Boundary Correction Methods in Kernel Density Estimation

Tom Alberts
$C_ou(r)a_n(t)$ Institute

joint work with R.J. Karunamuni
University of Alberta

November 29, 2007

Outline

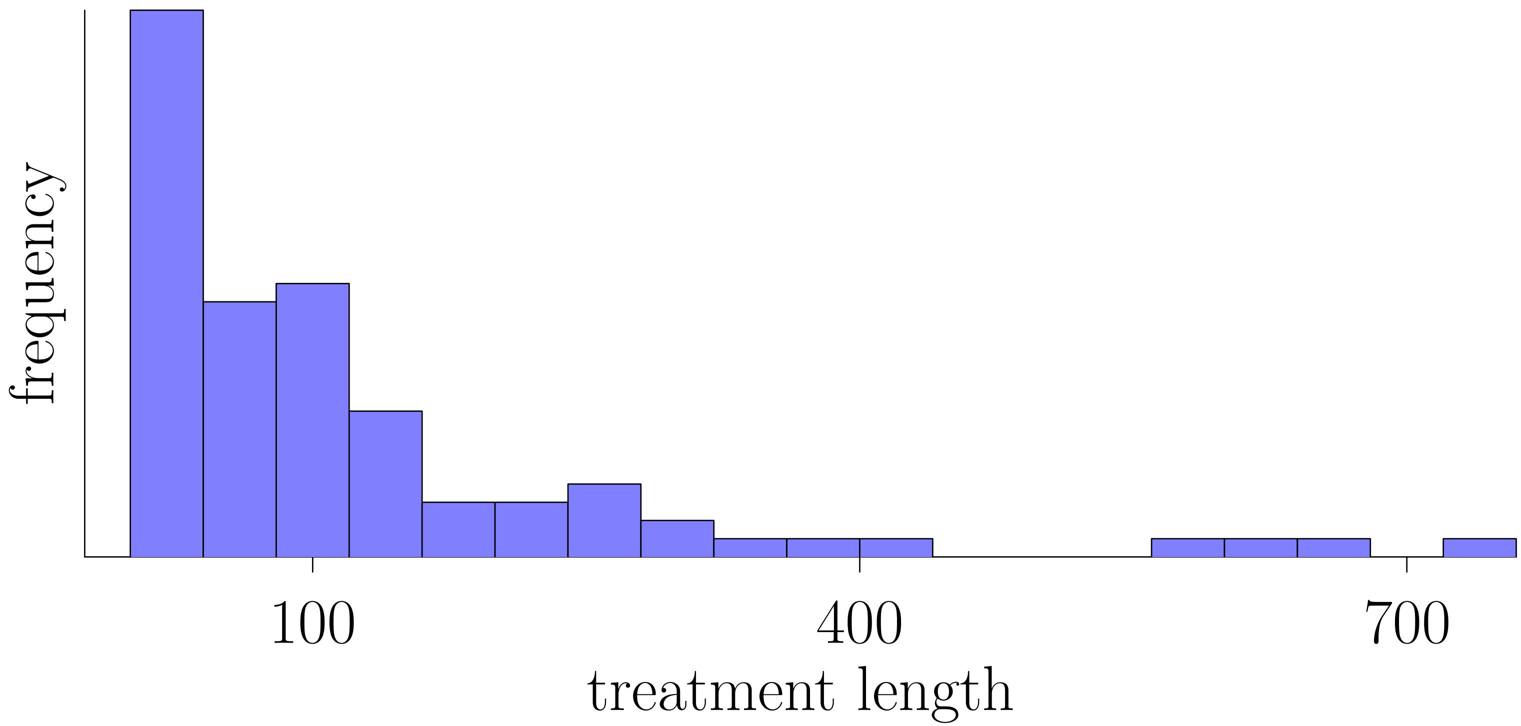
- Overview of Kernel Density Estimation
- Boundary Effects
- Methods for Removing Boundary Effects
- Karunamuni and Alberts Estimator

What is Density Estimation?

- Basic question: given an i.i.d. sample of data $X_1, X_2, \ldots, X_n$, can one estimate the distribution the data comes from?
- As usual, there are *parametric* and *non-parametric* estimators. Here we consider only non-parametric estimators.
- Assumptions on the distribution:
  - It has a probability density function, which we call f,
  - f is as smooth as we need, at least having continuous second derivatives.

Most Basic Estimator: the Histogram!

- Parameters: an origin x_0 and a bandwidth h
- Create bins $\ldots, [x_0 - h, x_0), [x_0, x_0 + h), [x_0 + h, x_0 + 2h), \ldots$



- Dataset: lengths (in days) of 86 spells of psychiatric treatments for patients in a study of suicide risks

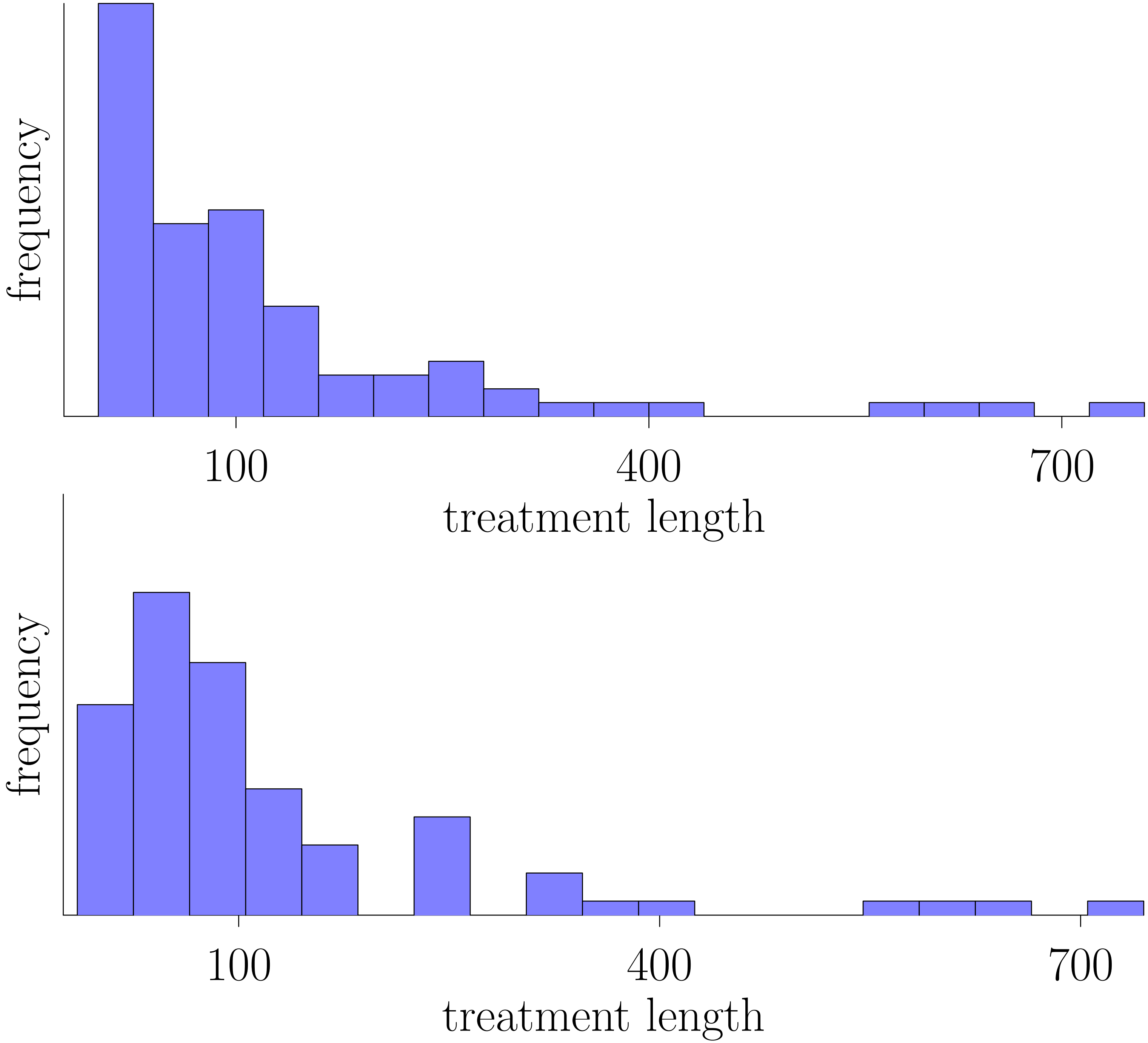
Most Basic Estimator: the Histogram!

- Can write the estimator as

$$f_n(x) = \frac{1}{nh}\#\{X_i : X_i \text{ in the same bin as x }\}$$

- Is it accurate? In the limit, yes.
- A consequence of the Strong Law of Large Numbers: as $n \to \infty$ and $h \to 0$, $f_n(x) \to f(x)$ almost surely.
- Advantages:
  – simple
  – computationally easy
  – well known by the general public
- Disadvantages:
  – depends very strongly on the choice of x_0 and h
  – ugly

Dependence on x_0



Making the Histogram a "Local" Estimator

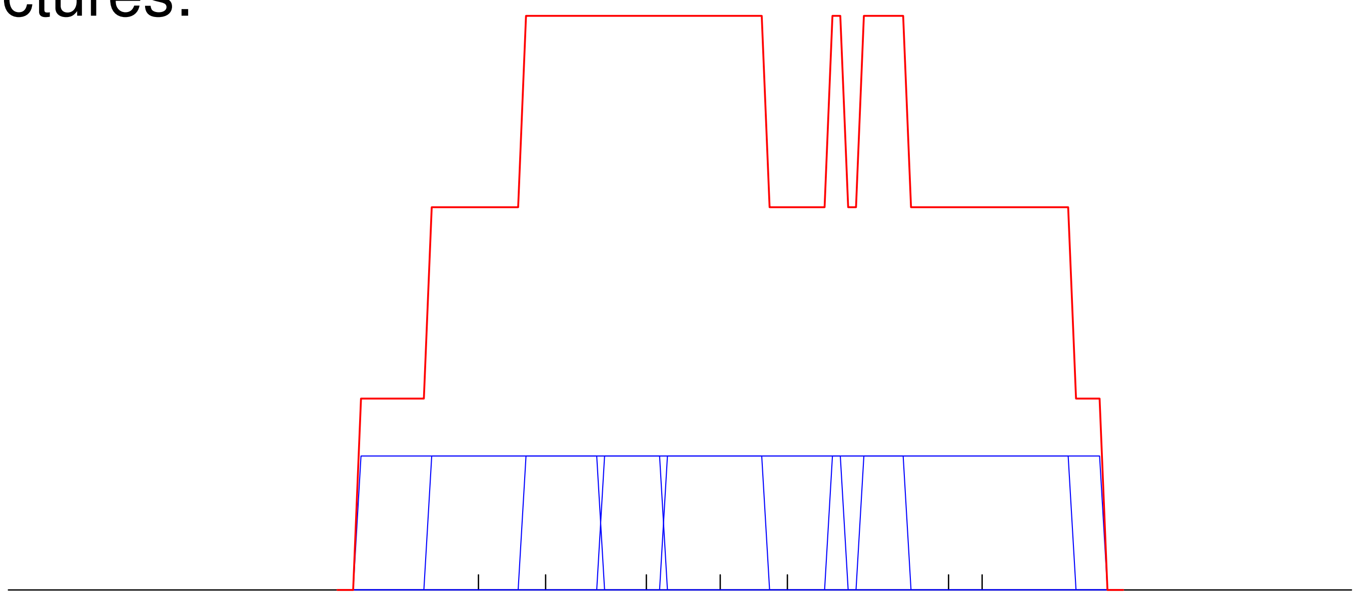
- There's an easy way to get rid of the dependence on x_0. Recall

$$f(x) = \lim_{h \downarrow 0} \frac{1}{2h} \mathbf{P}\left(x - h < X < x + h\right)$$

which can be naively estimated by

$$f_n(x) = \frac{1}{2nh} \#\{X_i : x - h \leq X_i \leq x + h\}$$

- In pictures:

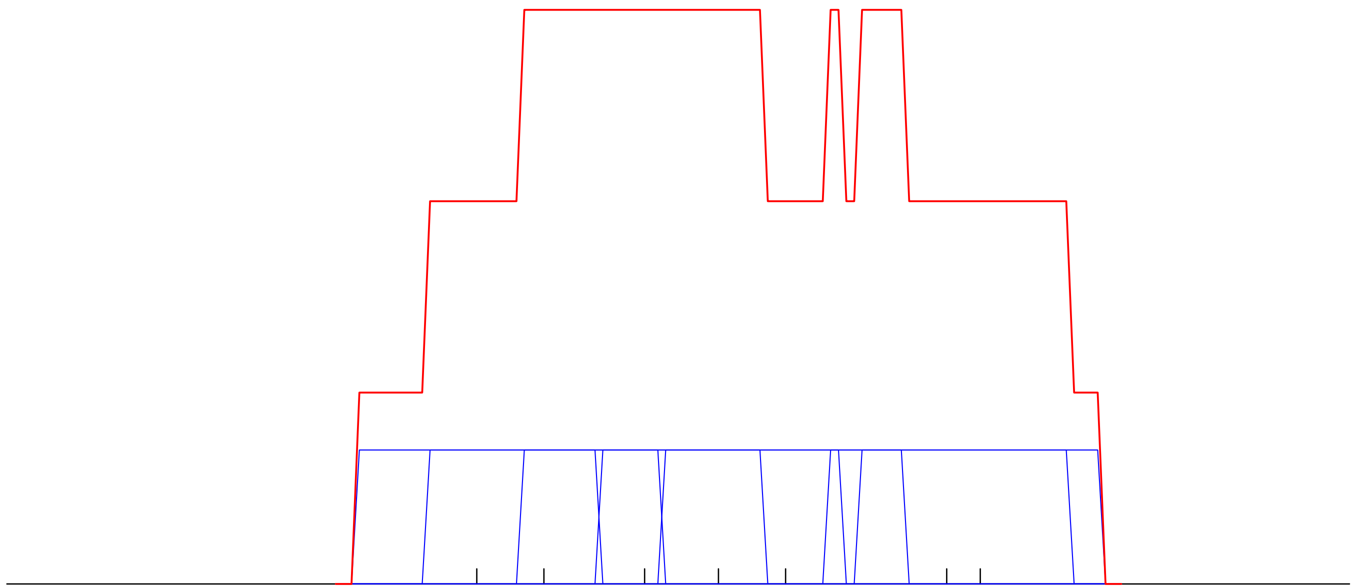


Making the Histogram a "Local" Estimator

- Let $K(x) = \frac{1}{2}\mathbf{1}\{-1 \leq x \leq 1\}$. Can also write the estimator as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

- This is the general form of a kernel density estimator.
- Nothing special about the choice $K(x) = \frac{1}{2}\mathbf{1}\{-1 \leq x \leq 1\}$
- Can use smooth K and get smooth kernel estimators.

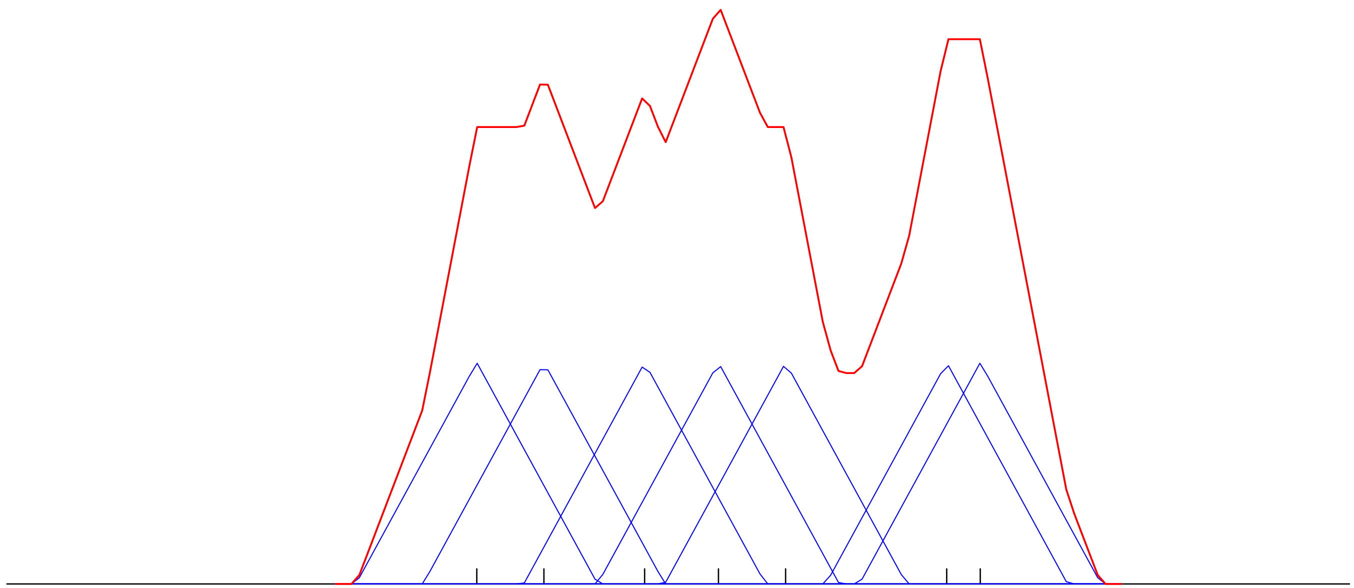


Making the Histogram a "Local" Estimator

- Let $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$. Can also write the estimator as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

- This is the general form of a kernel density estimator.
- Nothing special about the choice $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$
- Can use smooth K and get smooth kernel estimators.

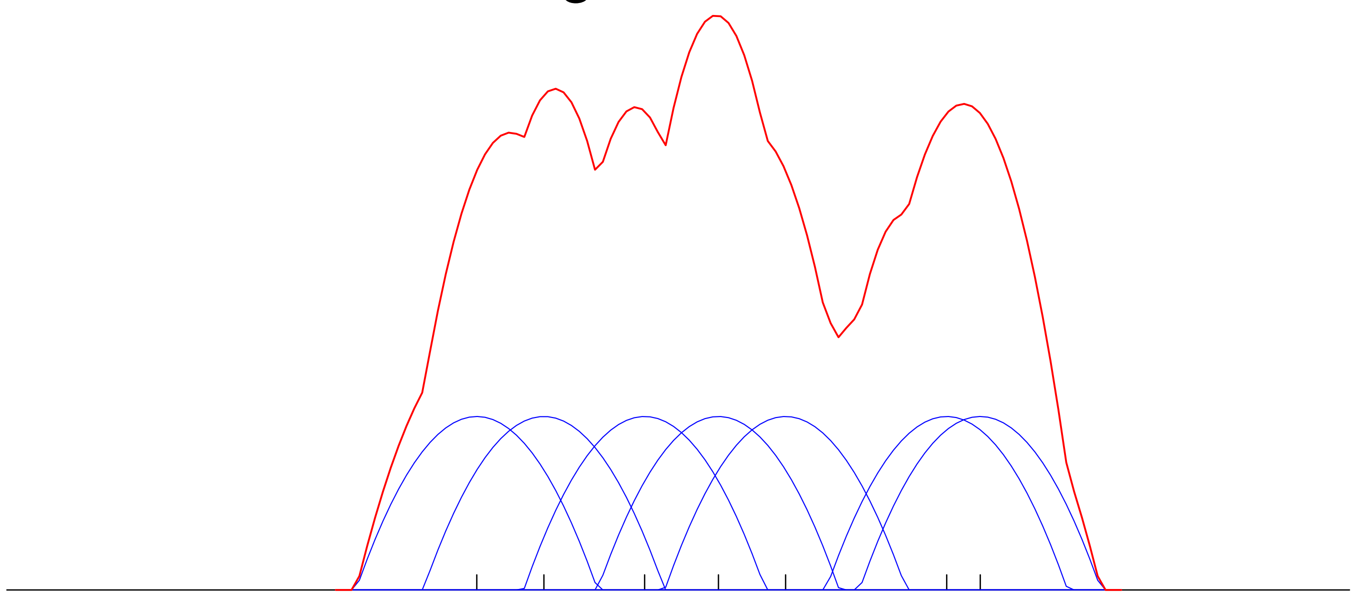


Making the Histogram a "Local" Estimator

- Let $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$. Can also write the estimator as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- This is the general form of a kernel density estimator.
- Nothing special about the choice $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$
- Can use smooth K and get smooth kernel estimators.

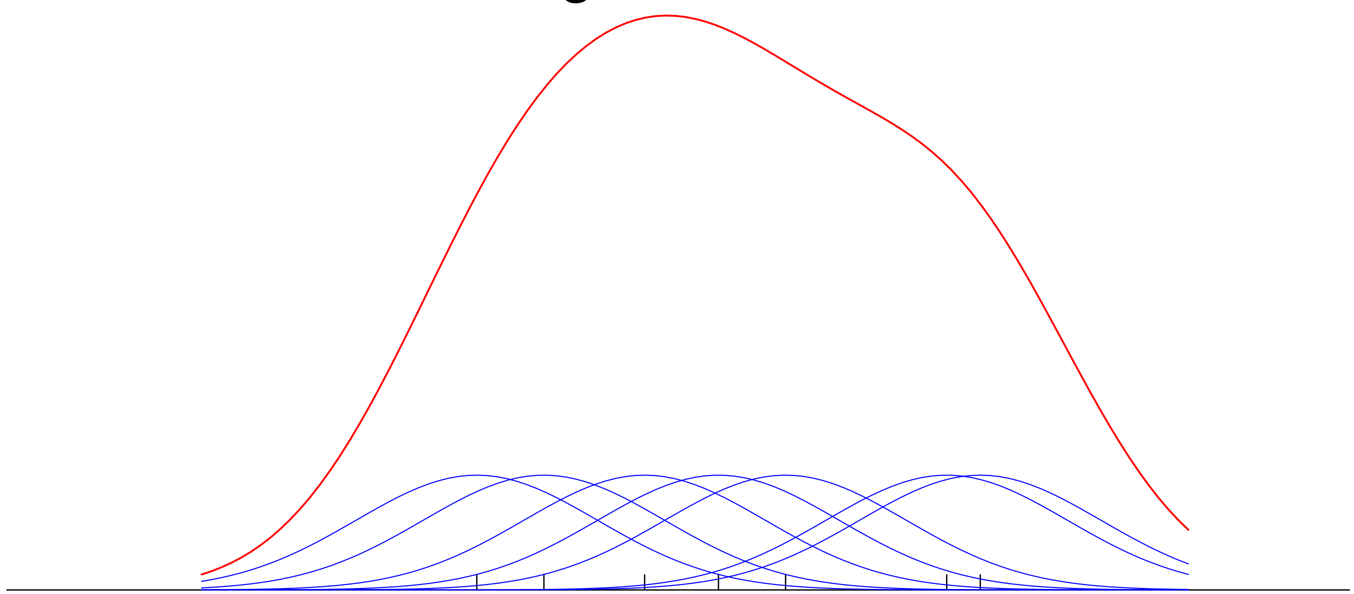


Making the Histogram a "Local" Estimator

- Let $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$. Can also write the estimator as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

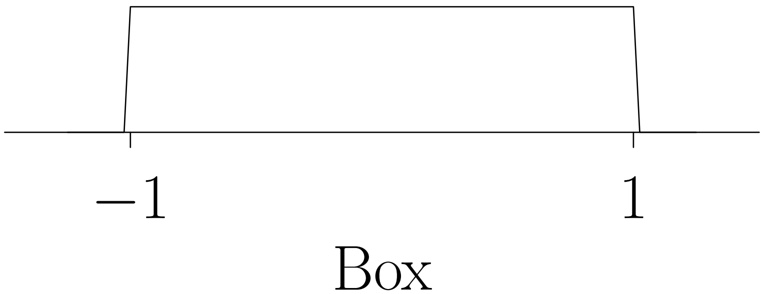
- This is the general form of a kernel density estimator.
- Nothing special about the choice $K(x) = \frac{1}{2}\mathbf{1}\{-1 \le x \le 1\}$
- Can use smooth K and get smooth kernel estimators.

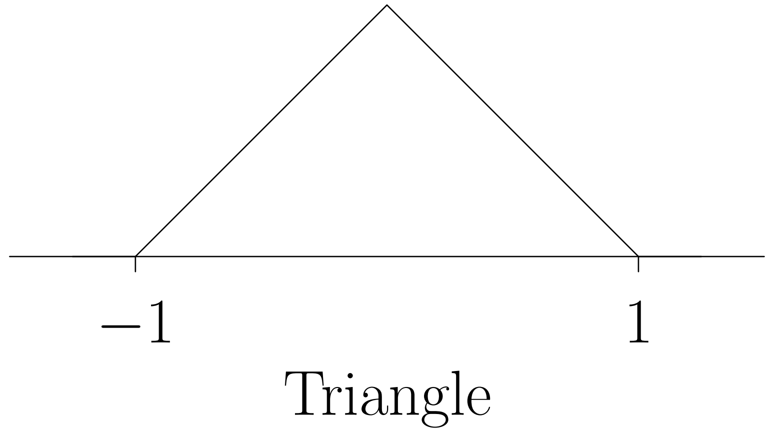


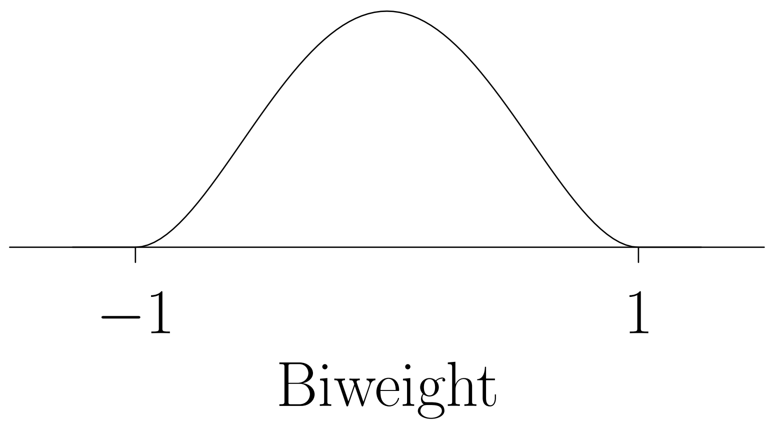
Properties of the Kernel

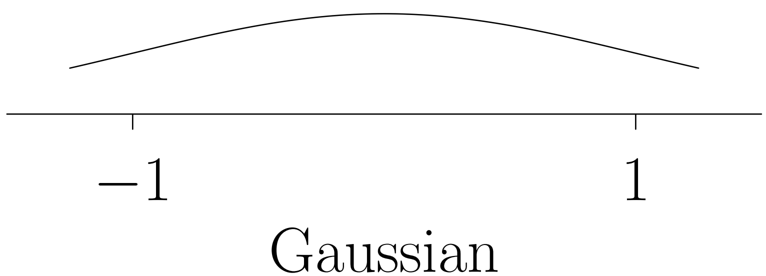
- What other properties should K satisfy?
  - positive
  - symmetric about zero
  - $\int K(t)dt = 1$
  - $\int tK(t)dt = 0$
  - $0 < \int t^2 K(t)dt < \infty$
- If K satisfies the above, it follows immediately that $f_n(x) \geq 0$ and $\int f_n(x)dx = 1$.

Different Kernels









Does the Choice of Kernel Matter?

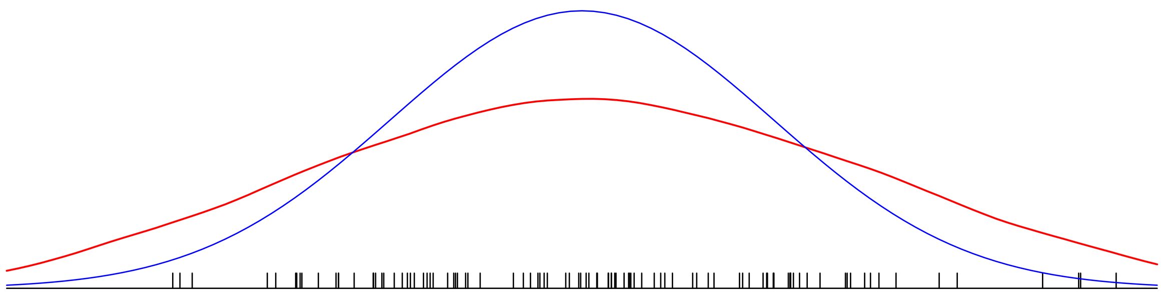
- For reasons that we will see, the optimal K should minimize

$$C(K) = \left(\int t^2 K(t) dt\right)^{2/5} \left(\int K(t)^2 dt\right)^{4/5}$$

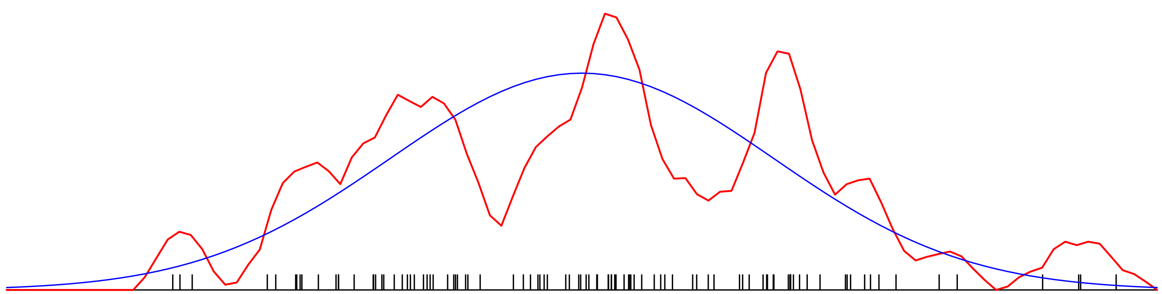
- It has been proven that the Epanechnikov kernel is the minimizer.
- However, for most other kernels $C(K)$ is not much larger than C(Epanechnikov). For the five presented here, the worst is the box estimator, but C(Box) $<$ $1.1C$(Epanechnikov)
- Therefore, usually choose kernel based on other considerations, i.e. desired smoothness.

How Does Bandwidth Affect the Estimator?

- The bandwidth h acts as a smoothing parameter.
- Choose h too small and spurious fine structures become visible.
- Choose h too large and many important features may be oversmoothed.



$h = 2$



$h = .2$

How Does Bandwidth Affect the Estimator?

- A common choice for the "optimal" value of h is

$$\left(\int t^2 K(t)dt\right)^{-2/5}\left(\int K(t)^2 dt\right)^{1/5}\left(\int f''(x)^2 dx\right)^{-1/5} n^{-1/5}$$

- Note the optimal choice still depends on the unknown f
- Finding a good estimator of h is probably the most important problem in kernel density estimation. But it's not the focus of this talk.

Measuring Error

- How do we measure the error of an estimator $f_n(x)$?
- Use Mean Squared Error throughout.
- Can measure error at a single point

$$\begin{aligned}\mathbf{E}\left[(f_n(x) - f(x))^2\right] &= (\mathbf{E}\left[f_n(x)\right] - f(x))^2 + \mathrm{Var}\left(f_n(x)\right) \\ &= \mathrm{Bias}^2 + \mathrm{Variance}\end{aligned}$$

- Can also measure error over the whole line by integrating

$$\int \mathbf{E}\left[(f_n(x) - f(x))^2\right] dx$$

- The latter is called *Mean Integrated Squared Error* (MISE).
- MISE has an integrated bias and variance part.

Bias and Variance

$$
\begin{aligned}
\mathbf{E}\left[f_n(x)\right] - f(x) &= \frac{1}{nh}\sum_{i=1}^{n} \mathbf{E}\left[K\left(\frac{x - X_i}{h}\right)\right] \\
&= \frac{1}{h}\mathbf{E}\left[K\left(\frac{x - X_i}{h}\right)\right] \\
&= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\
&= \int K(t)f(x - ht)dt - f(x) \\
&= \int K(t)(f(x-ht) - f(x))dt \\
&= \int K(t)\left(-htf'(x) + \frac{1}{2}h^2t^2f''(x) + \ldots\right)dt \\
&= -hf'(x)\int tK(t)dt + \frac{1}{2}h^2f''(x)\int t^2K(t)dt + \ldots \\
&= \frac{1}{2}h^2f''(x)\int t^2K(t)dt + \text{higher order terms in } h
\end{aligned}
$$

Bias and Variance

- Can work out the variance in a similar way

$$\mathbf{E}\left[f_n(x)\right] - f(x) = \frac{h^2}{2} f^{(2)}(x) \int t^2 K(t)dt + o(h^2)$$

$$\mathrm{Var}\left(f_n(x)\right) = \frac{1}{nh} f(x) \int K(t)^2 dt + o\left(\frac{1}{nh}\right)$$

- Notice how h affects the two terms in opposite ways.
- Can integrate out the bias and variance estimates above to get the MISE

$$\frac{h^4}{4}\left(\int t^2 K(t)dt\right)^2 \int f''(x)^2 dx + \frac{1}{nh}\int K(t)^2 dt$$

plus some higher order terms

Bias and Variance

- The optimal h from before was chosen so as to minimize the MISE.
- This minimum of the MISE turns out to be

$$\frac{5}{4}C(K)\left(\int f''(x)^2dx\right)^{1/5}n^{-4/5}$$

where $C(K)$ was the functional of the kernel given earlier. Thus we see we chose the "optimal" kernel to be the one that minimizes the MISE, all else held equal.

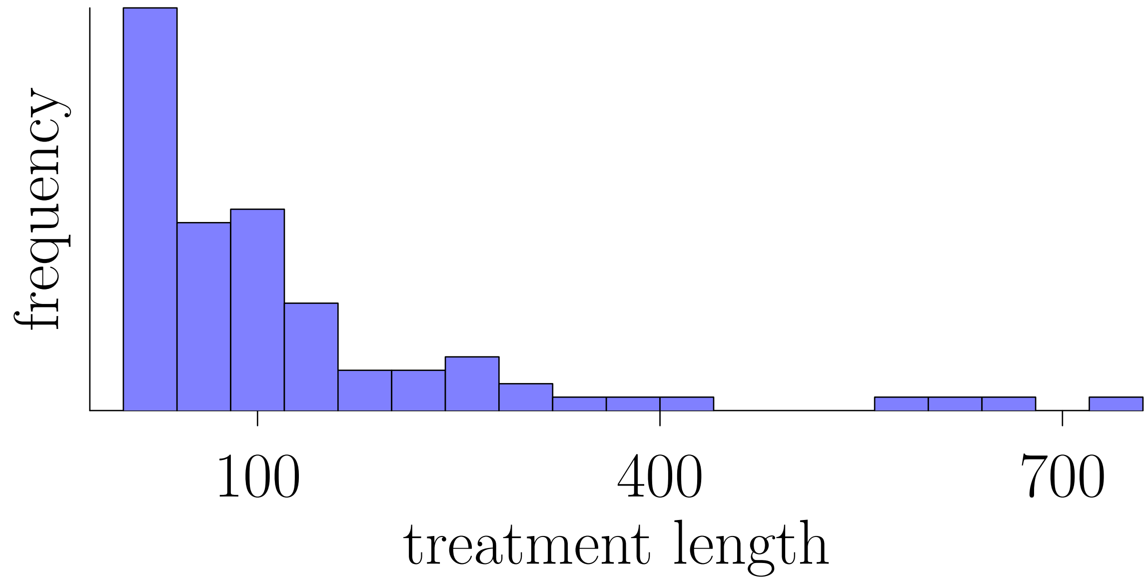
- Note that when using the optimal bandwidth, the MISE goes to zero like $n^{-4/5}$.

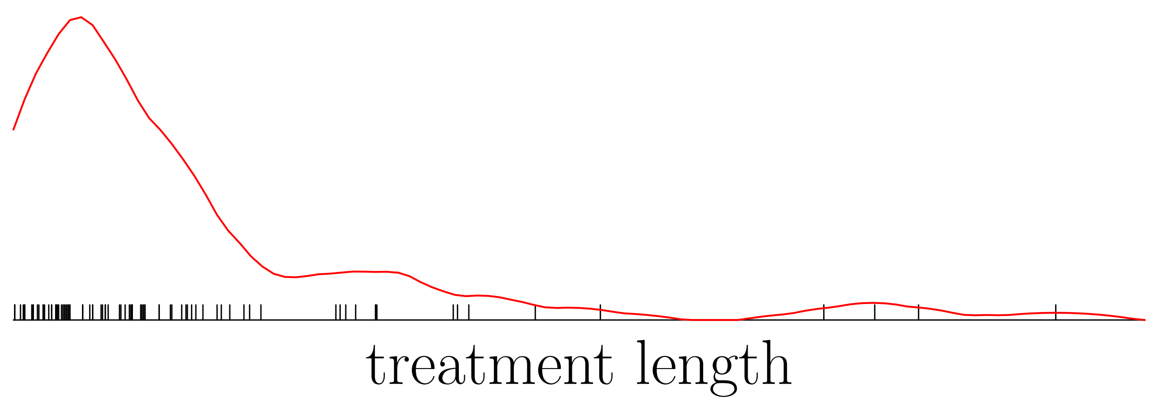
Boundary Effects

- All of these calculations implicitly assume that the density is supported on the entire real line.
- If it's not, then the estimator can behave quite poorly due to what are called *boundary effects*. Combatting these is the main focus of this talk.
- For simplicity, we'll assume from now on that f is supported on $[0, \infty)$.
- Then $[0, h)$ is called the boundary region.

Boundary Effects

- In the boundary region, f_n usually underestimates f.
- This is because f_n doesn't "feel" the boundary, and penalizes for the lack of data on the negative axis.





Boundary Effects

- For $x \in [0, h)$, the bias of $f_n(x)$ is of order $O(h)$ rather than $O(h^2)$.
- In fact it's even worse: $f_n(x)$ is not even a consistent estimator of $f(x)$.

$$\mathbf{E}\left[f_n(x)\right] = f(x)\int_{-1}^{c} K(t)dt - hf'(x)\int_{-1}^{c} tK(t)dt + \frac{h^2}{2}f''(x)\int_{-1}^{c} t^2K(t)dt + o(h^2)$$

$$\mathrm{Var}\left(f_n(x)\right) = \frac{f(x)}{nh}\int_{-1}^{c} K(t)^2dt + o\left(\frac{1}{nh}\right)$$

where $x = ch, 0 \le c \le 1$.

- Note the variance isn't much changed.

Methods for Removing Boundary Effects

- There is a vast literature on removing boundary effects. I briefly mention 4 common techniques:
  - Reflection of data
  - Transformation of data
  - Pseudo-Data Methods
  - Boundary Kernel Methods
- They all have their advantages and disadvantages.
- One disadvantage we don't like is that some of them, especially boundary kernels, can produce negative estimators.

Reflection of Data Method

- Basic idea: since the kernel estimator is penalizing for a lack of data on the negative axis, why not just put some there?
- Simplest way: just add $-X_1, -X_2, \ldots, -X_n$ to the data set.
- Estimator becomes:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right) \right\}$$

  for $x \geq 0$, $\hat{f}_n(x) = 0$ for $x < 0$.

- It is easy to show that $\hat{f}_n'(x) = 0$.
- Hence it's a very good method if the underlying density has $f'(0) = 0$.

Transformation of Data Method

- Take a one-to-one, continuous function $g : [0, \infty) \to [0, \infty)$.
- Use the regular kernel estimator with the transformed data set $\{g(X_1), g(X_2), \ldots, g(X_n)\}$.
- Estimator

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - g(X_i)}{h}\right)$$

- Note this isn't really estimating the pdf of X, but instead of $g(X)$.
- Leaves room for manipulation then. One can choose g to get the data to produce whatever you want.

Pseudo-Data Methods

- Due to Cowling and Hall, this generates data beyond the left endpoint of the support of the density.
- Kind of a “reflected transformation estimator”. It transforms the data into a new set, then puts this new set on the negative axis.

$$\hat{f}_n(x) = \frac{1}{nh}\left[\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^{m} K\left(\frac{x + X_{(-i)}}{h}\right)\right]$$

- Here $m \leq n$, and

$$X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + \frac{10}{3}X_{(i)}$$

where $X_{(t)}$ linearly interpolates among $0, X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

Boundary Kernel Method

- At each point in the boundary region, use a different kernel for estimating function.
- Usually the new kernels give up the symmetry property and put more weight on the positive axis.

$$\hat{f}_n(x) = \frac{1}{nh_c} \sum_{i=1}^{n} K_{(c/b(c))} \left(\frac{x - X_i}{h_c} \right)$$

where $x = ch, 0 \leq c \leq 1$, and $b(c) = 2 - c$. Also

$$K_{(c)}(t) = \frac{12}{(1+c)^4}(1+t) \left\{ (1-2c)t + \frac{3c^2 - 2c + 1}{2} \right\} \mathbf{1} \{-1 \leq t \leq c\}$$

Method of Karunamuni and Alberts

- Our method combines transformation and reflection.

$$\tilde{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x + g(X_i)}{h}\right) + K\left(\frac{x - g(X_i)}{h}\right) \right\}$$

  for some transformation g to be determined.

- We choose g so that the bias is of order $O(h^2)$ in the boundary region, rather than $O(h)$.

- Also choose g so that $g(0) = 0, g'(0) = 1$, g is continuous and increasing.

Method of Karunamuni and Alberts

- Do a **very** careful Taylor expansion of f and g in

$$\mathbf{E}\left[\tilde{f}_n(x)\right] = \frac{1}{h}\int \left\{K\left(\frac{x+g(y)}{h}\right) + K\left(\frac{x-g(y)}{h}\right)\right\} f(y)dy$$

  to compute the bias.

- Set the h coefficient of the bias to be zero requires

$$g''(0) = 2f'(0)\int_c^1 (t-c)K(t)dt \Big/ f(0)\left(c + 2\int_c^1 (t-c)K(t)dt\right).$$

  where $x = ch, 0 \leq c \leq 1$.

- Most novel feature: note that $g''(0)$ actually depends on x!
- What this means: at different points x, the data is transformed by a different amount.

Method of Karunamuni and Alberts

- Simplest possible g satisfying these conditions

$$g(y) = y + \frac{1}{2} d k_c' y^2 + \lambda_0 (d k_c')^2 y^3$$

where

$$\begin{aligned} d &= f^{(1)}(0) \, / f(0), \\ k_c' &= 2 \int_c^1 (t-c) K(t) dt \Bigg/ \left(c + 2 \int_c^1 (t-c) K(t) dt \right), \end{aligned}$$

and λ_0 is big enough so that g is strictly increasing.

- Note g really depends on c, so we write $g_c(y)$ instead.
- Hence the amount of transformation of the data depends on the point x at which we're estimating $f(x)$.
- Important feature: $k_c' \to 0$ as $c \uparrow 1$.

Method of Karunamuni and Alberts

- Consequently, $g_c(y) = y$ for $c \geq 1$.
- This means our estimator reduces to the regular kernel estimator at interior points!
- We like that feature: the regular kernel estimator does well at interior points so why mess with a good thing?
- Also note that our estimator is always positive.
- Moreover, by performing a careful Taylor expansion of the boundary, one can show the variance is still $O\left(\frac{1}{nh}\right)$.

$$\mathrm{Var}\left(\tilde{f}_n(x)\right) = \frac{f(0)}{nh}\left\{2\int_c^1 K(t)K(2c-t)dt + \int_{-1}^1 K^2(t)dt\right\} + o\left(\frac{1}{nh}\right)$$

Method of Karunamuni and Alberts

- Note that $g_c(y)$ requires a parameter $d = f'(0)/f(0)$.
- Of course we don't know this, so we have to estimate it somehow.
- We note $d = \frac{d}{dx} \log f(x)\Big|_{x=0}$, which we can estimate by

$$\hat{d} = \frac{\log f_n^*(h_1) - \log f_n^*(0)}{h_1}$$

where f_n^* is some other kind of density estimator.

- We follow methodology of Zhang, Karunamuni and Jones for this.
- Imporant feature: $d = 0$, then $g_c(y) = y$.
- This means our estimator reduces to the reflection estimator if $f'(0) = 0$!

Method of Karunamuni and Alberts

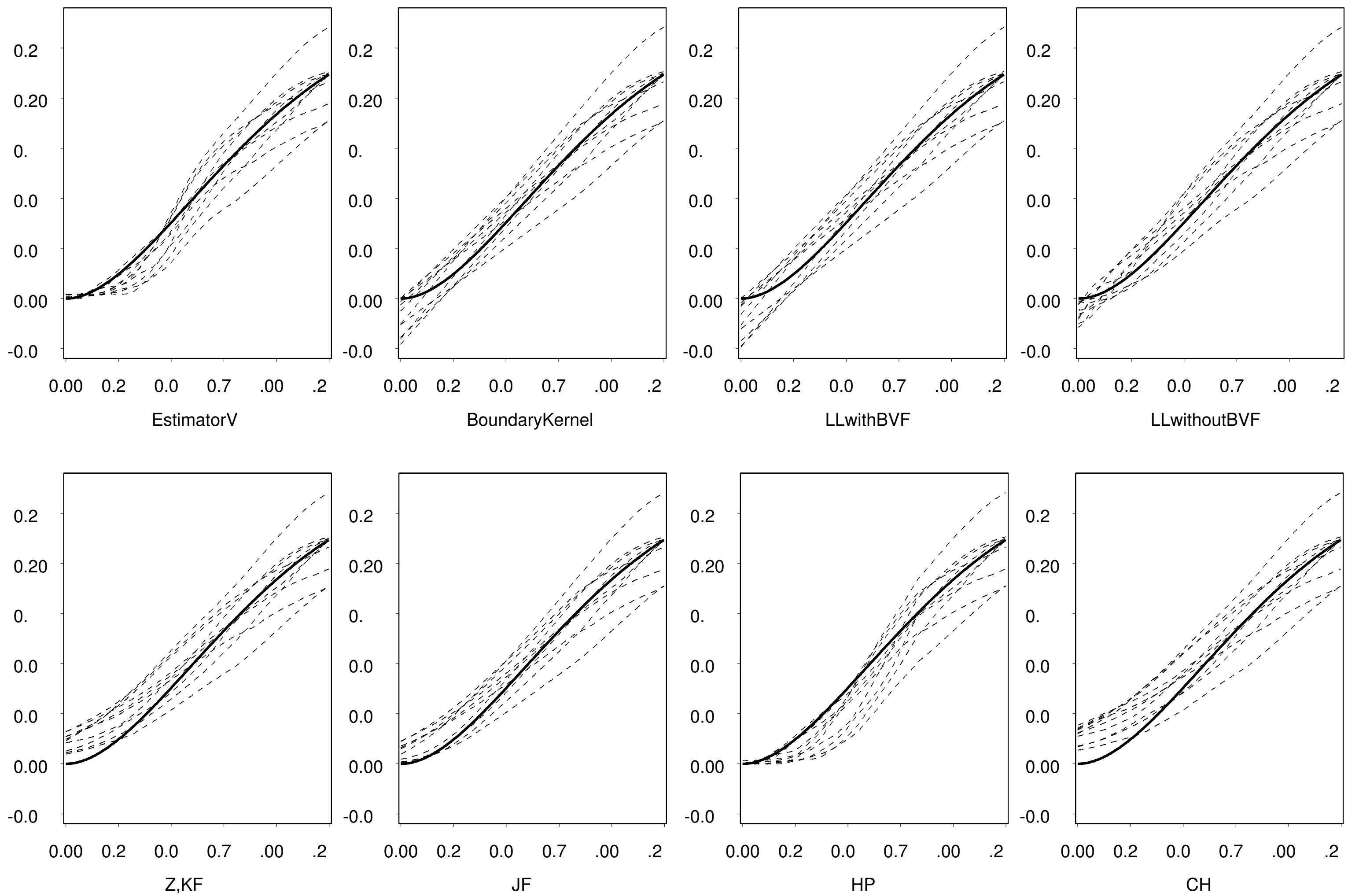
- I mention that our method can be generalized to having two distinct transformations involved.

$$\tilde{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x + g_1(X_i)}{h}\right) + K\left(\frac{x - g_2(X_i)}{h}\right) \right\}$$

- With both g_1 and g_2 there are many degrees of freedom.
- In another paper we investigated five different pairs of (g_1, g_2).
- As would be expected, no one pair did exceptionally well on all shapes of densities.
- The previous choice $g_1 = g_2 = g$ was the most consistent of all the choices, so we recommend it for practical use.

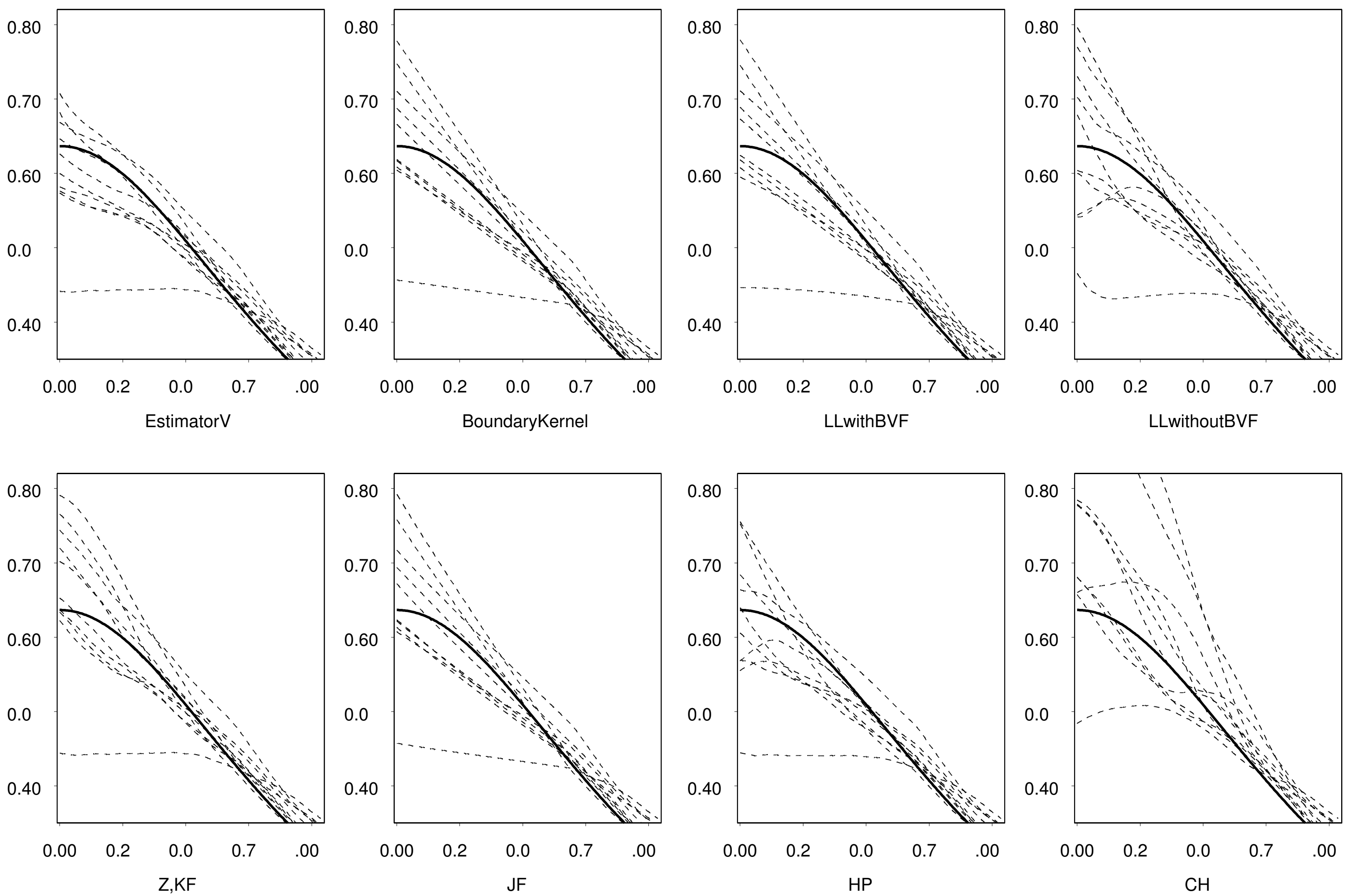
Simulations of Our Estimator

$$f(x) = \frac{x^2}{2}e^{-x}, x \geq 0, \text{ with } h = .832109$$



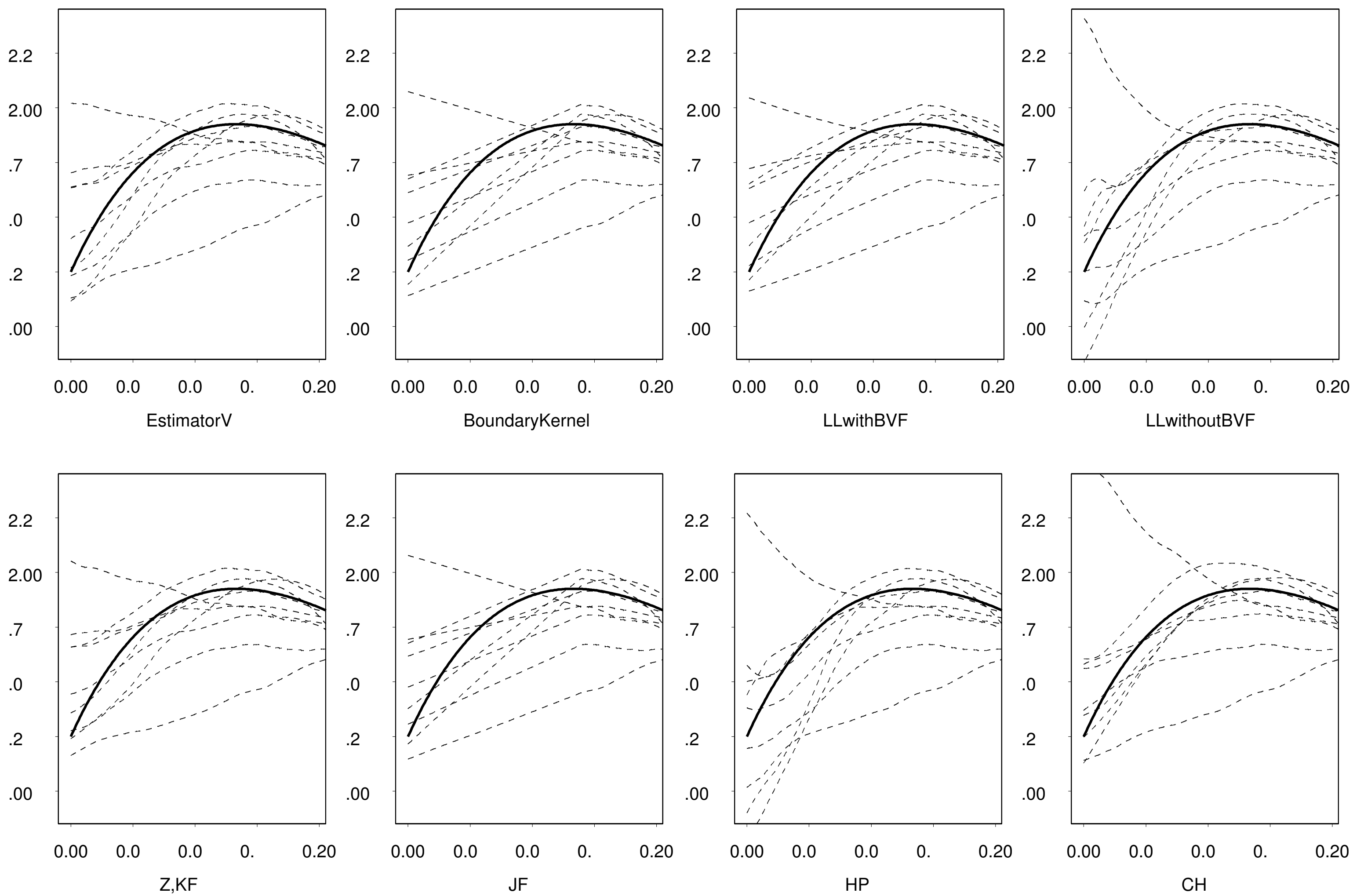
Simulations of Our Estimator

$$f(x) = \frac{2}{\pi(1+x^2)}, x \geq 0, \text{ with } h = .690595$$



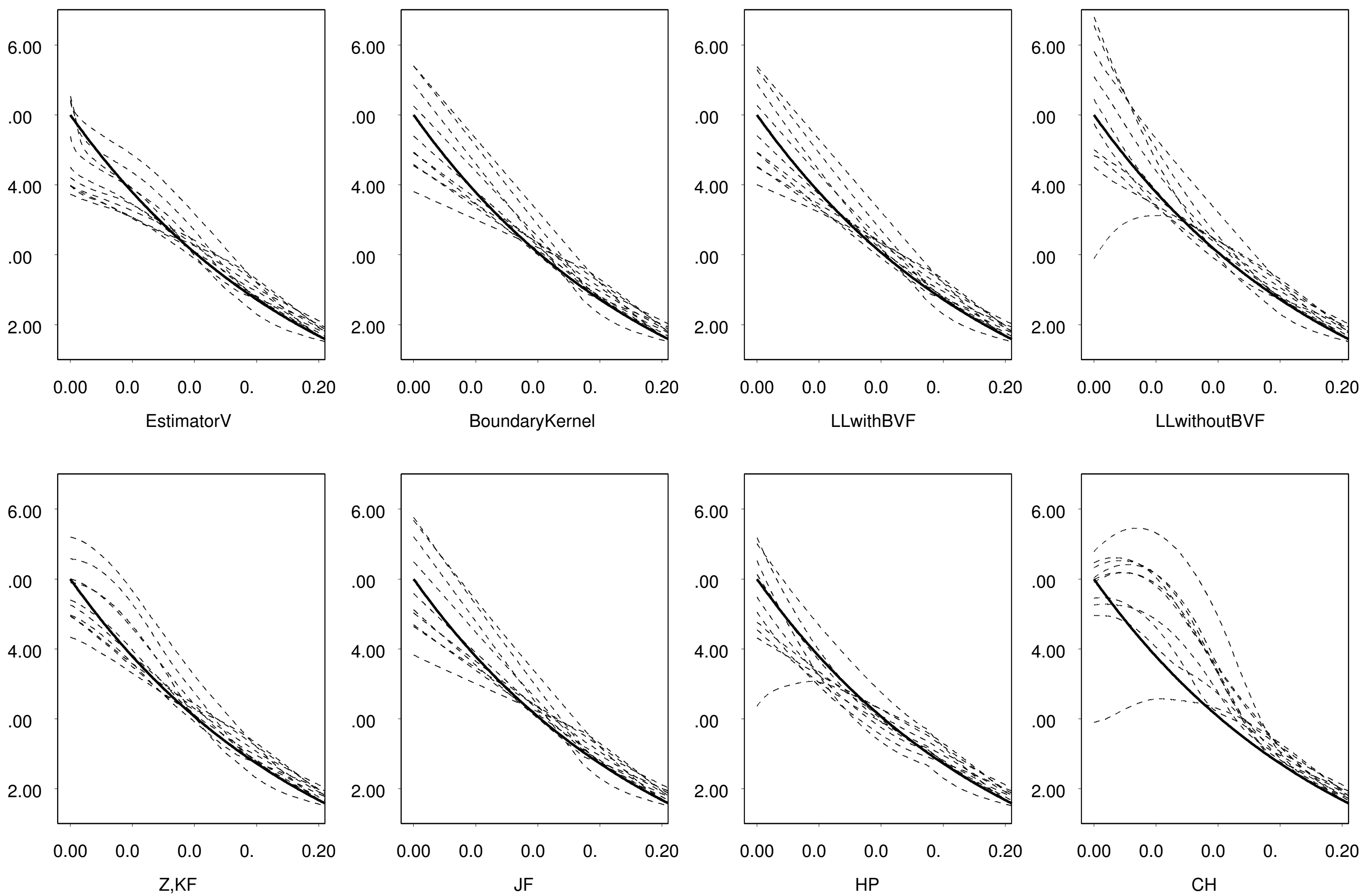
Simulations of Our Estimator

$$f(x) = \frac{5}{4}(1+15x)e^{-5x}, x \geq 0, \text{ with } h = .139332$$



Simulations of Our Estimator

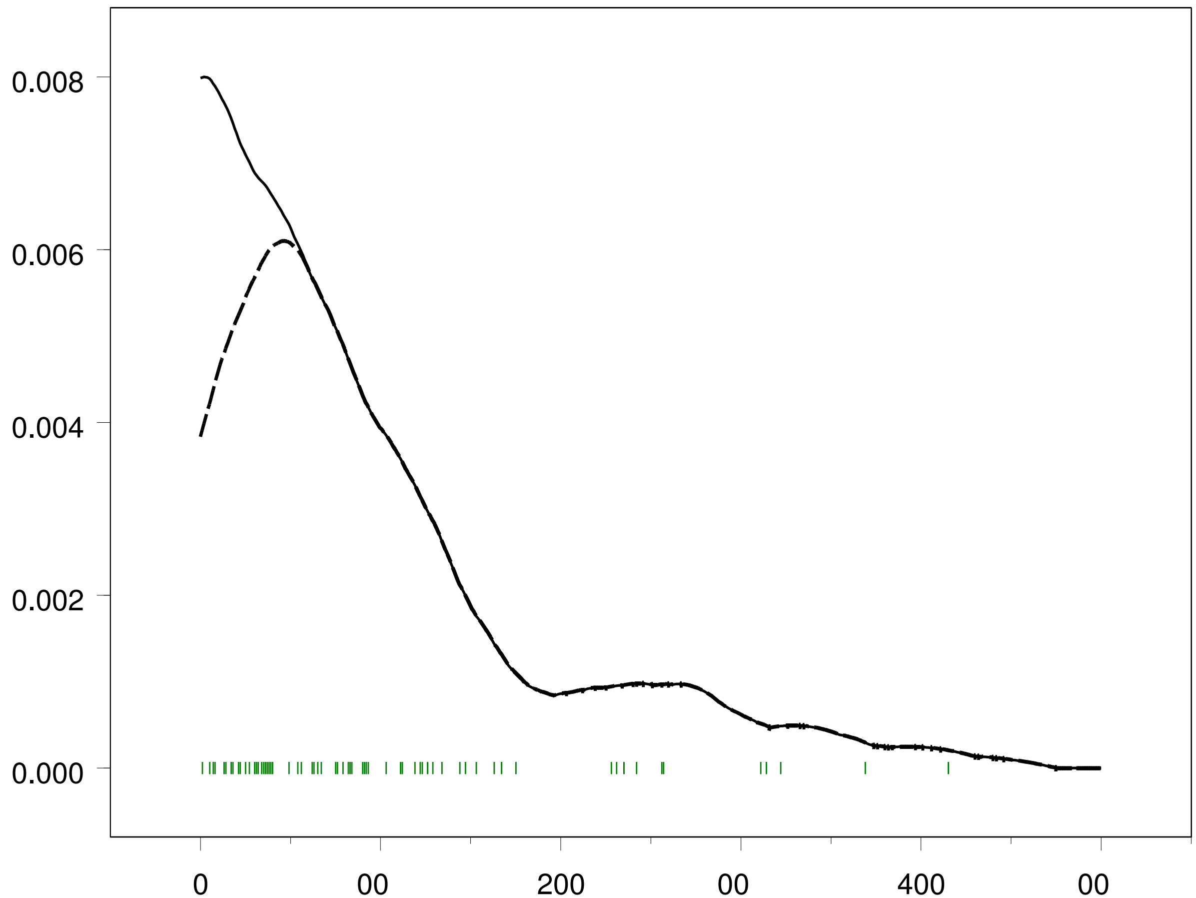
$$f(x) = 5e^{-5x}, x \geq 0, \text{ with } h = .136851$$



Our First Estimator $g_1 = g_2 = g$ on the Suicide Data

Dashed Line: Regular Kernel Estimator

Solid Line: Karunamuni and Alberts



Slides Produced With

Asymptote: The Vector Graphics Language

`http://asymptote.sf.net`

(freely available under the GNU public license)