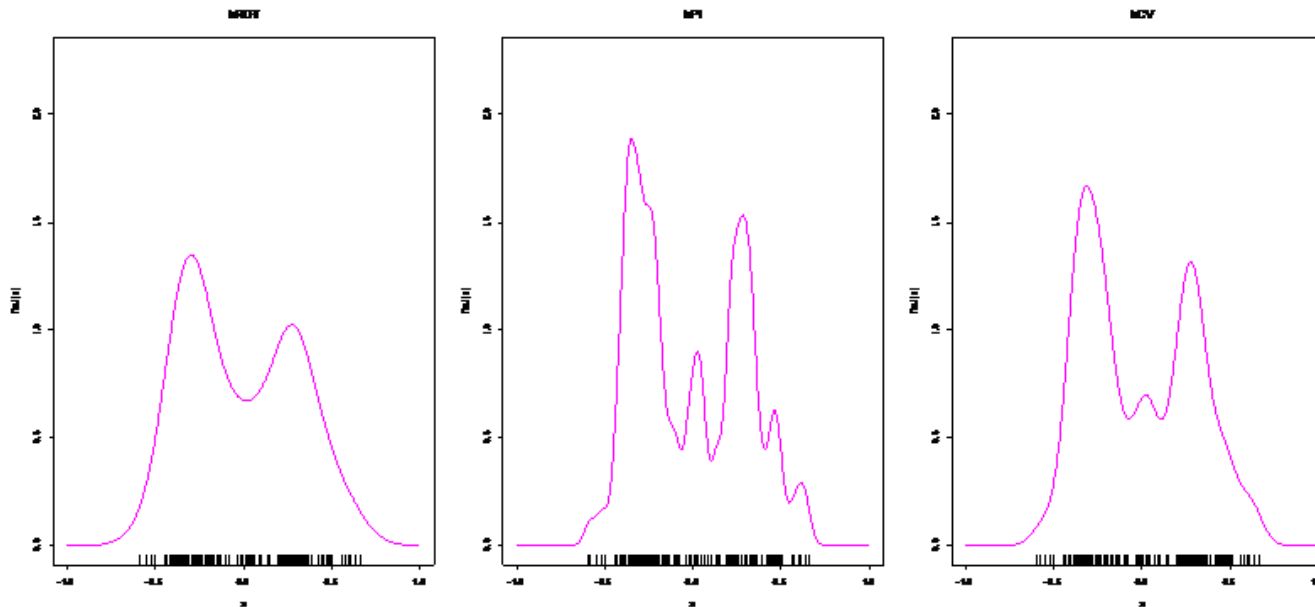
- ▶ The expression $IM\hat{S}E(h)$ can then be numerically minimized to determine the cross validation bandwidth, h_{CV}
- ▶ One must be careful however, because if there are two data points such that $X_n = X_m$, then the cross validation function will have a minimum at 0
- ▶ A solution is to use the estimator,

$$\frac{1}{N^2 h} \sum_{n=1}^N \sum_{m=1}^N (K \circ K) \left(\frac{X_n - X_m}{h} \right) - 2 \frac{1}{N h (N-1)} \sum_{n=1}^N \sum_{X_m \neq X_n} K \left(\frac{X_n - X_m}{h} \right)$$

Nonparametrics

Selecting the Bandwidth

- ▶ Senate Incumbent Positions Example continued:
 - We can determine that $h_{ROT} = 0.106$, $h_{PI} = 0.028$, and $h_{CV} = 0.059$



Nonparametrics

Selecting the Kernel

- ▶ Suppose that we plug the optimal bandwidth into the formula for the integrated mean squared error,

$$IMSE(\hat{f}, h) = \frac{5}{4}(\mu_2\nu_2^2)^{2/5} \left(\int_x f''(x)^2 dx \right)^{1/5} N^{-4/5} + o(N^{-1})$$

Nonparametrics

Selecting the Kernel

- ▶ Suppose that we plug the optimal bandwidth into the formula for the integrated mean squared error,

$$IMSE(\hat{f}, h) = \frac{5}{4}(\mu_2\nu_2^2)^{2/5} \left(\int_x f''(x)^2 dx \right)^{1/5} N^{-4/5} + o(N^{-1})$$

- ▶ The efficiency of the Kernel therefore depends on the constant $\mu_2^{2/5} \nu_2^{4/5}$.

Nonparametrics

Selecting the Kernel

- ▶ Suppose that we plug the optimal bandwidth into the formula for the integrated mean squared error,

$$IMSE(\hat{f}, h) = \frac{5}{4}(\mu_2\nu_2^2)^{2/5} \left(\int_x f''(x)^2 dx \right)^{1/5} N^{-4/5} + o(N^{-1})$$

- ▶ The efficiency of the Kernel therefore depends on the constant $\mu_2^{2/5} \nu_2^{4/5}$.
- ▶ We can choose K to solve the calculus of variations problem, minimizing $IMSE(K)$ subject to constraints based on (i) through (iv)

Nonparametrics

Selecting the Kernel

Kernel	μ_2	ν_2	Relative Efficiency
Uniform	$\frac{1}{3}$	$\frac{1}{2}$	1.060
Triangle	$\frac{1}{6}$	$\frac{2}{3}$	1.011
Epanechnikov	$\frac{1}{5}$	$\frac{3}{5}$	1.000
Gaussian	1	$\frac{1}{2\sqrt{\pi}}$	1.041

Nonparametrics

Selecting the Kernel

Kernel	μ_2	ν_2	Relative Efficiency
Uniform	$\frac{1}{3}$	$\frac{1}{2}$	1.060
Triangle	$\frac{1}{6}$	$\frac{2}{3}$	1.011
Epanechnikov	$\frac{1}{5}$	$\frac{3}{5}$	1.000
Gaussian	1	$\frac{1}{2\sqrt{\pi}}$	1.041

- ▶ Epanechnikov kernel is the most efficient, but the choice of a kernel in practice does not seem to matter much

Nonparametrics

Selecting the Kernel

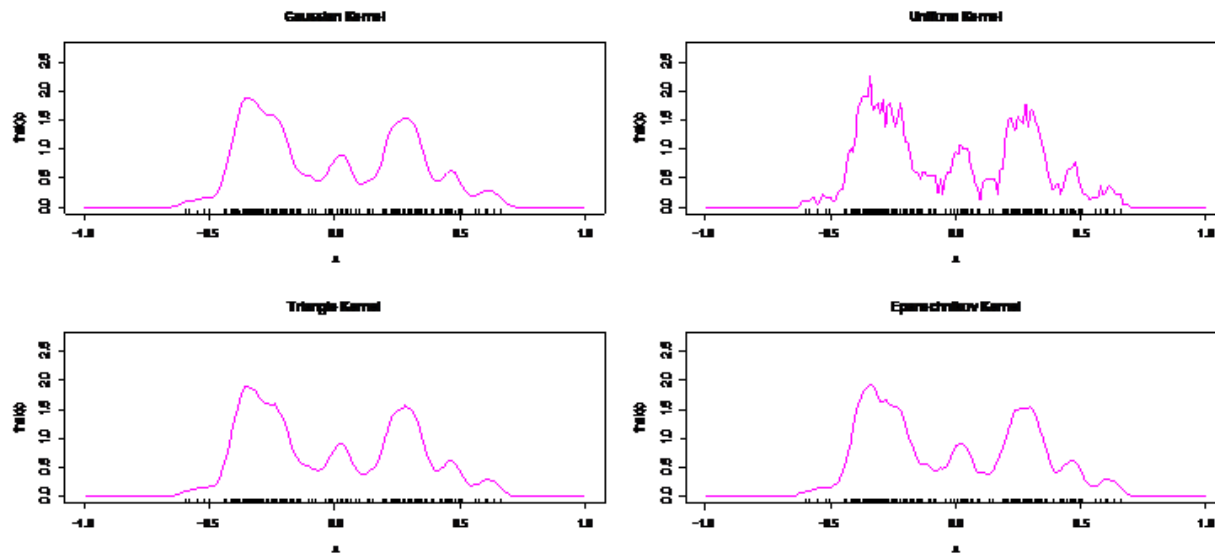
Kernel	μ_2	ν_2	Relative Efficiency
Uniform	$\frac{1}{3}$	$\frac{1}{2}$	1.060
Triangle	$\frac{1}{6}$	$\frac{2}{3}$	1.011
Epanechnikov	$\frac{1}{5}$	$\frac{3}{5}$	1.000
Gaussian	1	$\frac{1}{2\sqrt{\pi}}$	1.041

- ▶ Epanechnikov kernel is the most efficient, but the choice of a kernel in practice does not seem to matter much
- ▶ The effect of the kernel on mean-squared error is quite small, with the popular Gaussian kernel having an inefficiency of about 4%

Nonparametrics

Selecting the Kernel

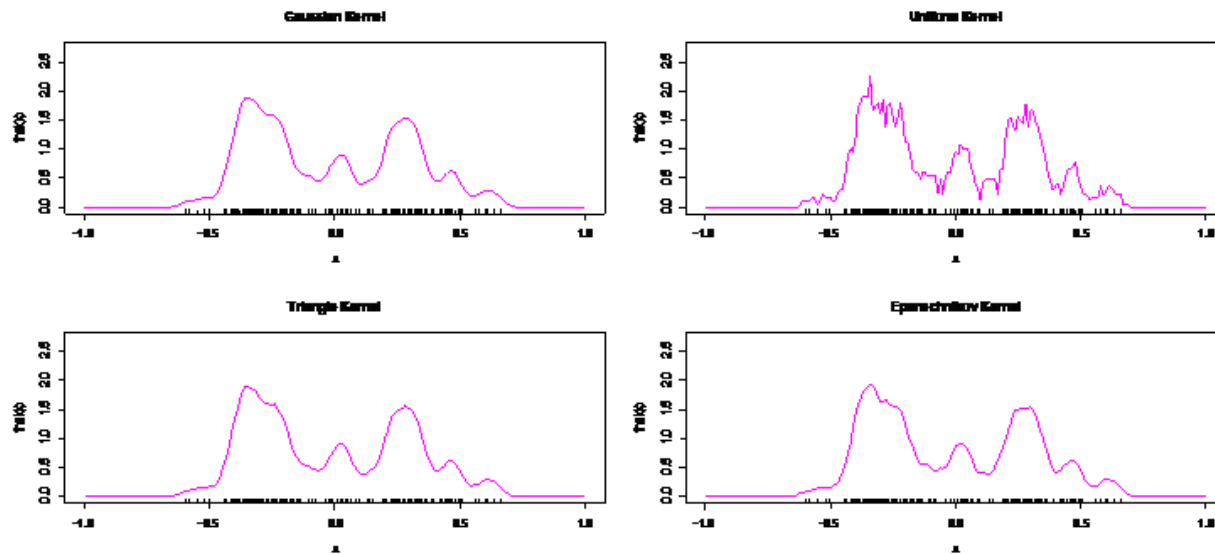
- ▶ Senate Incumbent Positions example continued ($h = h_{PI}$):



Nonparametrics

Selecting the Kernel

- ▶ Senate Incumbent Positions example continued ($h = h_{PI}$):

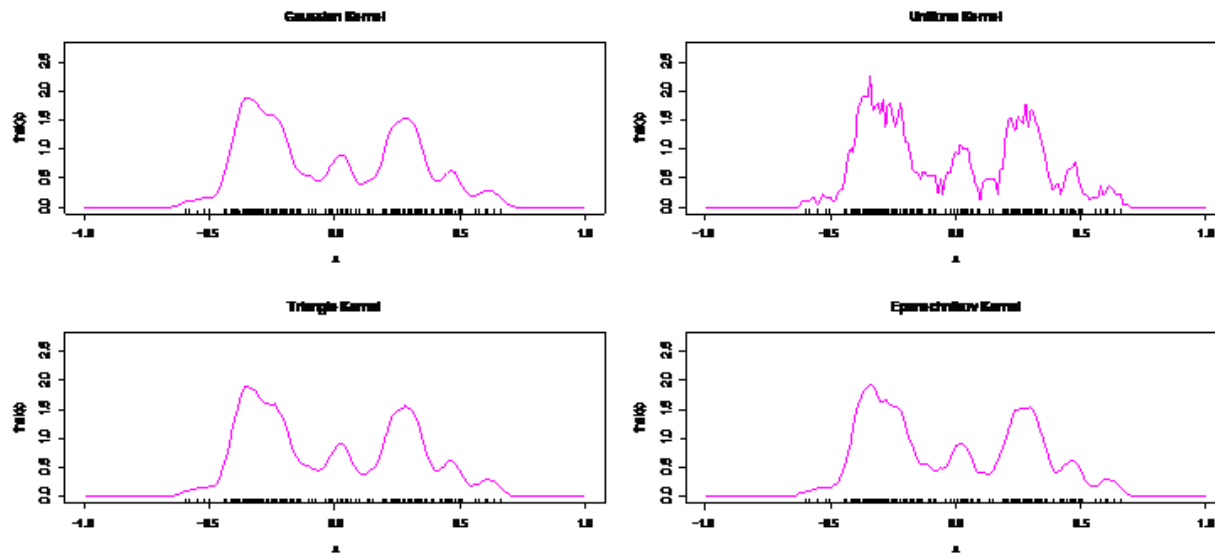


- ▶ Warning: different Kernels require different bandwidths

Nonparametrics

Selecting the Kernel

- ▶ Senate Incumbent Positions example continued ($h = h_{PI}$):



- ▶ Warning: different Kernels require different bandwidths
- ▶ In this case, $h_{Gaussian} = 0.028$, $h_{Unif.} = 0.028$, $h_{Tri.} = 0.069$, and $h_{Epan.} = 0.057$

Nonparametrics

Selecting the Kernel

- ▶ Notice that of all the kernels, the Gaussian kernel is the only one that has full support

Nonparametrics

Selecting the Kernel

- ▶ Notice that of all the kernels, the Gaussian kernel is the only one that has full support
- ▶ Full support is one reason that the Gaussian kernel is often chosen

Nonparametrics

Confidence Intervals

- ▶ Like most parametric estimators, kernel density estimators are consistent and asymptotically normal

Nonparametrics

Confidence Intervals

- ▶ Like most parametric estimators, kernel density estimators are consistent and asymptotically normal
- ▶ They do, however, converge at a slower rate than parametric estimators

$$\text{Var}(\hat{f}(x; h)) = \frac{1}{Nh} \nu_2 f_0(x) + O(N^{-1})$$

Nonparametrics

Confidence Intervals

- ▶ Like most parametric estimators, kernel density estimators are consistent and asymptotically normal
- ▶ They do, however, converge at a slower rate than parametric estimators

$$\text{Var}(\hat{f}(x; h)) = \frac{1}{Nh} \nu_2 f_0(x) + O(N^{-1})$$

- ▶ This implies that the estimator converges at the rate $(Nh)^{-1/2}$ rather than the usual $N^{-1/2}$

Nonparametrics

Confidence Intervals

- ▶ Like most parametric estimators, kernel density estimators are consistent and asymptotically normal
- ▶ They do, however, converge at a slower rate than parametric estimators

$$\text{Var}(\hat{f}(x; h)) = \frac{1}{Nh} \nu_2 f_0(x) + O(N^{-1})$$

- ▶ This implies that the estimator converges at the rate $(Nh)^{-1/2}$ rather than the usual $N^{-1/2}$
- ▶ When an optimal bandwidth is selected, the convergence rate is $N^{-2/5}$, which is of course slower than $N^{-1/2}$

Nonparametrics

Confidence Intervals

- ▶ Under the assumption that $h = h^*$, we can show that the kernel density estimator is asymptotically normally distributed:

$$\sqrt{hN}(\hat{f}(x; h) - f_0(x)) \xrightarrow{dist.} N\left(\frac{1}{2}\mu_2 f_0''(x)h^{5/2}N^{-1/2}, \nu_2 f_0(x)\right)$$

Nonparametrics

Confidence Intervals

- ▶ Under the assumption that $h = h^*$, we can show that the kernel density estimator is asymptotically normally distributed:

$$\sqrt{hN}(\hat{f}(x; h) - f_0(x)) \xrightarrow{dist.} N\left(\frac{1}{2}\mu_2 f_0''(x)h^{5/2}N^{-1/2}, \nu_2 f_0(x)\right)$$

- ▶ Notice that the asymptotic distribution is not centered at zero because (by construction) the bias and variance are of the same order

Nonparametrics

Confidence Intervals

- ▶ Under the assumption that $h = h^*$, we can show that the kernel density estimator is asymptotically normally distributed:

$$\sqrt{hN}(\hat{f}(x; h) - f_0(x)) \xrightarrow{dist.} N\left(\frac{1}{2}\mu_2 f_0''(x)h^{5/2}N^{-1/2}, \nu_2 f_0(x)\right)$$

- ▶ Notice that the asymptotic distribution is not centered at zero because (by construction) the bias and variance are of the same order
- ▶ We can eliminate the bias term by under-smoothing, selecting $h = cN^{-1/5+k}$ where $k > 0$

$$\sqrt{hN}(\hat{f}(x; h) - f_0(x)) \xrightarrow{dist.} N(0, \nu_2 f_0(x))$$

Nonparametrics

Confidence Intervals

- ▶ Two major approaches to conducting inferences for kernel density estimators—asymptotic formulas vs. the bootstrap

Nonparametrics

Confidence Intervals

- ▶ Two major approaches to conducting inferences for kernel density estimators—asymptotic formulas vs. the bootstrap
- ▶ Inference based on asymptotic formulas:
 - Asymptotic distribution w/ optimal smoothing

$$\hat{f}(x; h) - \frac{1}{2}\mu_2 \hat{f}_0''(x)h^{5/2}N^{-1/2} \pm 1.96\sqrt{\nu_2 \hat{f}(x)/\sqrt{hN}}$$

Nonparametrics

Confidence Intervals

- ▶ Two major approaches to conducting inferences for kernel density estimators—asymptotic formulas vs. the bootstrap
- ▶ Inference based on asymptotic formulas:
 - Asymptotic distribution w/ optimal smoothing

$$\hat{f}(x; h) - \frac{1}{2}\mu_2 \hat{f}_0''(x)h^{5/2}N^{-1/2} \pm 1.96\sqrt{\nu_2 \hat{f}(x)/\sqrt{hN}}$$

- Asymptotic distribution w/ under-smoothing

$$\hat{f}(x; h) \pm 1.96\sqrt{\nu_2 \hat{f}(x)/\sqrt{hN}}$$

Nonparametrics

Confidence Intervals

- ▶ Inference based on the bootstrap:
 - We sample S draws, with replacement, from $\{X_n\}_{n=1}^N$.

Nonparametrics

Confidence Intervals

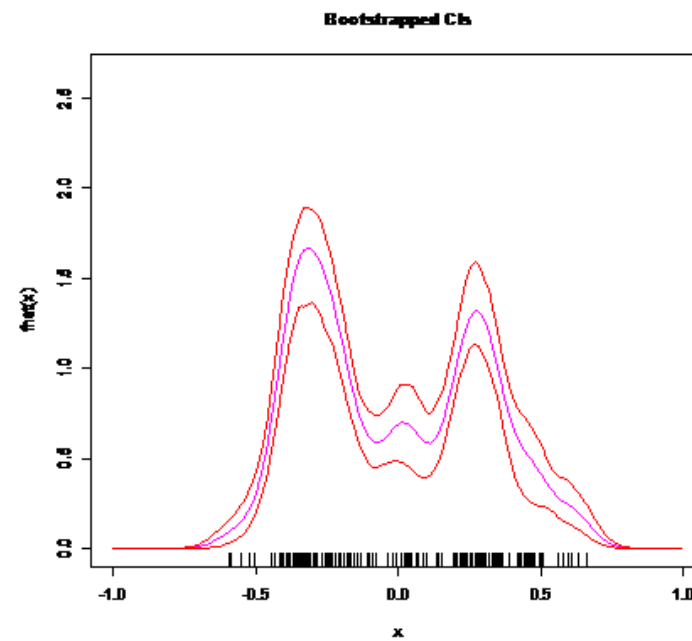
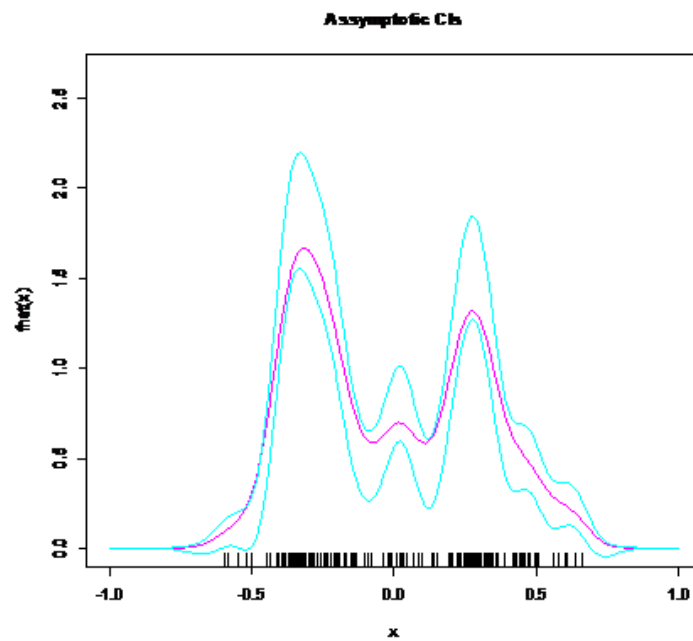
► Inference based on the bootstrap:

- We sample S draws, with replacement, from $\{X_n\}_{n=1}^N$.
- To compute the 95% confidence interval of $\hat{f}(x; h)$, we simply take the 2.5% and 97.5% quantiles of the empirical distribution $\hat{f}_s(x; h)$

Nonparametrics

Confidence Intervals

- Senate Incumbent Positions example continued:



Nonparametrics

Multivariate Density Estimation

- ▶ Consider now the problem of estimating a d -dimensional density $f_0(x)$

Nonparametrics

Multivariate Density Estimation

- ▶ Consider now the problem of estimating a d -dimensional density $f_0(x)$
- ▶ Define the multivariate kernel density estimator to be,

$$\hat{f}(x; h) = \frac{1}{Nh^d} \sum_{n=1}^N \prod_{i=1}^d K\left(\frac{X_{n,i} - X_i}{h}\right)$$

Nonparametrics

Multivariate Density Estimation

- ▶ Consider now the problem of estimating a d -dimensional density $f_0(x)$
- ▶ Define the multivariate kernel density estimator to be,

$$\hat{f}(x; h) = \frac{1}{Nh^d} \sum_{n=1}^N \prod_{i=1}^d K\left(\frac{X_{n,i} - X_i}{h}\right)$$

- ▶ Bias:

$$\text{Bias}[\hat{f}(x; h)] = \frac{1}{2}h^2\mu_2 \sum_{i=1}^d f_d''(x_1, \dots, x_d) + o(h^2)$$

Nonparametrics

Multivariate Density Estimation

- ▶ Consider now the problem of estimating a d -dimensional density $f_0(x)$
- ▶ Define the multivariate kernel density estimator to be,

$$\hat{f}(x; h) = \frac{1}{Nh^d} \sum_{n=1}^N \prod_{i=1}^d K\left(\frac{X_{n,i} - X_i}{h}\right)$$

- ▶ Bias:

$$Bias[\hat{f}(x; h)] = \frac{1}{2}h^2 \mu_2 \sum_{i=1}^d f_d''(x_1, \dots, x_d) + o(h^2)$$

- ▶ Variance:

$$Var(\hat{f}(x; h)) = \frac{1}{h^d N} f(x) \nu_2^d + o(N^{-1}h^{-d})$$

Nonparametrics

Multivariate Density Estimation

► IMSE:

$$\frac{1}{h^d N} \nu_2^d + \frac{1}{4} \mu_2^2 h^4 \int_x \left(\sum_{i=1}^d f_d''(x) \right)^2 dx + o(h^4) + o(N^{-1} h^{-d})$$

Nonparametrics

Multivariate Density Estimation

► IMSE:

$$\frac{1}{h^d N} \nu_2^d + \frac{1}{4} \mu_2^2 h^4 \int_x \left(\sum_{i=1}^d f_d''(x) \right)^2 dx + o(h^4) + o(N^{-1} h^{-d})$$

► FOC for IMSE for optimal bandwidth,

$$h^* = \left(\frac{d \nu_2^d}{\mu_2^2 \int_x \left(\sum_{i=1}^d f_d''(x) \right)^2 dx} \right)^{1/(4+d)} N^{-1/(4+d)}$$

Nonparametrics

Multivariate Density Estimation

► IMSE:

$$\frac{1}{h^d N} \nu_2^d + \frac{1}{4} \mu_2^2 h^4 \int_x \left(\sum_{i=1}^d f_d''(x) \right)^2 dx + o(h^4) + o(N^{-1} h^{-d})$$

► FOC for IMSE for optimal bandwidth,

$$h^* = \left(\frac{d \nu_2^d}{\mu_2^2 \int_x \left(\sum_{i=1}^d f_d''(x) \right)^2 dx} \right)^{1/(4+d)} N^{-1/(4+d)}$$

► Optimal bandwidth yields an IMSE with an error of size $N^{-4/(4+d)}$

Nonparametrics

Multivariate Density Estimation

- ▶ The rate of convergence decreases as d increases (curse of dimensionality!)

Nonparametrics

Multivariate Density Estimation

- ▶ The rate of convergence decreases as d increases (curse of dimensionality!)
- ▶ Curse of dimensionality is not a drawback of KDEs, but a drawback of the nonparametric density estimation problem (i.e. KDEs achieve optimal rates under maintained assumptions about the derivatives of the density)

Nonparametrics

Multivariate Density Estimation

- ▶ The rate of convergence decreases as d increases (curse of dimensionality!)
- ▶ Curse of dimensionality is not a drawback of KDEs, but a drawback of the nonparametric density estimation problem (i.e. KDEs achieve optimal rates under maintained assumptions about the derivatives of the density)
- ▶ No alternative estimator (k-NN, splines, etc.) can do better under maintained assumptions

Nonparametrics

Multivariate Density Estimation

- ▶ The rate of convergence decreases as d increases (curse of dimensionality!)
- ▶ Curse of dimensionality is not a drawback of KDEs, but a drawback of the nonparametric density estimation problem (i.e. KDEs achieve optimal rates under maintained assumptions about the derivatives of the density)
- ▶ No alternative estimator (k-NN, splines, etc.) can do better under maintained assumptions
- ▶ Same problem holds for kernel regression, kernel binary choice, etc.

Nonparametrics

Multivariate Density Estimation

- ▶ The rate of convergence decreases as d increases (curse of dimensionality!)
- ▶ Curse of dimensionality is not a drawback of KDEs, but a drawback of the nonparametric density estimation problem (i.e. KDEs achieve optimal rates under maintained assumptions about the derivatives of the density)
- ▶ No alternative estimator (k-NN, splines, etc.) can do better under maintained assumptions
- ▶ Same problem holds for kernel regression, kernel binary choice, etc.
- ▶ One solution: avoid fully nonparametric problems

Nonparametrics

Take Away Points

- ▶ Purely nonparametric problems are difficult due to curse of dimensionality

Nonparametrics

Take Away Points

- ▶ Purely nonparametric problems are difficult due to curse of dimensionality
- ▶ Best ways to avoid the curse of dimensionality ($N^{-2/(d+4)}$):
 - Focus on a finite dimensional parameter of interest ($N^{-1/2}$)

Nonparametrics

Take Away Points

- ▶ Purely nonparametric problems are difficult due to curse of dimensionality
- ▶ Best ways to avoid the curse of dimensionality ($N^{-2/(d+4)}$):
 - Focus on a finite dimensional parameter of interest ($N^{-1/2}$)
 - Focus on a one-dimensional function of interest ($N^{-2/5}$)

Nonparametrics

Take Away Points

- ▶ Purely nonparametric problems are difficult due to curse of dimensionality
- ▶ Best ways to avoid the curse of dimensionality ($N^{-2/(d+4)}$):
 - Focus on a finite dimensional parameter of interest ($N^{-1/2}$)
 - Focus on a one-dimensional function of interest ($N^{-2/5}$)
- ▶ How would we report high-dimensional functions? (we would end up focusing on low dimensional problems anyway)

Nonparametrics

Take Away Points

- ▶ Every nonparametric problem is different:
 - We can derive large sample approximation, obtain formulas for optimal bandwidth choices, formulas for standard errors, obtain efficiency bounds, one problem at a time

Nonparametrics

Take Away Points

- ▶ Every nonparametric problem is different:
 - We can derive large sample approximation, obtain formulas for optimal bandwidth choices, formulas for standard errors, obtain efficiency bounds, one problem at a time
 - Better solution is to focus on methods which most easily generalize (cross validation and the bootstrap)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates
- ▶ “Fully” nonparametric estimation
 - In the one dimensional case, convergence was slower than (\sqrt{N})

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates
- ▶ “Fully” nonparametric estimation
 - In the one dimensional case, convergence was slower than (\sqrt{N})
 - In higher dimensions, convergence became slower and slower ($N^{2/(4+d)}$)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates
- ▶ “Fully” nonparametric estimation
 - In the one dimensional case, convergence was slower than (\sqrt{N})
 - In higher dimensions, convergence became slower and slower ($N^{2/(4+d)}$)
- ▶ $y_n = g_0(x_n) + \varepsilon_n$ (nonparametric regression)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates
- ▶ “Fully” nonparametric estimation
 - In the one dimensional case, convergence was slower than (\sqrt{N})
 - In higher dimensions, convergence became slower and slower ($N^{2/(4+d)}$)
- ▶ $y_n = g_0(x_n) + \varepsilon_n$ (nonparametric regression)
- ▶ $\Pr(y_n = 1|x_n) = G_0(x_n)$ (nonparametric binary choice)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ Some “easy” semiparametric estimators that did not require estimating an infinite dimensional quantity (e.g. OLS w/ heteroskedasticity)
- ▶ These estimators retain parametric (\sqrt{N}) convergence rates
- ▶ “Fully” nonparametric estimation
 - In the one dimensional case, convergence was slower than (\sqrt{N})
 - In higher dimensions, convergence became slower and slower ($N^{2/(4+d)}$)
- ▶ $y_n = g_0(x_n) + \varepsilon_n$ (nonparametric regression)
- ▶ $\Pr(y_n = 1|x_n) = G_0(x_n)$ (nonparametric binary choice)
- ▶ In both cases, curse of dimensionality ($N^{2/(4+d)}$)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ In some cases, we can estimate the parameter of interest at the parametric rate (\sqrt{N}), but require estimating an infinite dimensional quantity in the process (semiparametric estimation)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ In some cases, we can estimate the parameter of interest at the parametric rate (\sqrt{N}), but require estimating an infinite dimensional quantity in the process (semiparametric estimation)
 - $\Pr(y_n = 1|x_n) = G_0(\beta_0'x_n)$ (semiparametric binary choice)

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ In some cases, we can estimate the parameter of interest at the parametric rate (\sqrt{N}), but require estimating an infinite dimensional quantity in the process (semiparametric estimation)
 - $\Pr(y_n = 1|x_n) = G_0(\beta_0'x_n)$ (semiparametric binary choice)
 - $y_n = g_0(x_n, t_n) + \varepsilon_n, ATE = E[g_0(x_n, 1) - g_0(x_n, 0)]$

Nonparametrics

Semiparametric and Nonparametric Models

- ▶ In some cases, we can estimate the parameter of interest at the parametric rate (\sqrt{N}), but require estimating an infinite dimensional quantity in the process (semiparametric estimation)
 - $\Pr(y_n = 1|x_n) = G_0(\beta'_0 x_n)$ (semiparametric binary choice)
 - $y_n = g_0(x_n, t_n) + \varepsilon_n$, $ATE = E[g_0(x_n, 1) - g_0(x_n, 0)]$
- ▶ In other cases, the parameter of interest will be infinite dimensional, but we will control for nuisance variables using a parametric component
 - $y_n = g_0(x_n) + \beta'_0 z_n + \varepsilon_n$ (partially linear model)

Nonparametrics

Kernel Regression

- ▶ Consider the relationship, $y_n = g_0(x_n) + \varepsilon_n$, where (x_n, ε_n) are iid and $E[\varepsilon_n | x_n] = 0$.

Nonparametrics

Kernel Regression

- ▶ Consider the relationship, $y_n = g_0(x_n) + \varepsilon_n$, where (x_n, ε_n) are iid and $E[\varepsilon_n|x_n] = 0$.
- ▶ The (locally constant) Kernel estimator is defined by,

$$\hat{g}(x; h) = \frac{1}{N} \sum_{n=1}^N w_n(x; h) y_n$$

$$\text{where } w_n(x; h) = \frac{\frac{1}{h} K\left(\frac{x-x_n}{h}\right)}{\frac{1}{Nh} \sum_{m=1}^N K\left(\frac{x-x_m}{h}\right)}$$

Nonparametrics

Kernel Regression

- ▶ Consider the relationship, $y_n = g_0(x_n) + \varepsilon_n$, where (x_n, ε_n) are iid and $E[\varepsilon_n|x_n] = 0$.
- ▶ The (locally constant) Kernel estimator is defined by,

$$\hat{g}(x; h) = \frac{1}{N} \sum_{n=1}^N w_n(x; h) y_n$$

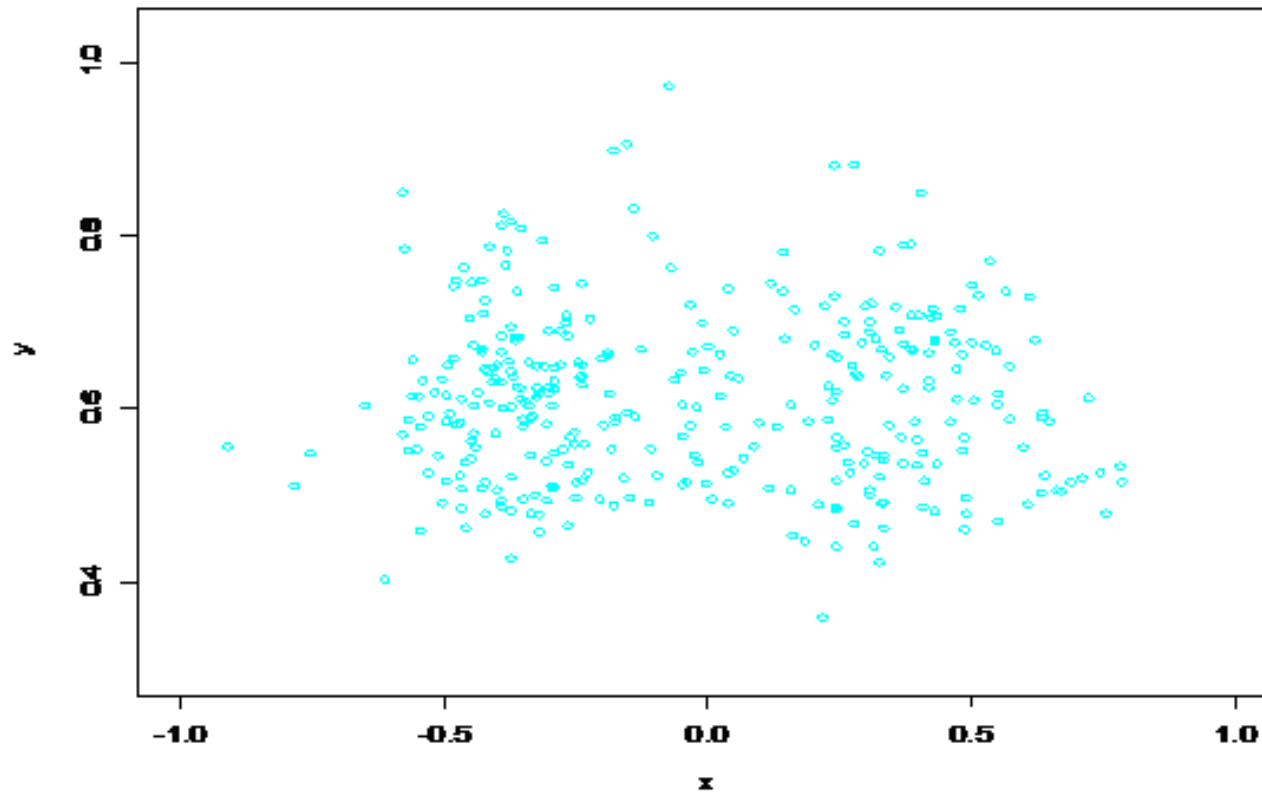
where $w_n(x; h) = \frac{\frac{1}{h} K\left(\frac{x-x_n}{h}\right)}{\frac{1}{Nh} \sum_{m=1}^N K\left(\frac{x-x_m}{h}\right)}$

- ▶ Motivation: to evaluate $E[y_n|x]$, look at average value of y_n for x_n 's that are close to x (weighted by their closeness)

Nonparametrics

Kernel Regression

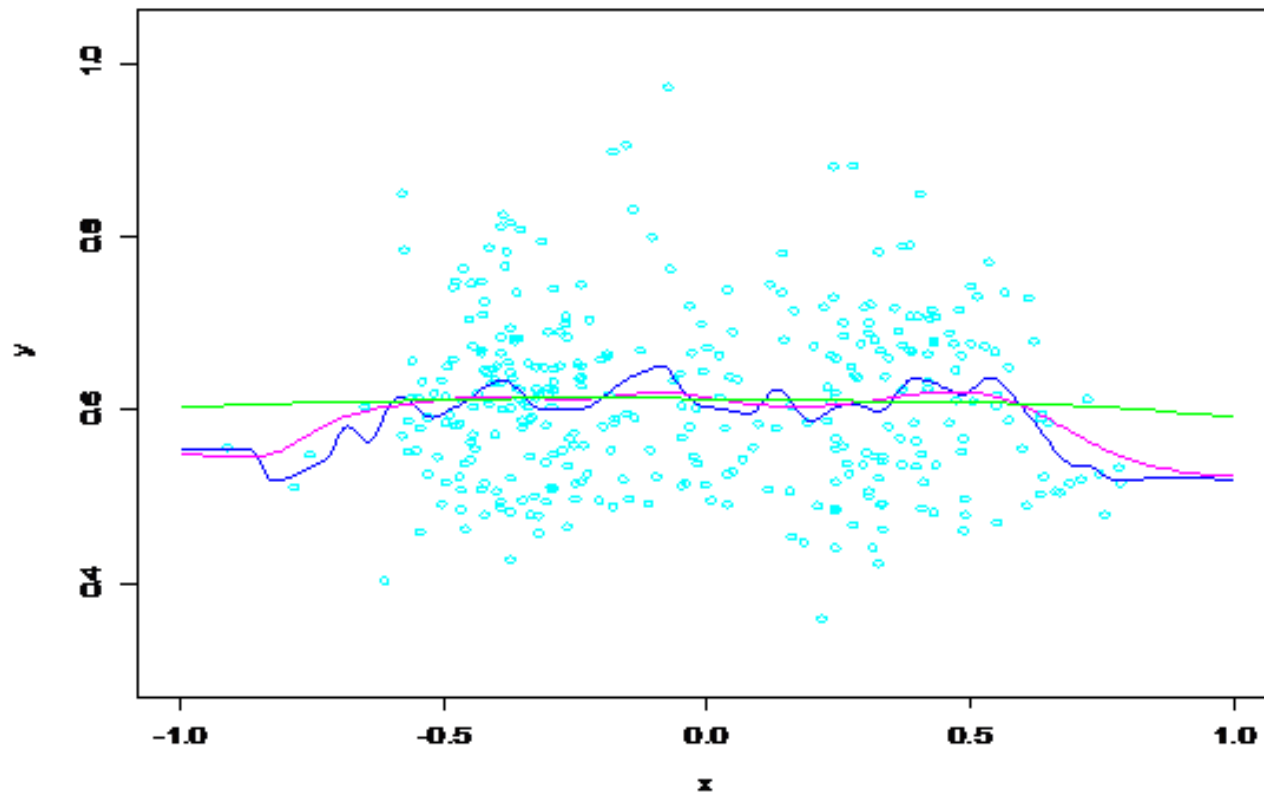
- ▶ Example: Effect of Position on Vote Share for Senate Incumbents



Nonparametrics

Kernel Regression

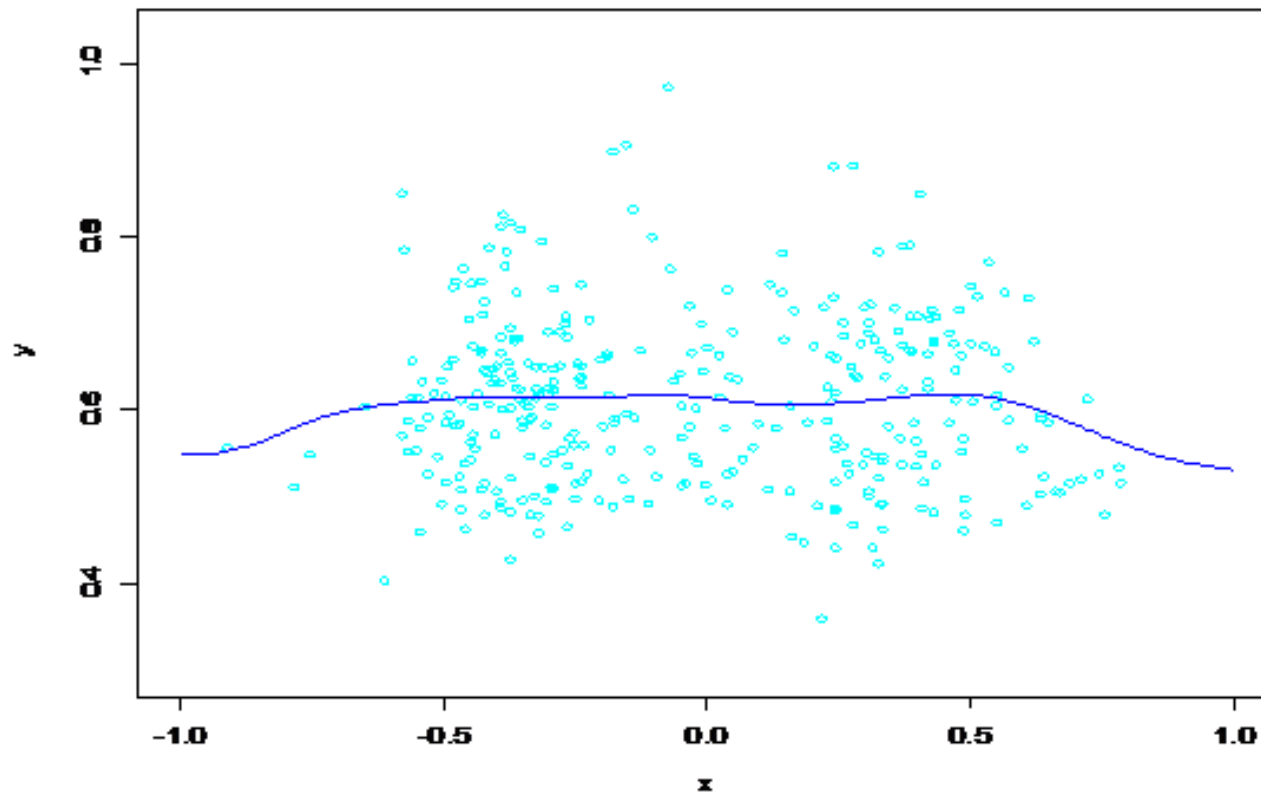
- ▶ Example: Effect of Position on Vote Share for Senate Incumbents ($h = .03$, $h = .01$, and $h = .3$)



Nonparametrics

Kernel Regression

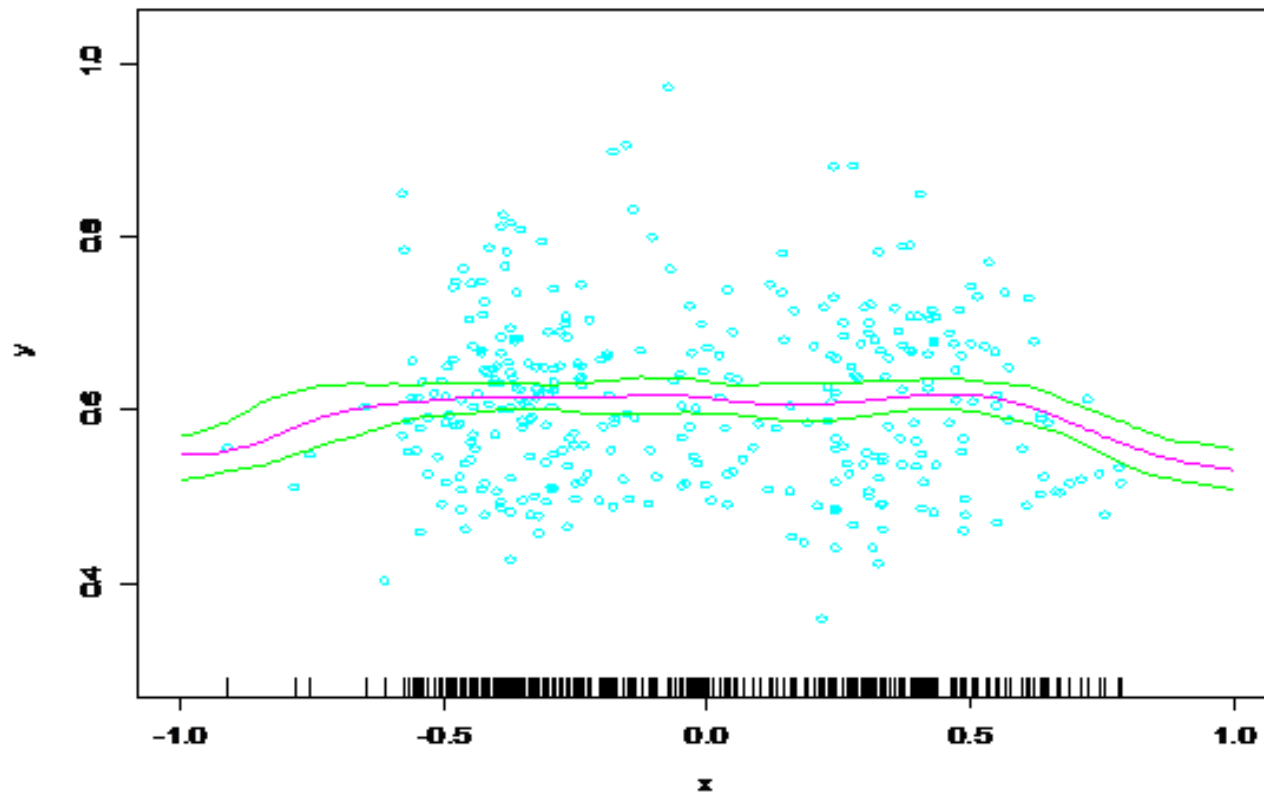
- ▶ Example: Effect of Position on Vote Share for Senate Incumbents ($h_{ROT} = 0.126$)



Nonparametrics

Kernel Regression

- ▶ Example: Effect of Position on Vote Share for Senate Incumbents w/ Bootstrapped Standard Errors:



Nonparametrics

Multivariate Kernel Regression

- ▶ Consider the relationship, $y_n = g_0(x_n) + \varepsilon_n$, where (x_n, ε_n) are iid and $E[\varepsilon_n | x_n] = 0$

Nonparametrics

Multivariate Kernel Regression

- ▶ Consider the relationship, $y_n = g_0(x_n) + \varepsilon_n$, where (x_n, ε_n) are iid and $E[\varepsilon_n|x_n] = 0$
- ▶ The Kernel estimator is defined by,

$$\hat{g}(x; h) = \frac{1}{N} \sum_{n=1}^N w_n(x; h) y_n$$

where,

$$w_n(x; h) = \frac{\frac{1}{h^d} \prod_{i=1}^d K\left(\frac{x_i - x_{n,i}}{h}\right)}{\frac{1}{Nh^d} \sum_{m=1}^N \prod_{i=1}^d K\left(\frac{x_i - x_{m,i}}{h}\right)}$$

Nonparametrics

Kernel Binary Choice

- ▶ Consider the relationship, $\Pr(y_n = 1|x_n) = G_0(x_n)$

Nonparametrics

Kernel Binary Choice

- ▶ Consider the relationship, $\Pr(y_n = 1|x_n) = G_0(x_n)$
- ▶ The Kernel estimator is defined by,

$$\hat{G}(x; h) = \frac{1}{N} \sum_{n=1}^N w_n(x; h) y_n$$

where,

$$w_n(x; h) = \frac{\frac{1}{h^d} \prod_{i=1}^d K\left(\frac{x_i - x_{n,i}}{h}\right)}{\frac{1}{Nh^d} \sum_{m=1}^N \prod_{i=1}^d K\left(\frac{x_i - x_{m,i}}{h}\right)}$$

Nonparametrics

Kernel Binary Choice

- ▶ Consider the relationship, $\Pr(y_n = 1|x_n) = G_0(x_n)$
- ▶ The Kernel estimator is defined by,

$$\hat{G}(x; h) = \frac{1}{N} \sum_{n=1}^N w_n(x; h) y_n$$

where,

$$w_n(x; h) = \frac{\frac{1}{h^d} \prod_{i=1}^d K\left(\frac{x_i - x_{n,i}}{h}\right)}{\frac{1}{Nh^d} \sum_{m=1}^N \prod_{i=1}^d K\left(\frac{x_i - x_{m,i}}{h}\right)}$$

- ▶ Notice that this is the same estimator as the Kernel regression estimator

Nonparametrics

Semiparametric Binary Choice

- ▶ Parametric binary choice (i.e. probit)

$$\Pr(y_n = 1|x_n) = \Phi(\beta' x_n)$$

- ▶ Nonparametric binary choice

$$\Pr(y = 1|x) = G_0(x)$$

- ▶ Semiparametric binary choice

$$\Pr(y_n = 1|x_n) = G_0(\beta' x_n)$$

Nonparametrics

Semiparametric Binary Choice

► Why consider semiparametric binary choice model?

- Parametric binary choice (i.e. probit)

$$\Pr(y_n = 1|x_n) = \Phi(\beta' x_n)$$

- Marginal effect of x_k

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k \phi(\beta' x)$$

Nonparametrics

Semiparametric Binary Choice

► Why consider semiparametric binary choice model?

- Parametric binary choice (i.e. probit)

$$\Pr(y_n = 1|x_n) = \Phi(\beta' x_n)$$

- Marginal effect of x_k

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k \phi(\beta' x)$$

- Magnitude of $\beta_k \phi(\beta' x)$ is largest when $\beta' x = 0$

Nonparametrics

Semiparametric Binary Choice

- ▶ Why consider semiparametric binary choice model?
 - Parametric binary choice (i.e. probit)

$$\Pr(y_n = 1|x_n) = \Phi(\beta'x_n)$$

- Marginal effect of x_k

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k \phi(\beta'x)$$

- Magnitude of $\beta_k \phi(\beta'x)$ is largest when $\beta'x = 0$
- ▶ Same will hold for any symmetric unimodal density

Nonparametrics

Semiparametric Binary Choice

- ▶ Why consider semiparametric binary choice model?
 - Parametric binary choice (i.e. probit)

$$\Pr(y_n = 1|x_n) = \Phi(\beta'x_n)$$

- Marginal effect of x_k

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k \phi(\beta'x)$$

- Magnitude of $\beta_k \phi(\beta'x)$ is largest when $\beta'x = 0$
- ▶ Same will hold for any symmetric unimodal density
- ▶ Fully nonparametric model is too general (hard to report results) and suffers from curse of dimensionality

Nonparametrics

Semiparametric Binary Choice

- For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta' x)$$

Nonparametrics

Semiparametric Binary Choice

- ▶ For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta'x)$$

- ▶ If g is not symmetric, then $\frac{\partial}{\partial x_k} \Pr(y_n = 1|x)$ need not peak when $\beta'x = 0$

Nonparametrics

Semiparametric Binary Choice

- ▶ For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta'x)$$

- ▶ If g is not symmetric, then $\frac{\partial}{\partial x_k} \Pr(y_n = 1|x)$ need not peak when $\beta'x = 0$
- ▶ In an application to campaigning, this assumption implies that moderate voters are most sensitive to campaigning (maybe)

Nonparametrics

Semiparametric Binary Choice

- ▶ For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta' x)$$

- ▶ If g is not symmetric, then $\frac{\partial}{\partial x_k} \Pr(y_n = 1|x)$ need not peak when $\beta' x = 0$
- ▶ In an application to campaigning, this assumption implies that moderate voters are most sensitive to campaigning (maybe)
- ▶ In an application to GOTV, this implies that voters with a predicted probability of voting of 0.5 are most sensitive to GOTV (maybe)

Nonparametrics

Semiparametric Binary Choice

- ▶ For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta' x)$$

- ▶ If g is not symmetric, then $\frac{\partial}{\partial x_k} \Pr(y_n = 1|x)$ need not peak when $\beta' x = 0$
- ▶ In an application to campaigning, this assumption implies that moderate voters are most sensitive to campaigning (maybe)
- ▶ In an application to GOTV, this implies that voters with a predicted probability of voting of 0.5 are most sensitive to GOTV (maybe)
- ▶ These are assumptions!

Nonparametrics

Semiparametric Binary Choice

- ▶ For semiparametric binary choice model,

$$\frac{\partial}{\partial x_k} \Pr(y_n = 1|x) = \beta_k g(\beta' x)$$

- ▶ If g is not symmetric, then $\frac{\partial}{\partial x_k} \Pr(y_n = 1|x)$ need not peak when $\beta' x = 0$
- ▶ In an application to campaigning, this assumption implies that moderate voters are most sensitive to campaigning (maybe)
- ▶ In an application to GOTV, this implies that voters with a predicted probability of voting of 0.5 are most sensitive to GOTV (maybe)
- ▶ These are assumptions!
- ▶ Semiparametric binary choice model allows us to relax/test these assumptions

Nonparametrics

Semiparametric Binary Choice

- ▶ The Semiparametric Kernel Estimator:
 - Define $z_n = \beta' x_n$

Nonparametrics

Semiparametric Binary Choice

► The Semiparametric Kernel Estimator:

- Define $z_n = \beta' x_n$
- If we knew β , we would have $\Pr(y_n = 1 | z_n) = G_0(z_n)$

Nonparametrics

Semiparametric Binary Choice

► The Semiparametric Kernel Estimator:

- Define $z_n = \beta' x_n$
- If we knew β , we would have $\Pr(y_n = 1 | z_n) = G_0(z_n)$
- We can form,

$$\hat{G}(z) = \hat{P}(y_n = 1 | z) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n - z}{h}\right) y_n}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n - z}{h}\right)}$$

Nonparametrics

Semiparametric Binary Choice

- Approach we use embeds kernel estimator in log-likelihood

$$\hat{\beta} = \arg \min_{\beta: \beta_0=0, \beta_1=1} \frac{1}{N} \sum_{n=1}^N y_n \log \hat{G}(\beta' x_n; \beta) + (1 - y_n) \log(1 - \hat{G}(\beta' x_n; \beta))$$

where,

$$z_n(\beta) = \beta' x_n$$

$$\hat{G}(z; \beta) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n(\beta) - z}{h}\right) y_n}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n(\beta) - z}{h}\right)}$$

Nonparametrics

Semiparametric Binary Choice

- Approach we use embeds kernel estimator in log-likelihood

$$\hat{\beta} = \arg \min_{\beta: \beta_0=0, \beta_1=1} \frac{1}{N} \sum_{n=1}^N y_n \log \hat{G}(\beta' x_n; \beta) + (1 - y_n) \log(1 - \hat{G}(\beta' x_n; \beta))$$

where,

$$z_n(\beta) = \beta' x_n$$

$$\hat{G}(z; \beta) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n(\beta) - z}{h}\right) y_n}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{z_n(\beta) - z}{h}\right)}$$

- We must impose the restrictions $\beta_1 = 0$ and $\beta_2 = 1$ for identification

Nonparametrics

Semiparametric Binary Choice

- ▶ The estimator is \sqrt{N} -consistent and asymptotically normal for β_0

Nonparametrics

Semiparametric Binary Choice

- ▶ The estimator is \sqrt{N} -consistent and asymptotically normal for β_0
- ▶ Large sample properties of this (and many other estimators) follow from Andrews's (1994) MINPIN theorem:

$$\hat{\beta} \in \arg \max_{\beta \in B} \frac{1}{N} \sum_{n=1}^N \psi(x_n, y_n; \beta, \hat{G}(\beta))$$

Nonparametrics

Semiparametric Binary Choice

► We have,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{dist.} N(0, C^{-1}BC^{-1})$$

where,

$$C = E \left[\frac{\partial^2 \psi}{\partial \beta^2} (x_n, y_n; \beta_0, G_0) \right]$$

$$B = E \left[\frac{\partial \psi}{\partial \beta} (x_n, y_n; \beta_0, G_0) \frac{\partial \psi}{\partial \beta} (x_n, y_n; \beta_0, G_0)' \right]$$

Nonparametrics

Semiparametric Binary Choice

- Where can then estimate,

$$\hat{C} = \frac{1}{N} \sum_{n=1}^N \frac{\partial^2 \psi}{\partial \beta^2} (x_n, y_n; \hat{\beta}, \hat{G})$$

$$\hat{B} = \frac{1}{N} \sum_{n=1}^N \frac{\partial \psi}{\partial \beta} (x_n, y_n; \hat{\beta}, \hat{G}) \frac{\partial \psi}{\partial \beta} (x_n, y_n; \hat{\beta}, \hat{G})'$$

Nonparametrics

Semiparametric Binary Choice

- ▶ Where can then estimate,

$$\hat{C} = \frac{1}{N} \sum_{n=1}^N \frac{\partial^2 \psi}{\partial \beta^2} (x_n, y_n; \hat{\beta}, \hat{G})$$

$$\hat{B} = \frac{1}{N} \sum_{n=1}^N \frac{\partial \psi}{\partial \beta} (x_n, y_n; \hat{\beta}, \hat{G}) \frac{\partial \psi}{\partial \beta} (x_n, y_n; \hat{\beta}, \hat{G})'$$

- ▶ Alternatively, we can use the bootstrap to conduct inference about β_0 and G_0 (especially if G_0 is of direct interest)

Nonparametrics

Semiparametric Binary Choice

- ▶ Selecting the Bandwidth:
 - “Lazy” rule of thumb:
 - ▶ For each β , form $z_n = \beta' x_n$

Nonparametrics

Semiparametric Binary Choice

- ▶ Selecting the Bandwidth:

- “Lazy” rule of thumb:

- ▶ For each β , form $z_n = \beta' x_n$

- ▶ Compute h based on normal reference rule for the density of z_n (notice that there is a different h each time the objective function is evaluated at β)

Nonparametrics

Semiparametric Binary Choice

▶ Selecting the Bandwidth:

- “Lazy” rule of thumb:

- ▶ For each β , form $z_n = \beta' x_n$
- ▶ Compute h based on normal reference rule for the density of z_n (notice that there is a different h each time the objective function is evaluated at β)
- ▶ This is an ad-hoc rule, but at least you get the rates correct (i.e. $h = cN^{-1/5}$)

Nonparametrics

Semiparametric Binary Choice

▶ Selecting the Bandwidth:

- “Lazy” rule of thumb:

- ▶ For each β , form $z_n = \beta' x_n$
- ▶ Compute h based on normal reference rule for the density of z_n (notice that there is a different h each time the objective function is evaluated at β)
- ▶ This is an ad-hoc rule, but at least you get the rates correct (i.e. $h = cN^{-1/5}$)

- Cross-validation:

$$\frac{1}{N} \sum_{n=1}^N y_n \log \hat{G}_{(n)}(z_n; h) + (1 - y_n) \log(1 - \hat{G}_{(n)}(z_n; h))$$

where $\hat{G}_{(n)}$ is the leave-one-out estimator

Nonparametrics

Semiparametric Binary Choice

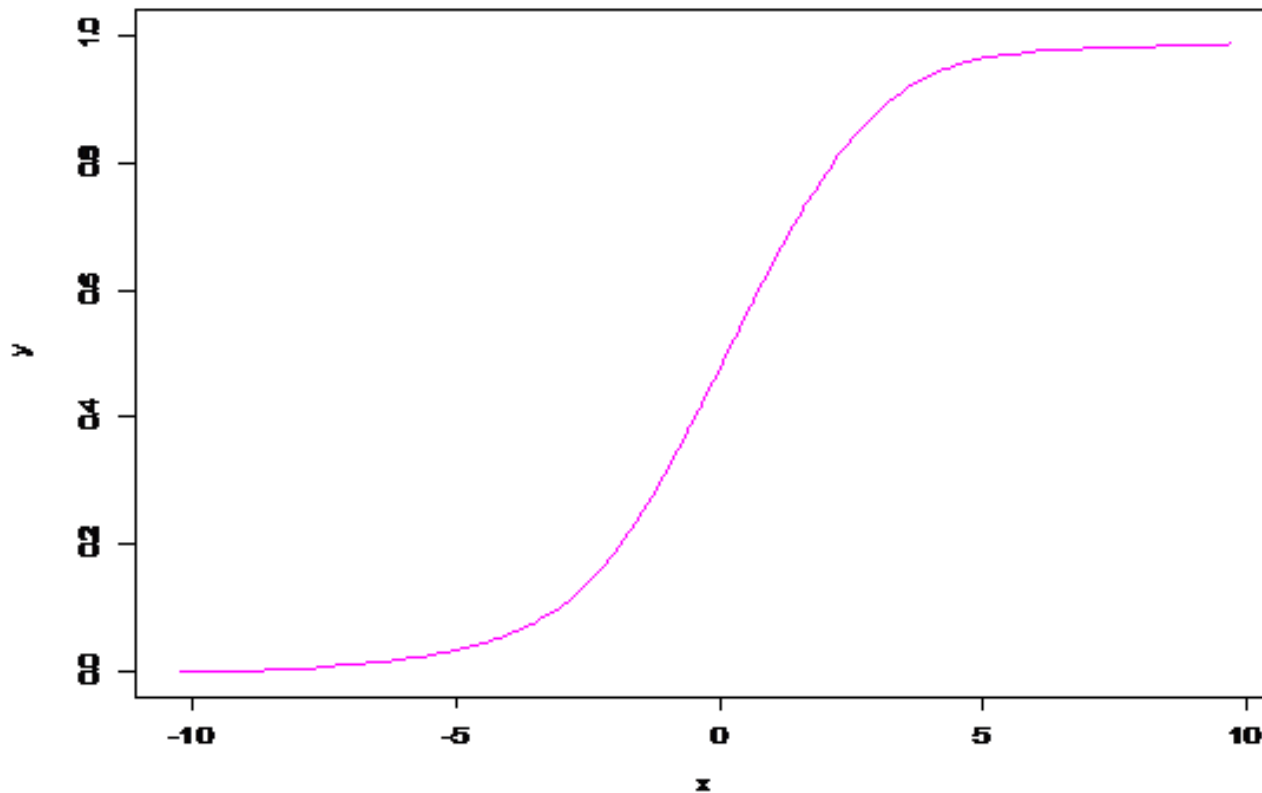
- ▶ Example: Semiparametric Model of Presidential Vote in 2004 Election (coefficient estimates)

	Probit		Probit (Normalized)		Semiparametric (Normalized)	
	est.	se	est.	est.	se	boot se
(Intercept)	-0.056	(0.146)	0.000	0.000		
Prox. Diff.	0.393	(0.044)	1.000	1.000		
Party Dem.	-0.967	(0.228)	-2.460	-2.341	(0.949)	[0.811]
Party Rep.	1.049	(0.210)	2.668	2.766	(0.652)	[0.699]
Black	-1.336	(0.361)	-3.398	-3.374	(1.022)	[0.992]
Female	0.161	(0.171)	0.411	0.427	(0.424)	[0.452]
South	0.227	(0.196)	0.577	0.553	(0.459)	[0.485]

Nonparametrics

Semiparametric Binary Choice

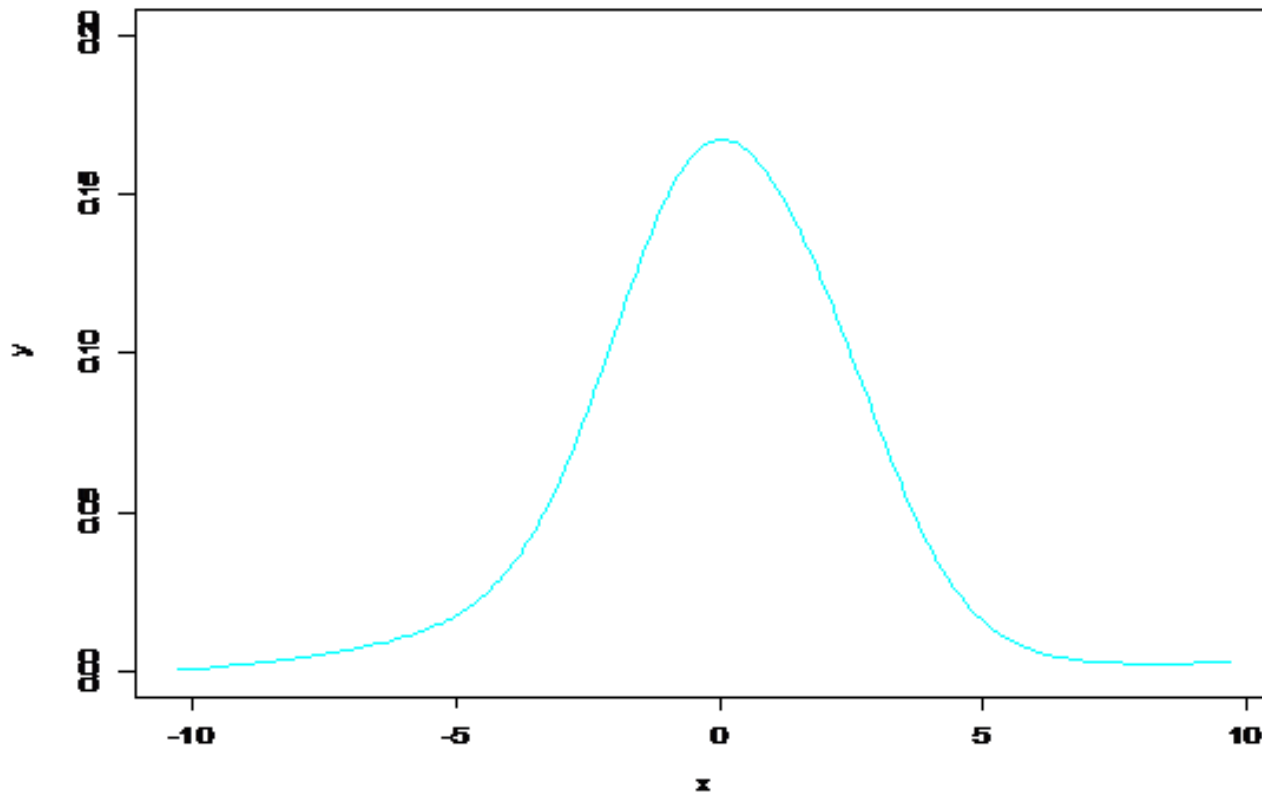
- ▶ Example: Semiparametric Model of Presidential Vote in 2004 Election (estimate of G)



Nonparametrics

Semiparametric Binary Choice

- ▶ Example: Semiparametric Model of Presidential Vote in 2004 Election (estimate of g)



Nonparametrics

Partially Linear Models

- ▶ Kernel regression estimators are of limited use on their own because most social science applications involve multiple explanatory variables

Nonparametrics

Partially Linear Models

- ▶ Kernel regression estimators are of limited use on their own because most social science applications involve multiple explanatory variables
- ▶ As with binary choice, fully nonparametric approach suffers from the curse of dimensionality

Nonparametrics

Partially Linear Models

- ▶ Kernel regression estimators are of limited use on their own because most social science applications involve multiple explanatory variables
- ▶ As with binary choice, fully nonparametric approach suffers from the curse of dimensionality
- ▶ Partially linear model provides a way of having multiple regressors with one degree of nonparametrics

$$y_n = \beta_0' z_n + g_0(x_n) + \varepsilon_n$$

Nonparametrics

Partially Linear Models

- ▶ Kernel regression estimators are of limited use on their own because most social science applications involve multiple explanatory variables
- ▶ As with binary choice, fully nonparametric approach suffers from the curse of dimensionality
- ▶ Partially linear model provides a way of having multiple regressors with one degree of nonparametrics

$$y_n = \beta_0' z_n + g_0(x_n) + \varepsilon_n$$

- ▶ If β_0 is of interest and g_0 is a nuisance parameter, we have a semi-parametric model

Nonparametrics

Partially Linear Models

- ▶ Kernel regression estimators are of limited use on their own because most social science applications involve multiple explanatory variables
- ▶ As with binary choice, fully nonparametric approach suffers from the curse of dimensionality
- ▶ Partially linear model provides a way of having multiple regressors with one degree of nonparametrics

$$y_n = \beta_0' z_n + g_0(x_n) + \varepsilon_n$$

- ▶ If β_0 is of interest and g_0 is a nuisance parameter, we have a semi-parametric model
- ▶ If g_0 is of interest and β_0 is a nuisance parameter, we have a low-dimensional nonparametric model (picture w/ controls)

Nonparametrics

Partially Linear Models

- ▶ If we knew the value of β , we could define $w_n = y_n - \beta' z_n$ yielding $w_n = g_0(x_n) + \varepsilon_n$

Nonparametrics

Partially Linear Models

- ▶ If we knew the value of β , we could define $w_n = y_n - \beta' z_n$ yielding $w_n = g_0(x_n) + \varepsilon_n$
- ▶ Applying the Kernel regression estimator:

$$\hat{g}(x; \beta) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right) (y_n - \beta' z_n)}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right)}$$

Nonparametrics

Partially Linear Models

- ▶ If we knew the value of β , we could define $w_n = y_n - \beta' z_n$ yielding $w_n = g_0(x_n) + \varepsilon_n$
- ▶ Applying the Kernel regression estimator:

$$\hat{g}(x; \beta) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right) (y_n - \beta' z_n)}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right)}$$

- ▶ We can plug this estimator into the equation above to obtain $y_n - \hat{g}(x_n; \beta) = \beta' z_n + \varepsilon_n$

Nonparametrics

Partially Linear Models

- ▶ If we knew the value of β , we could define $w_n = y_n - \beta' z_n$ yielding $w_n = g_0(x_n) + \varepsilon_n$
- ▶ Applying the Kernel regression estimator:

$$\hat{g}(x; \beta) = \frac{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right) (y_n - \beta' z_n)}{\frac{1}{hN} \sum_{n=1}^N K\left(\frac{x_n - x}{h}\right)}$$

- ▶ We can plug this estimator into the equation above to obtain $y_n - \hat{g}(x_n; \beta) = \beta' z_n + \varepsilon_n$
- ▶ We can then estimate β_0 using least squares,

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{N} \sum_{n=1}^N (y_n - \hat{g}(x_n; \beta) - \beta' z_n)^2$$

Nonparametrics

Partially Linear Models

- For this model (unlike the semiparametric binary choice model) we can apply some computational tricks:

$$\tilde{y}_n = y_n - \frac{\frac{1}{hN} \sum_{m=1}^N K\left(\frac{x_m - x_n}{h}\right) y_m}{\frac{1}{hN} \sum_{m=1}^N K\left(\frac{x_m - x_n}{h}\right)}$$

$$\tilde{z}_n = z_n - \frac{\frac{1}{hN} \sum_{m=1}^N K\left(\frac{x_m - x_n}{h}\right) z_m}{\frac{1}{hN} \sum_{m=1}^N K\left(\frac{x_m - x_n}{h}\right)}$$

Nonparametrics

Partially Linear Models

► We have that,

$$\hat{\beta} = \left[\frac{1}{N} \sum_{n=1}^N \tilde{z}_n \tilde{z}_n' \right]^{-1} \left[\frac{1}{N} \sum_{n=1}^N \tilde{z}_n \tilde{y}_n \right]$$

Nonparametrics

Partially Linear Models

- ▶ We have that,

$$\hat{\beta} = \left[\frac{1}{N} \sum_{n=1}^N \tilde{z}_n \tilde{z}_n' \right]^{-1} \left[\frac{1}{N} \sum_{n=1}^N \tilde{z}_n \tilde{y}_n \right]$$

- ▶ Application: estimate effect of incumbent position on senate vote share with controls

Nonparametrics

Average Treatment Effects

► $y_n = \alpha'x_n + \beta t_n + \varepsilon_n, E[\varepsilon_n|x_n, t_n] = 0$

Nonparametrics

Average Treatment Effects

- ▶ $y_n = \alpha'x_n + \beta t_n + \varepsilon_n, E[\varepsilon_n|x_n, t_n] = 0$
- ▶ $ATE = E[y_n|t_n = 1] - E[y_n|t_n = 0] = E[E[y_n|x_n, t_n = 1]] - E[E[y_n|x_n, t_n = 0]] = E[\alpha'x_n] + \beta - E[\alpha'x_n] = \beta$

Nonparametrics

Average Treatment Effects

- ▶ $y_n = \alpha'x_n + \beta t_n + \varepsilon_n, E[\varepsilon_n|x_n, t_n] = 0$
- ▶ $ATE = E[y_n|t_n = 1] - E[y_n|t_n = 0] = E[E[y_n|x_n, t_n = 1]] - E[E[y_n|x_n, t_n = 0]] = E[\alpha'x_n] + \beta - E[\alpha'x_n] = \beta$
- ▶ $\widehat{ATE} = \widehat{TE} = \hat{\beta}$

Nonparametrics

Average Treatment Effects

► $y_n = g_0(x_n, t_n) + \varepsilon_n, E[\varepsilon_n | x_n, t_n] = 0$

Nonparametrics

Average Treatment Effects

- ▶ $y_n = g_0(x_n, t_n) + \varepsilon_n, E[\varepsilon_n | x_n, t_n] = 0$
- ▶ $ATE = E[y_n | t_n = 1] - E[y_n | t_n = 0] = E[g_0(x_n, 1) - g_0(x_n, 0)]$

Nonparametrics

Average Treatment Effects

- ▶ $y_n = g_0(x_n, t_n) + \varepsilon_n, E[\varepsilon_n | x_n, t_n] = 0$
- ▶ $ATE = E[y_n | t_n = 1] - E[y_n | t_n = 0] = E[g_0(x_n, 1) - g_0(x_n, 0)]$
- ▶ $\widehat{ATE} = \frac{1}{N} \sum_{n=1}^N (\hat{g}(x_n, 1) - \hat{g}(x_n, 0))$

Nonparametrics

Average Treatment Effects

- ▶ $y_n = g_0(x_n, t_n) + \varepsilon_n, E[\varepsilon_n | x_n, t_n] = 0$
- ▶ $ATE = E[y_n | t_n = 1] - E[y_n | t_n = 0] = E[g_0(x_n, 1) - g_0(x_n, 0)]$
- ▶ $\widehat{ATE} = \frac{1}{N} \sum_{n=1}^N (\hat{g}(x_n, 1) - \hat{g}(x_n, 0))$
- ▶ $\hat{g}(x, 1) = \frac{\frac{1}{N} \sum_{n=1}^N K\left(\frac{x-x_n}{h}\right) t_n y_n}{\frac{1}{N} \sum_{n=1}^N K\left(\frac{x-x_n}{h}\right) t_n}$
- ▶ $\hat{g}(x, 0) = \frac{\frac{1}{N} \sum_{n=1}^N K\left(\frac{x-x_n}{h}\right) (1-t_n) y_n}{\frac{1}{N} \sum_{n=1}^N K\left(\frac{x-x_n}{h}\right) (1-t_n)}$

Nonparametrics

k-Nearest Neighbor Estimator

- ▶ Consider the multivariate nonparametric regression problem:

$$y_n = g_0(x_n) + \varepsilon_n$$

Nonparametrics

k-Nearest Neighbor Estimator

- ▶ Consider the multivariate nonparametric regression problem:

$$y_n = g_0(x_n) + \varepsilon_n$$

- ▶ The k-NN estimator is given by,

$$\hat{g}(x; k) = \frac{1}{k} \sum_{n=1}^N 1_{nk}(x) y_n$$

where $I_{nk} = 1 \Leftrightarrow x_n$ is one of the k closest points to x

Nonparametrics

k-Nearest Neighbor Estimator

- ▶ Consider the multivariate nonparametric regression problem:

$$y_n = g_0(x_n) + \varepsilon_n$$

- ▶ The k-NN estimator is given by,

$$\hat{g}(x; k) = \frac{1}{k} \sum_{n=1}^N 1_{nk}(x) y_n$$

where $I_{nk} = 1 \Leftrightarrow x_n$ is one of the k closest points to x

- ▶ Issues:
 - Selecting k

Nonparametrics

k-Nearest Neighbor Estimator

- ▶ Consider the multivariate nonparametric regression problem:

$$y_n = g_0(x_n) + \varepsilon_n$$

- ▶ The k-NN estimator is given by,

$$\hat{g}(x; k) = \frac{1}{k} \sum_{n=1}^N 1_{nk}(x) y_n$$

where $I_{nk} = 1 \Leftrightarrow x_n$ is one of the k closest points to x

- ▶ Issues:
 - Selecting k
 - Computation

Nonparametrics

Sieve estimator

► $\hat{g}(x) = \sum_{i=1}^m a_i h_i(x)$, where $\{h_i(x)\}_{i=1}^{\infty}$ are basis functions

Nonparametrics

Sieve estimator

- ▶ $\hat{g}(x) = \sum_{i=1}^m a_i h_i(x)$, where $\{h_i(x)\}_{i=1}^{\infty}$ are basis functions
- ▶ For example, if $h_i(x) = x^i$, we have $\hat{g}_0(x) = \sum_{i=1}^m a_i x^i$

Nonparametrics

Sieve estimator

- ▶ $\hat{g}(x) = \sum_{i=1}^m a_i h_i(x)$, where $\{h_i(x)\}_{i=1}^{\infty}$ are basis functions
- ▶ For example, if $h_i(x) = x^i$, we have $\hat{g}_0(x) = \sum_{i=1}^m a_i x^i$
- ▶ Issues:
 - Selecting m

Nonparametrics

Sieve estimator

- ▶ $\hat{g}(x) = \sum_{i=1}^m a_i h_i(x)$, where $\{h_i(x)\}_{i=1}^{\infty}$ are basis functions
- ▶ For example, if $h_i(x) = x^i$, we have $\hat{g}_0(x) = \sum_{i=1}^m a_i x^i$
- ▶ Issues:
 - Selecting m
 - Becomes very complicated in higher dimensions

Nonparametrics

Smoothing Splines

- Solve problem,

$$\hat{g} = \arg \max_g \frac{1}{N} \sum_{n=1}^N (y_n - g(x_n))^2 - \lambda \int_x (g''(x))^2 dx$$

Nonparametrics

Smoothing Splines

- ▶ Solve problem,

$$\hat{g} = \arg \max_g \frac{1}{N} \sum_{n=1}^N (y_n - g(x_n))^2 - \lambda \int_x (g''(x))^2 dx$$

- ▶ Solution is a cubic spline with knots at all the data points

Nonparametrics

Smoothing Splines

- ▶ Solve problem,

$$\hat{g} = \arg \max_g \frac{1}{N} \sum_{n=1}^N (y_n - g(x_n))^2 - \lambda \int_x (g''(x))^2 dx$$

- ▶ Solution is a cubic spline with knots at all the data points
- ▶ Computation involves linear algebra

Nonparametrics

Smoothing Splines

- ▶ Solve problem,

$$\hat{g} = \arg \max_g \frac{1}{N} \sum_{n=1}^N (y_n - g(x_n))^2 - \lambda \int_x (g''(x))^2 dx$$

- ▶ Solution is a cubic spline with knots at all the data points
- ▶ Computation involves linear algebra
- ▶ Easy to impose shape restrictions (i.e. monotonicity) – becomes quadratic programming problem

Nonparametrics

Smoothing Splines

- ▶ Solve problem,

$$\hat{g} = \arg \max_g \frac{1}{N} \sum_{n=1}^N (y_n - g(x_n))^2 - \lambda \int_x (g''(x))^2 dx$$

- ▶ Solution is a cubic spline with knots at all the data points
- ▶ Computation involves linear algebra
- ▶ Easy to impose shape restrictions (i.e. monotonicity) – becomes quadratic programming problem
- ▶ Issues:
 - Selecting λ (smoothing parameter)

Nonparametrics

LASSO w/ Interactions

- Sieve estimator, with additional regularization:

$$\hat{g}(x) = \sum_{i=1}^I \hat{a}_i h_i(x)$$

$$\hat{a} = \arg \max_a \frac{1}{N} \sum_{n=1}^N (y_n - \sum_{i=1}^I a_i h_i(x_n))^2 - \lambda \sum_{i=1}^I |a_i|$$

Nonparametrics

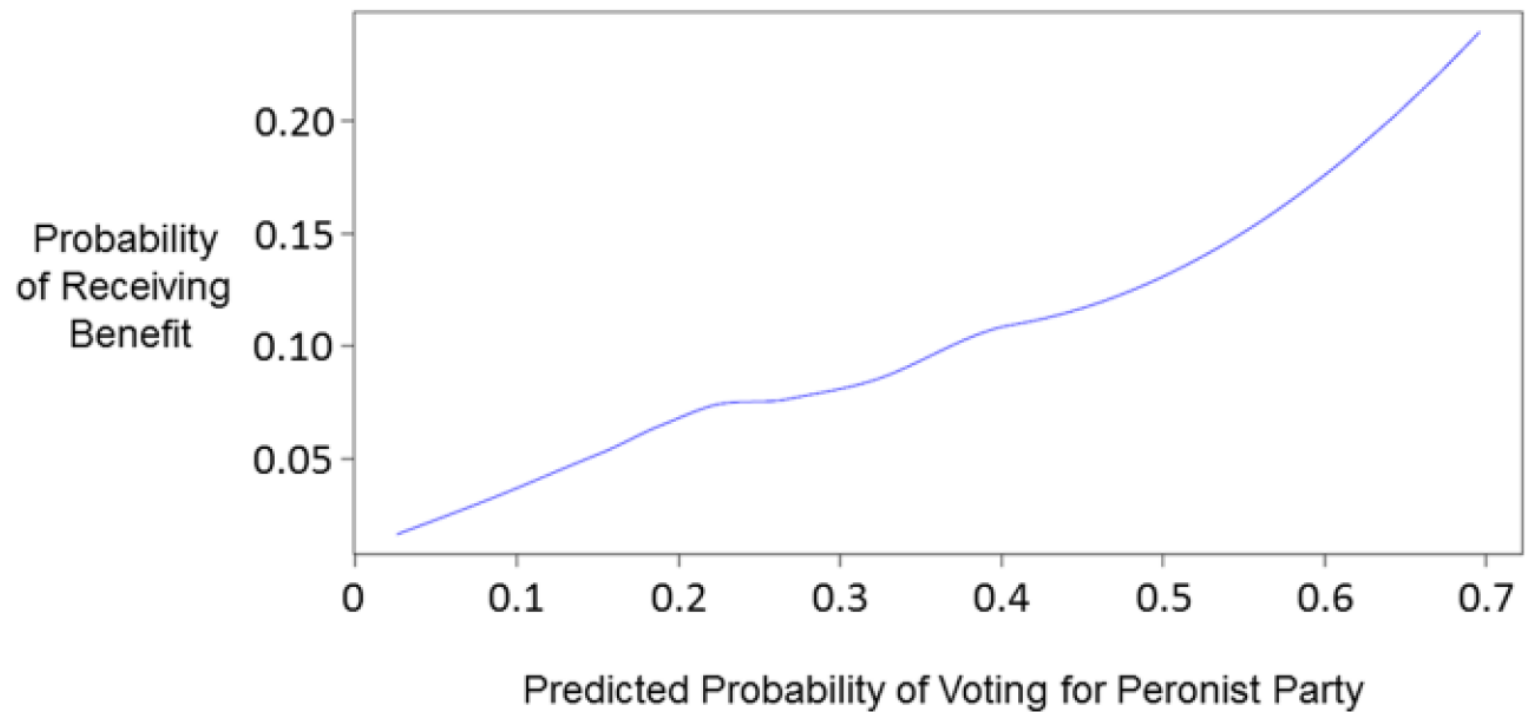
Alternative Nonparametric Estimators

Nonparametric Estimator	Complexity Parameter	Advantages
Kernels	h (variance of kernel)	Incorporate into semiparametric model
k-NN	k (# of matches)	Discrete regressors
Sieves	m (degree)	Incorporate into semiparametric model
Smoothing Splines	λ (degree of smoothing)	Shape restrictions
LASSO w/ interactions	Penalty term	Mostly for prediction
Trees (basic)	# of splits	Interpretability?
Trees (other)	Various	Mostly for prediction

Nonparametrics

More Applications

- ▶ Nichter and Peress (2017) (Kernel Regression):



Nonparametrics

More Applications

- ▶ Nichter and Peress (2017) (Kernel Regression):
- ▶ Alternatives:
 - k-NN and smoothing splines would work fine too

Nonparametrics

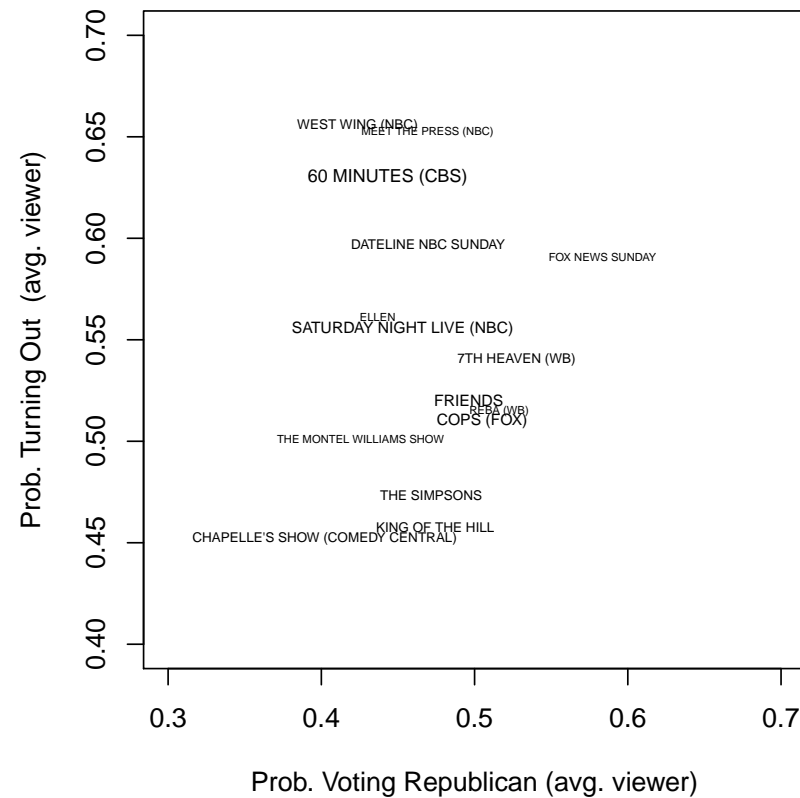
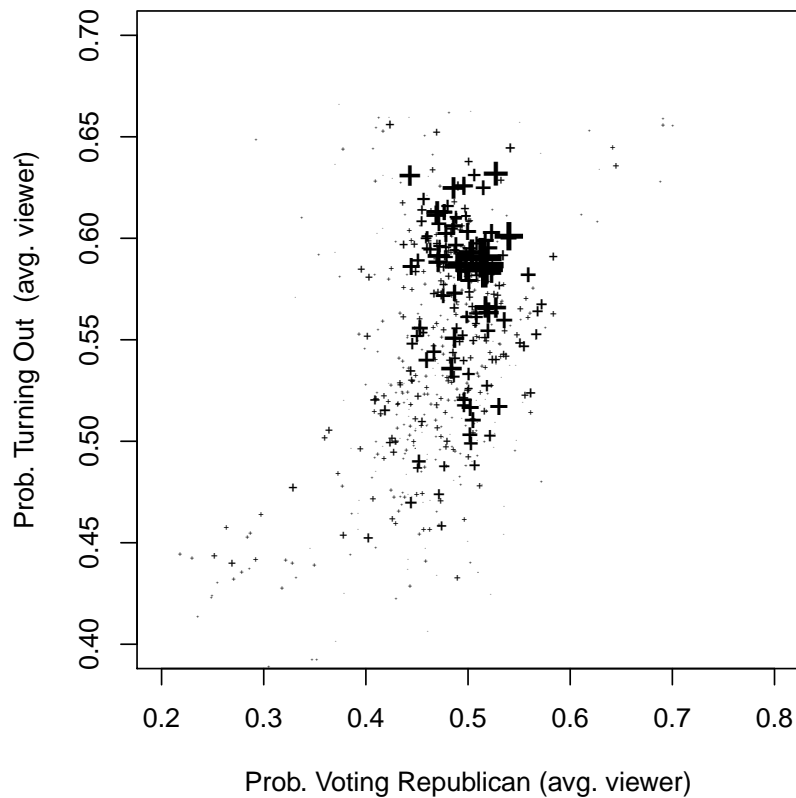
More Applications

- ▶ Nichter and Peress (2017) (Kernel Regression):
- ▶ Alternatives:
 - k-NN and smoothing splines would work fine too
 - Sieves and locally linear Kernel estimators probably less convincing

Nonparametrics

More Applications

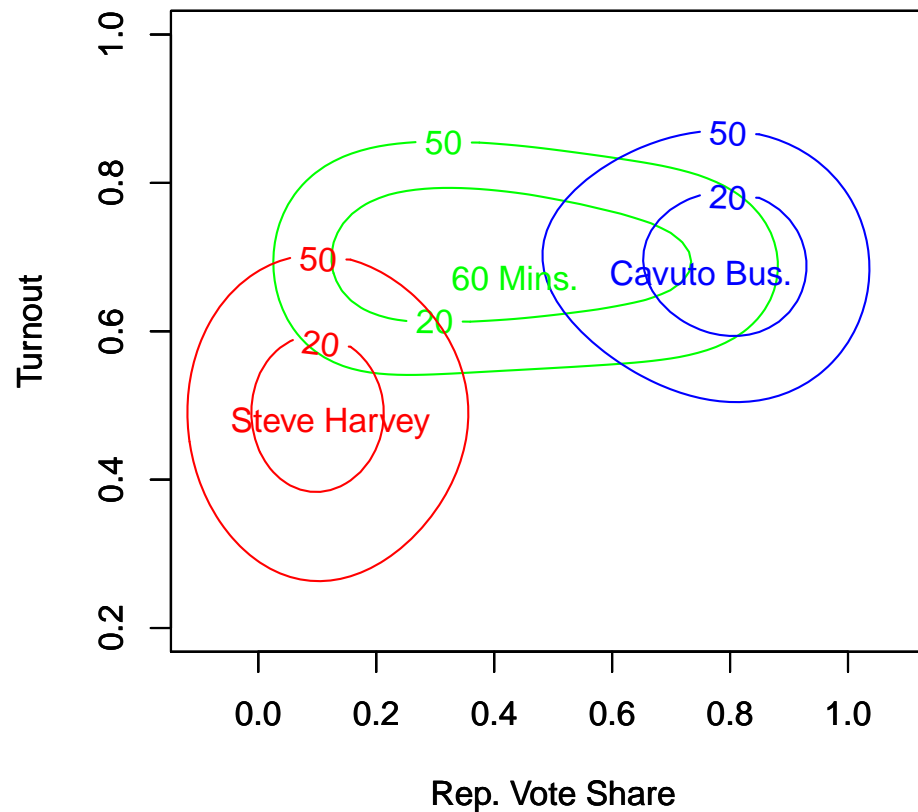
- ▶ Lovett and Peress (2015) (Multivariate Kernel Density Estimation):



Nonparametrics

More Applications

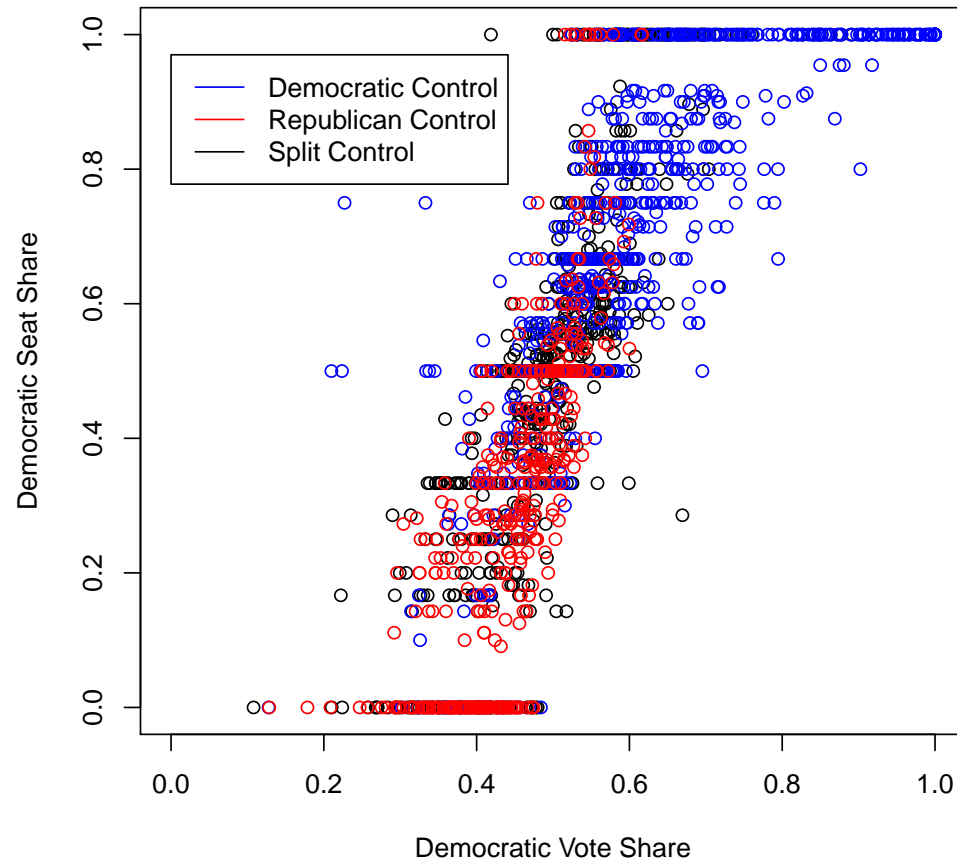
- ▶ Lovett and Peress (2015) (Multivariate Kernel Density Estimation):



Nonparametrics

More Applications

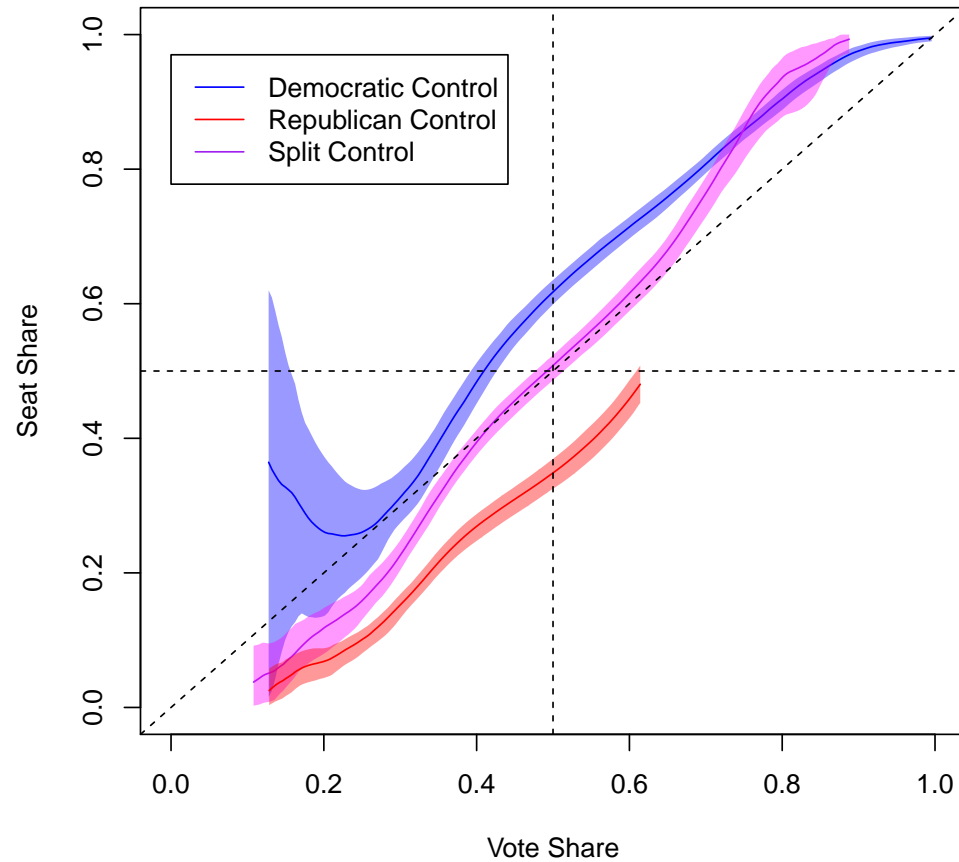
- ▶ Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

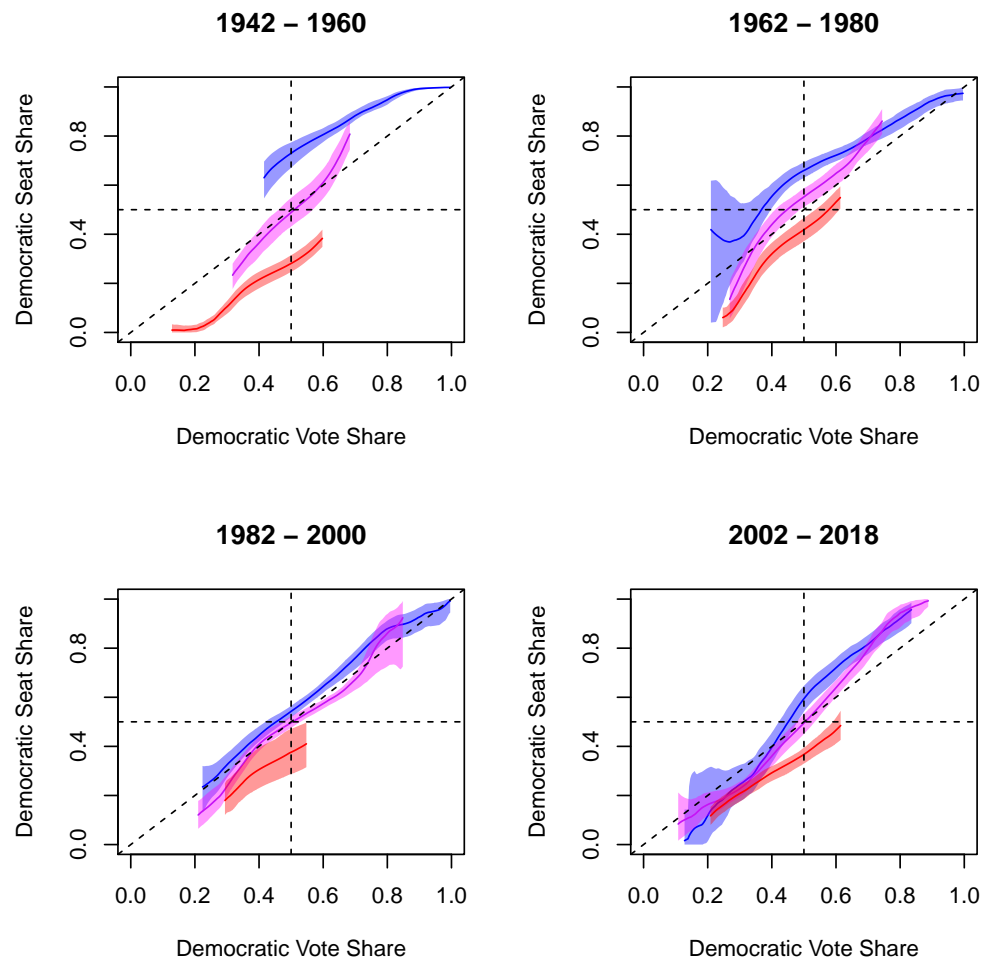
- ▶ Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

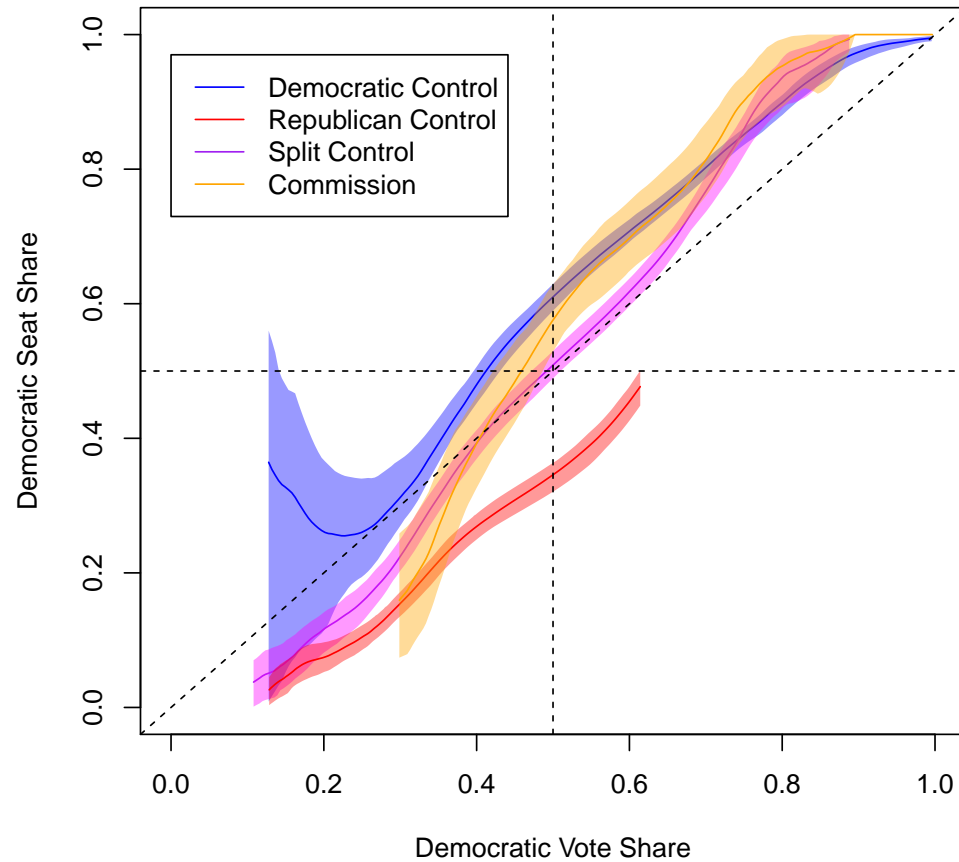
- Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

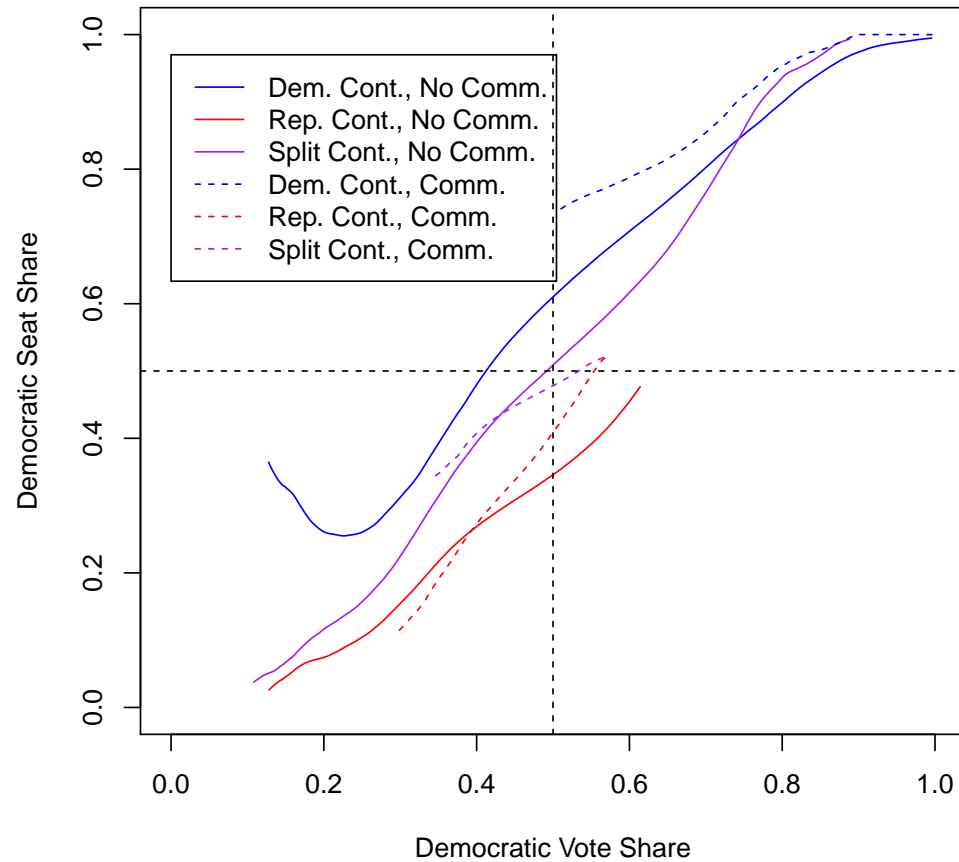
- ▶ Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

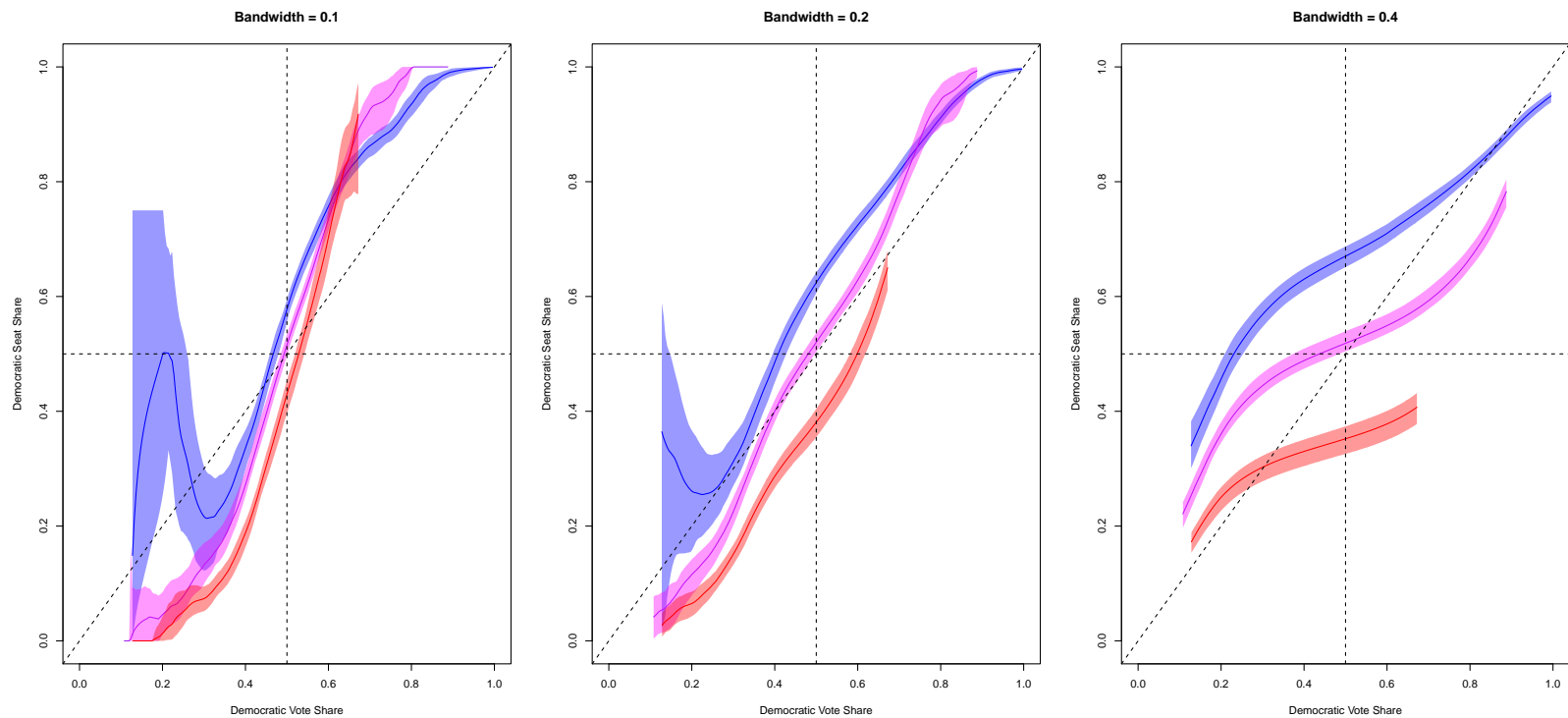
- ▶ Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

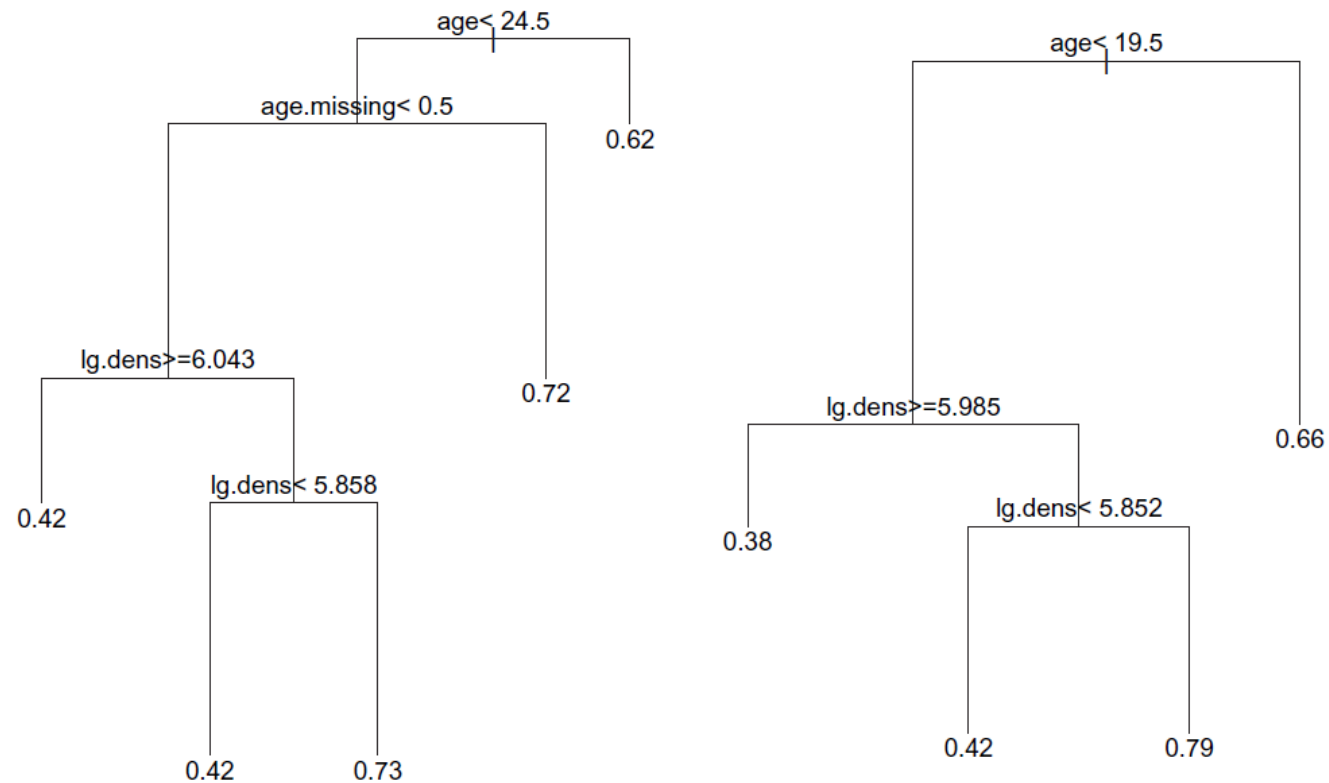
- Peress and Zhao (2020) (Kernel Regression):



Nonparametrics

More Applications

- ▶ Imai and Strauss (2011) (Tree-based Methods):



Nonparametrics

More Applications

- ▶ Imai and Strauss (2011) (Tree-based Methods):
- ▶ Alternatives:
 - Kernel regression—less interpretable and potential issues with discrete variables

Nonparametrics

More Applications

- ▶ Imai and Strauss (2011) (Tree-based Methods):
- ▶ Alternatives:
 - Kernel regression—less interpretable and potential issues with discrete variables
 - k-NN—less interpretable, but otherwise viable

Nonparametrics

More Applications

- ▶ Imai and Strauss (2011) (Tree-based Methods):
- ▶ Alternatives:
 - Kernel regression—less interpretable and potential issues with discrete variables
 - k-NN—less interpretable, but otherwise viable
 - Sieves—many terms and harder to deal with discrete variables

References

- Andrews, Donald W. K. 1994. "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity." *Econometrica* 62:43–72.
- Imai, Kosuke and Aaron Strauss. 2011. "Planning the Optimal Get-out-the-vote Campaign Using Randomized Field Experiments." *Political Analysis* 19:1–19.
- Lovett, Michell and Michael Peress. 2015. "Targeting Political Advertising on Television." *Quarterly Journal of Political Science* 10:391–432.
- Nichter, Simeon and Michael Peress. 2017. "Request Fulfilling: When Citizens Ask for Clientelist Benefits." *Comparitive Political Studies* 50:1086–1117.
- Peress, Michael and Yangzi Zhao. 2020. "How Many Seats in Congress Is Control of Redistricting Worth?" *Legislative Studies Quarterly* 45:433–468.
- Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.