Density estimation

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester

Introduction/Problem Statement

Let X be a r.v. (discrete or continuous) $\sim f(x;\boldsymbol{\theta}) \equiv f(x)$, $x \in \mathcal{X}$

Problem: Estimation of the p.d.f. or p.m.f. f from a random sample.

Let $\mathbf{X} = (X_1, X_2 \ldots X_n)$ be a random sample, where $X_1, X_2 \ldots X_n$ are i.i.d. r.vs. $\sim f(x)$ and $\mathbf{x} = (x_1, x_2 \ldots x_n)$ observations/data.

For every $\mathbf{y} = (y_1, y_2 \ldots y_n) \in \mathcal{X}^n$ let

$$f(\mathbf{y}) = \prod_{i=1}^{n} f(y)$$

denote the sampling distribution; i.e. the distribution of $\mathbf{X}$.

Well-known methods

- Histogram
- Naive estimator
- Kernels

Quantities for comparing estimators

$$\forall x \in \mathcal{X},\ f(x) \to \quad \hat{f}(x) \text{ (depends on } \mathbf{X}) \to \quad \text{R.V.}$$

- Thus $X_1, \ldots, X_n$ are the random variables (with observed values $x_1, \ldots, x_n$) and x is a fixed value for which we wish to find $f(x)$.
- **Bias**: $\mathrm{Bias}(\hat{f}(x)) = \mathbb{E}\left[\hat{f}(x)\right] - f(x)$
- **Variance**: $\mathrm{Var}(\hat{f}(x)) = \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]$
- **Mean squared error**:

$$\mathrm{MSE}(\hat{f}(x)) = \mathrm{Bias}(\hat{f}(x))^2 + \mathrm{Var}(\hat{f}(x)) \ (*)$$

Problem: $\mathrm{MSE}(\hat{f}(x))$ concerns only one $x \in \mathcal{X}$.

- **Mean integrated squared error**:

$$\begin{aligned} \mathrm{MISE}(\hat{f}) &= \int_{\mathcal{X}} \mathrm{MSE}(\hat{f}(x))\mathrm{d}x \\ &= \int_{\mathcal{X}} \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] \mathrm{d}x \\ &= \mathbb{E}\left[\int_{\mathcal{X}} \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x\right] \end{aligned}$$

Comments

1. Proof of (*)

$$\begin{aligned}\mathrm{MSE}(\hat{f}(x)) &= \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] \\ &= \mathrm{Var}\left[\left\{\hat{f}(x) - f(x)\right\}\right] + \mathbb{E}\left[\hat{f}(x) - f(x)\right]^2 \\ &= \mathrm{Bias}(\hat{f}(x))^2 + \mathrm{Var}(\hat{f}(x))\,,\end{aligned}$$

since $f(x)$ is constant.

2. All expected values are computed w.r.t. the sampling distribution $f(\mathbf{x})$ (since they concern r.vs. which are functions of $X_1, X_2, \ldots X_n$). Thus,

$$\mathbb{E}[\hat{f}(x)] = \int_{\mathcal{X}^n} \hat{f}(x) f(\mathbf{x}) \mathrm{d}\mathbf{x}\,.$$

1. Histogram

Construction:
1. We compute # observations in each bin/interval.
2. We make a bar with height equal to the frequency of the values in that bin.

Be careful: Despite the simplicity in its construction, a wrong histogram might lead to wrong impressions.

We need to choose values for

- bin width
- # bins
- left boundary of the first bin

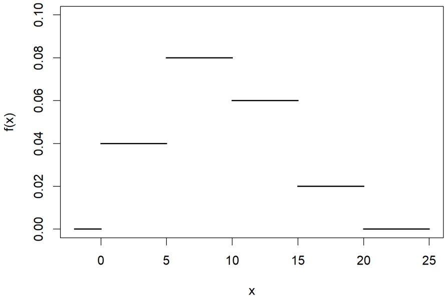
Knowing two of the above we can compute the third.

$$\hat{f}(x) = \frac{1}{n}\frac{\#\text{observations in the same bin as } x}{\text{bin width containing } x}$$

Histogram - Example

$n = 10 \;\rightarrow\; 2, 6, 8, 11, 14, 3, 5, 13, 19, 5$

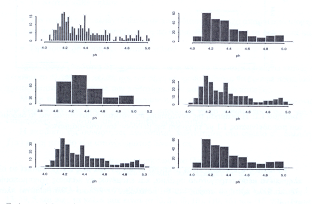
width 5, start 0: $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20)$



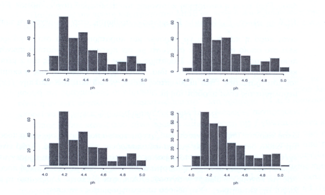
$$\hat{f}(x) = \begin{cases} 0.04 & x \in [0, 5) \\ 0.08 & x \in [5, 10) \\ 0.06 & x \in [10, 15) \\ 0.02 & x \in [15, 20) \end{cases}$$

Histogram - Example (cont'd)

Bin width?



Histogram start?



Bin width $\rightarrow$ $\min \mathrm{MISE}$

Histogram (cont'd)

If b_0: bin start

j bin: $[b_{j-1}, b_j)$

n_j: frequency of j bin

$h = b_j - b_{j-1}$: width

κ: number of bins, then

$$\hat{f}(x) = \frac{1}{n}\frac{n_j}{h}\ , \quad x \in [b_{j-1}, b_j)\ ,$$

where n, h are constants and
$n_j \sim Bin\,(n, F(b_j) - F(b_{j-1}) = \mathbb{P}\,(b_{j-1} \leq X < b_j))$. Thus

$$\begin{aligned}\mathbb{E}\big[\hat{f}(x)\big] &= \mathbb{E}\left[\frac{1}{n}\frac{n_j}{h}\right] = \frac{1}{nh}\mathbb{E}[n_j] \\ &= \frac{n[F(b_j) - F(b_{j-1})]}{nh} = \frac{[F(b_j) - F(b_{j-1})]}{h}\end{aligned}$$

Comments: 1) The expected value is the same $\forall x \in [b_{j-1}, b_j)$.
2) In general $\mathbb{E}\big[\hat{f}(x)\big] \neq f(x)$.

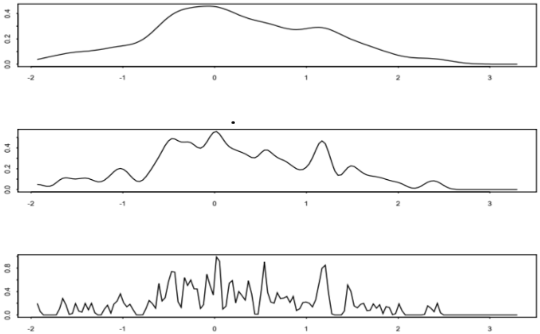
Histogram (cont'd)

- It can be shown that $\mathrm{Bias}(\hat{f}(x)) \approx \frac{1}{2} f'(x)\left[h - 2(x - b_{j-1})\right]$.
  To be unbiased $f'(x) = 0$, that is $f(x)$ is uniform on the interval under study.
- Also, $\mathrm{Var}(\hat{f}(x)) \approx \frac{f(x)}{nh}$.
- Thus, $\mathrm{MSE}(\hat{f}(x)) \approx \frac{1}{4}[f'(x)]^2\left[h - 2(x - b_{j-1})\right]^2 + \frac{f(x)}{nh}$.
  Note: as h increases, the variance decreases but the bias also increases!
- Finally, $\mathrm{MISE}(\hat{f}) \approx \frac{R(f')h^2}{12} + \frac{1}{nh}$, where $R(f') = \int_{\mathcal{X}} [f'(u)]^2 \mathrm{d}u$.
  Taking the derivative w.r.t. h

$$\frac{2R(f')h}{12} - \frac{1}{nh^2} = 0 \Rightarrow$$

$$h_{opt} = \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3}$$

$R(f')$ measures how smooth f is (the smoother f is the smaller $R(f')$ is). The smoother f is, the larger width we need - since for such a distribution its general image can be represented with a larger size window.



In the first graph above, we get a small value of $R(f')$, while in the last graph $R(f')$ is large.

Histogram (cont'd)

Problem: h_{opt} depends on unknown f!!

- For a normal distribution $h_{opt} = 3.491\sigma n^{-1/3}$ (see Appendix p. 28-29).

  If σ unknown

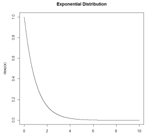
  $$h_{opt} = 3.491 s n^{-1/3} \ (s^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2) \text{ or}$$

  $$h_{opt} = 3.491\frac{IQR}{1.345}n^{-1/3}(IQR = \text{Interquantile Range})$$ - see next page

  The second type is preferred in the presence of outliers.
- If there is a symmetry but not normality (e.g. Student with few d.f.)

  $$h_{opt} = 3.491\tilde{s}n^{-1/3} \ ,$$

  where $\tilde{s} = \min\{IQR/1.345, s\}$
- If there is no symmetry, start with a distribution that makes sense $\rightarrow$ find $R(f') \rightarrow h_{opt}$. E.g.



Histogram (cont'd)

Note: For any normal distribution, 50% of the values lies approx. 0.6725 standard deviations of the mean:

$$IQR = Q_3 - Q_1 = 0.6725\sigma - (-0.6725\sigma) = 1.345\sigma$$

Optimum start This is not so important unless n is small.

See right panel on p. 7 (same width, different starts).

Averaging histograms: Same width different starts.

Usually $b_0 = \min x_i - h/2$

Conclusions (Histogram)

(+) Simple
(+) No assumptions for its construction
(-) The result is not smooth since the density is the same in each bin
(-) Difficult to generalize in higher dimensions

2. Naive estimator

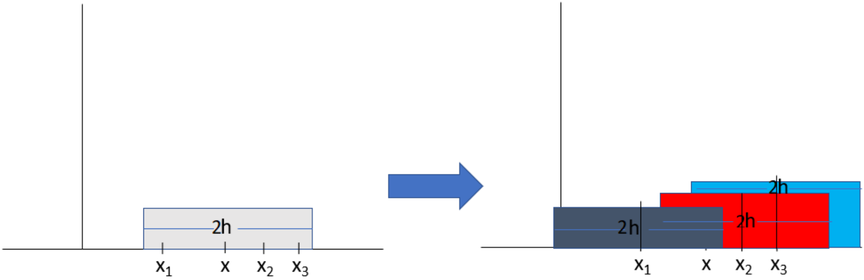
- Definition of density function: $f(x) = \lim_{h \to 0} \frac{\mathbb{P}(x-h<X<x+h)}{2h}$
- To estimate f we will use the sample equivalent, that is we define a small interval and count # observations in it.

$$\hat{f}(x) = \frac{1}{n} \frac{\#\text{observations in } \ (x-h, x+h)}{2h}$$

- Choice of h? Small $h \to$ non-smooth, large $h \to$ uniform.
- Difference with histogram: instead of having specific bins and counting how many observations fall in them, here we measure how many observations are in a specific distance from value x- that is we draw a box of width $2h$ and center x and count how many observations fall into it.

2. Naive estimator (cont'd)

$\Leftrightarrow$ We draw the box $(-h, +h)$ with width $2h$ and height $1/2nh$ around each observation x_i and for each x we measure how many boxes contain this x.



2. Naive estimator (cont'd)

- $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}W\left(\frac{x-x_i}{h}\right)$, where

$$W(y) = \begin{cases} 1/2 & y \in [-1,1] \\ 0 & \text{else} \end{cases}$$

  i.e. $W(y) \sim \mathrm{U}(-1,1)$, thus the estimate has jumps at $x_i \pm h$.
- Disadvantage: the result is not smooth, which in turn means that the derivatives do not exist everywhere.
- What if we choose another W?

3. Kernels

The disadvantage of the naive estimator is mainly due to the choice of the uniform distribution. The method of Kernels uses a kernel $K(x)$ in the place of $W(x)$:

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

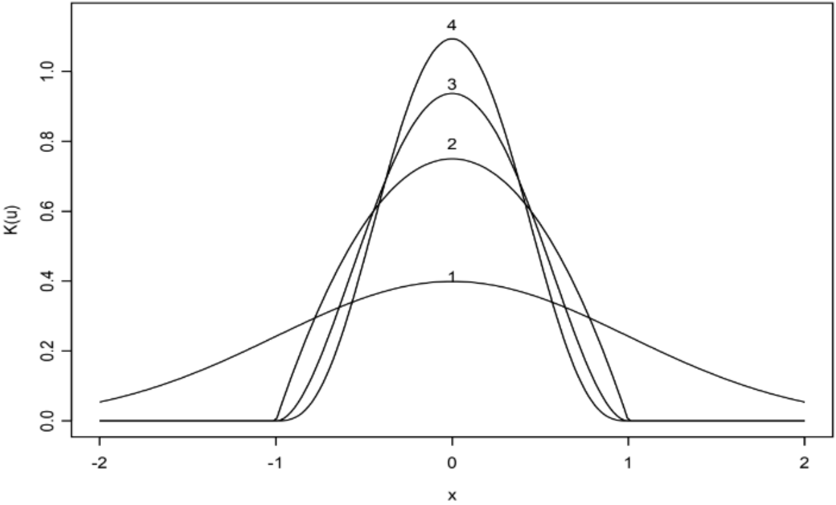
where the function $K(x)$ has the following properties:

1. $K(x) \geq 0 \ , \forall x$
2. $\int_{-\infty}^{\infty} K(x) \mathrm{d}x = 1$
3. $\int_{-\infty}^{\infty} x K(x) \mathrm{d}x = 0$
4. $\int_{-\infty}^{\infty} x^2 K(x) \mathrm{d}x \ (= \sigma_K^2) < \infty$

Widely-used Kernels

1. Gaussian $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \ x \in \mathbb{R}$
2. Epanechnikov $\frac{3}{4}(1 - x^2), \ |x| < 1$
3. Biweight $\frac{15}{16}(1 - x^2)^2, \ |x| < 1$
4. Triweight $\frac{35}{32}(1 - x^2)^3, \ |x| < 1$
5. Uniform ($\rightarrow$ Naive Estimator) $1/2, \ |x| < 1$

Kernels - Illustration



Kernels - Comments

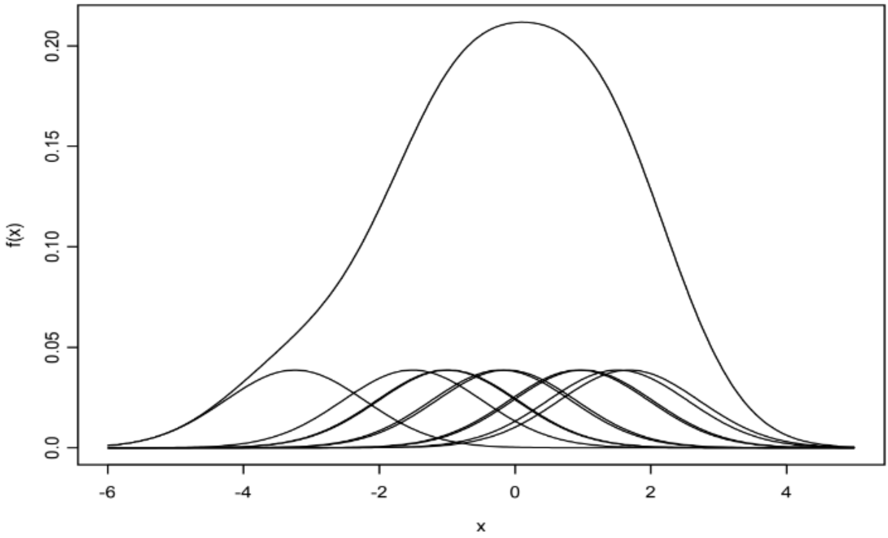
1. $\hat{f}(x)$ is a pdf. Indeed:
   i) $\hat{f}(x) \geq 0 \, , \forall x$
   ii)

$$\int_{-\infty}^{\infty} \hat{f}(x) \mathrm{d}x = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x - x_i}{h}\right) \mathrm{d}x$$

$$\overset{u=(x-x_i)/h}{=} \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} hK(u) \mathrm{d}u = \frac{nh}{nh} = 1$$

2. $h =$? (width/window/length)
   Big $h \rightarrow$ smooth estimate $\rightarrow$ losing information
   Small $h \rightarrow$ non-smooth estimate $\rightarrow$ non-descriptive
3. Which kernel?
4. How does the method work? For each observation we draw a kernel centered around this observation. Given the fact that each kernel is symmetric with mode/mean value 0, we assign bigger weight exactly on that observation and smaller as we move apart. The final estimator is the sum of the weights arising from each kernel.

Kernels - Illustration how it works



Kernels - Properties - Expected value

$$
\begin{aligned}
\mathbb{E}\left[\hat{f}(x)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} h^{-1} K\left(\frac{x-X_i}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbb{E}\left[K\left(\frac{x-X_i}{h}\right)\right] \underset{\text{call them } Y}{\overset{X_i \text{ i.i.d.}}{=}} \frac{1}{h}\mathbb{E}_Y\left[K\left(\frac{x-Y}{h}\right)\right] \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\mathrm{d}y \underset{\substack{y=x-hu\\ dy=-hdu\\ (h>0)}}{\overset{u=(x-y)/h}{=}} -\int_{\infty}^{-\infty} K(u)f(x-hu)\mathrm{d}u \\
&= -\lim_{a\to\infty}\int_{a}^{-a} K(u)f(x-hu)\mathrm{d}u = \lim_{a\to\infty} -\int_{a}^{-a} K(u)f(x-hu)\mathrm{d}u \\
&= \lim_{a\to\infty}\int_{-a}^{a} K(u)f(x-hu)\mathrm{d}u = \int_{-\infty}^{+\infty} K(u)f(x-hu)\mathrm{d}u \neq f(x)
\end{aligned}
$$

(i.e. there is bias which is independent of n)

Kernels - Properties (cont'd) - Bias

$$
\begin{aligned}
\mathrm{Bias}(\hat{f}(x)) &= \mathbb{E}\left[\hat{f}(x)\right] - f(x) = \int_{-\infty}^{+\infty} K(t) f(x-ht)\mathrm{d}t - f(x) \\
&\overset{\int_{-\infty}^{+\infty} K(t)\mathrm{d}t=1}{=} \int_{-\infty}^{+\infty} K(t)[f(x-ht) - f(x)]\mathrm{d}t \\
&\overset{\text{Taylor } (*)}{=} \int_{-\infty}^{+\infty} K(t)\left[-htf'(x) + \frac{1}{2}h^2t^2f''(x) + O(h^3)\right]\mathrm{d}t \\
&\overset{\int_{-\infty}^{+\infty} tK(t)\mathrm{d}t=0}{=} \frac{1}{2}h^2 f''(x)\int_{-\infty}^{+\infty} t^2K(t)\mathrm{d}t + O(h^3) \\
&\approx \frac{1}{2}h^2 f''(x)\sigma_K^2\,,
\end{aligned}
$$

where (*) $f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + O(h^3)\,.$

Kernels - Properties (cont'd) - Variance

$$
\begin{aligned}
\mathrm{Var}(\hat{f}(x)) &= \mathrm{Var}\left[\sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right)\right] \underset{\text{call them } Y}{\overset{X_i \text{ i.i.d.}}{=}} n\mathrm{Var}_Y\left[\frac{1}{nh} K\left(\frac{x - Y}{h}\right)\right] \\
&= \frac{1}{nh^2} \mathrm{Var}_Y\left[K\left(\frac{x - Y}{h}\right)\right] \\
&= \frac{1}{nh^2}\left[\mathbb{E}_Y\left[K^2\left(\frac{x - Y}{h}\right)\right] - \left(\mathbb{E}_Y\left[K\left(\frac{x - Y}{h}\right)\right]\right)^2\right] \\
&= \frac{1}{nh^2}\left[\int_{-\infty}^{\infty} K^2\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y - \left(\int_{-\infty}^{\infty} K\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y\right)^2\right] \\
&= n^{-1}\left[\int_{-\infty}^{\infty} h^{-2} K^2\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y - \left(\int_{-\infty}^{\infty} h^{-1} K\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y\right)^2\right] \\
&= n^{-1}\left[\int_{-\infty}^{\infty} h^{-2} K^2\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y - \left(\mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]
\end{aligned}
$$

Kernels - Properties (cont'd) - Variance

$$
\begin{aligned}
&= n^{-1}\int_{-\infty}^{\infty} h^{-2}K^2\left(\frac{x-y}{h}\right) f(y)\mathrm{d}y - n^{-1}\left(f(x)+\mathrm{Bias}(\hat{f}(x))\right)^2 \\
&\overset{y=x-ht}{=} n^{-1}h^{-1}\int_{-\infty}^{\infty} f(x-ht)K^2(t)\mathrm{d}t - n^{-1}\left(f(x)+O(h^2)\right)^2 \\
&\overset{\text{Taylor}, n \text{ large}}{\approx} n^{-1}h^{-1}\int_{-\infty}^{\infty}\left\{f(x)-htf'(x)+\ldots\right\}K^2(t)\mathrm{d}t + O(n^{-1}) \\
&\overset{h \text{ small}}{\approx} n^{-1}h^{-1}f(x)\int_{-\infty}^{\infty}K^2(t)\mathrm{d}t + O(n^{-1}) \\
&\approx \frac{f(x)R(K)}{nh},
\end{aligned}
$$

where $R(K) = \int_{-\infty}^{\infty} K^2(t)\mathrm{d}t$.

Kernels - Choice of optimum h

$$\begin{aligned}\mathrm{MSE}(\hat{f}(x)) &= \mathrm{Var}(\hat{f}(x)) + \mathrm{Bias}(\hat{f}(x))^2 \\ &\approx \frac{f(x)R(K)}{nh} + \frac{1}{4}h^4(f''(x))^2\sigma_K^4\end{aligned}$$

$$\mathrm{MISE}(\hat{f}) \approx \frac{R(K)}{nh} + \frac{1}{4}h^4 R(f'')\sigma_K^4 \ ,$$

where $R(f'') = \int_{-\infty}^{\infty}(f''(x))^2 \mathrm{d}x.$ (measures the rapidity of fluctuations in f.)

Thus, big $h \rightarrow$ Bias $\uparrow$ Variance $\downarrow$, small $h \rightarrow$ Bias $\downarrow$ Variance $\uparrow$

$$\begin{aligned}\frac{\mathrm{dMISE}(\hat{f})}{\mathrm{d}h} &= \frac{-R(K) + nh^5 R(f'')\sigma_K^4}{nh^2} = 0 \\ \Rightarrow h_{\mathrm{opt}} &= \left(\frac{R(K)}{nR(f'')\sigma_K^4}\right)^{1/5} \text{ and} \\ \mathrm{MISE}_{\mathrm{opt}} &= \tfrac{5}{4}[\sigma_K \underset{\text{kernel}}{\underset{\downarrow}{R(K)}}]^{4/5} \underset{\text{f}}{\underset{\downarrow}{R(f'')^{1/5}}} \underset{\text{sample size}}{\underset{\downarrow}{n^{-4/5}}}\end{aligned}$$

Kernels - Inefficiency

The quantity $\sigma_K R(K)$ is minimized for the Epanechnikov kernel and becomes equal to $3/(5\sqrt{5})$.

The quantity $in := \frac{\sigma_K R(K)}{3/(5\sqrt{5})}$ is called inefficiency.

It turns out that

$$\begin{aligned} in &= 1 \quad \text{Epanechnikov} \\ &= 1.0061 \quad \text{Biweight} \\ &= 1.0135 \quad \text{Triweight} \\ &= 1.0513 \quad \text{Gaussian} \\ &= 1.0758 \quad \text{Uniform} \;\; (\approx 7\% \text{ error}) \end{aligned}$$

In other words, the choice of Kernel is not that important.

Kernels - Practical computation of h_{opt}

	Gaussian Kernel	**Other Kernel**
Normal population	$h = 1.059 s n^{-1/5}$ (see Appendix p. 30-32)	$h = c1.059 s n^{-1/5}$
Symmetric population (e.g. Student) with heavier tails than normal	$h = 1.059 \tilde{s} n^{-1/5}$ $\tilde{s} = \min\left(\frac{\mathrm{IQR}}{1.345}, s\right)$	$h = c1.059 \tilde{s} n^{-1/5}$
Not Normal population	$h' = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5}$ $\frac{R(K)}{\sigma_K^4} = \frac{1}{2\sqrt{\pi}}$ $R(f'')$: initial estimate	h=ch'

Kernel	**c**
Epanechnikov	2.214
Biweight	2.693
Triweight	2.978
Gaussian	1.000
Uniform	1.740

Appendix

$f \to \mathcal{N}(0, \sigma^2)$ (centered data), $\phi \to \mathcal{N}(0, 1)$, i.e. $f(x) = 1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)$ and $\phi(x) = 1/(\sqrt{2\pi}) \exp\left(-\frac{x^2}{2}\right)$.

Then $f'(x) = 1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(-\frac{x}{\sigma^2}\right) = f(x)\left(-\frac{x}{\sigma^2}\right)$ and $\phi'(x) = \phi(x)(-x)$.

Thus,

$$
\begin{aligned}
R(f') &= \int_{-\infty}^{\infty} (f'(x))^2 \mathrm{d}x = \int_{-\infty}^{\infty} \left(f(x)\left(-\frac{x}{\sigma^2}\right)\right)^2 \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(-\frac{x}{\sigma^2}\right)\right)^2 \mathrm{d}x \\
&\underset{\mathrm{d}x=\sigma \mathrm{d}y}{\overset{x^2/\sigma^2=y^2}{=}} \int_{-\infty}^{\infty} \left(\phi(y)\left(-\frac{y}{\sigma}\right)\right)^2 \frac{1}{\sigma} \mathrm{d}y \\
&= \sigma^{-3} \int_{-\infty}^{\infty} (\phi'(y))^2 \mathrm{d}y
\end{aligned}
$$

Appendix (cont'd)

$$
\begin{aligned}
R(f') &= \sigma^{-3}\int_{-\infty}^{\infty} 1/(2\pi)\exp\left(-y^2\right)y^2 \mathrm{d}y \\
&\underset{\mathrm{d}y=\mathrm{d}x/\sqrt{2}}{\overset{y=x/\sqrt{2}}{=}} \sigma^{-3}\int_{-\infty}^{\infty} 1/(2\pi)\exp\left(-\frac{x^2}{2}\right)\frac{x^2}{2}\frac{1}{\sqrt{2}}\mathrm{d}x \\
&= \frac{1}{4\pi\sqrt{2}}\sigma^{-3}\sqrt{2\pi}\int_{-\infty}^{\infty} x^2\phi(x)\mathrm{d}x = \frac{1}{4\pi\sqrt{2}}\sigma^{-3}\sqrt{2\pi} \\
&= \frac{1}{4}\pi^{-1/2}\sigma^{-3}.
\end{aligned}
$$

Thus, for the histogram, under normality assumption,

$$
h_{opt} = \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3} = 2 \times 3^{1/3}\pi^{1/6}\sigma n^{-1/3} = 3.491\sigma n^{-1/3}
$$

Appendix (cont'd)

$f \to \mathcal{N}(0, \sigma^2)$ (centered data), $\phi \to \mathcal{N}(0, 1)$, i.e.
$f(x) = 1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)$ and $\phi(x) = 1/(\sqrt{2\pi}) \exp\left(-\frac{x^2}{2}\right)$.

Then $f'(x) = 1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(-\frac{x}{\sigma^2}\right) = f(x)\left(-\frac{x}{\sigma^2}\right)$ and
$f''(x) = f'(x)\left(-\frac{x}{\sigma^2}\right) + f(x)\left(-\frac{1}{\sigma^2}\right) = f(x)\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)$.

Thus,

$$
\begin{aligned}
R(f'') &= \int_{-\infty}^{\infty} (f''(x))^2 \mathrm{d}x = \int_{-\infty}^{\infty} \left(f(x)\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\right)^2 \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(1/(\sqrt{2\pi\sigma^2}) \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\right)^2 \mathrm{d}x \\
&\underset{\mathrm{d}x = \sigma \mathrm{d}y}{\overset{x^2/\sigma^2 = y^2}{=}} \int_{-\infty}^{\infty} \left(\phi(y)\left(\frac{y^2}{\sigma^2} - \frac{1}{\sigma^2}\right)\right)^2 \frac{1}{\sigma} \mathrm{d}y \\
&= \sigma^{-5} \int_{-\infty}^{\infty} (\phi''(y))^2 \mathrm{d}y
\end{aligned}
$$

Appendix (cont'd)

But $\phi'(x) = \phi(x)(-x)$ and $\phi''(x) = \phi(x)x^2 + \phi(x)(-1) = \phi(x)(x^2 - 1)$.
Thus,

$$
\begin{aligned}
R(f'') &= \sigma^{-5} \int_{-\infty}^{\infty} (\phi(y)(y^2 - 1))^2 \mathrm{d}y \\
&= \sigma^{-5} \int_{-\infty}^{\infty} 1/(2\pi) \exp\left(-y^2\right)(y^4 - 2y^2 + 1)\mathrm{d}y \\
&\underset{\mathrm{d}y=\mathrm{d}x/\sqrt{2}}{\overset{y=x/\sqrt{2}}{=}} \sigma^{-5} \int_{-\infty}^{\infty} 1/(2\pi) \exp\left(-\frac{x^2}{2}\right)(x^4/4 - 2x^2/2 + 1)1/\sqrt{2}\mathrm{d}x \\
&= \frac{\sigma^{-5}}{\sqrt{2\pi}\sqrt{2}} \left[\frac{1}{4}\int_{-\infty}^{\infty} x^4\phi(x)\mathrm{d}x - \int_{-\infty}^{\infty} x^2\phi(x)\mathrm{d}x + 1\right] \\
&= \frac{1}{2}\pi^{-1/2}\sigma^{-5}\left[\frac{1}{4}3 - 1 + 1\right] \\
&= \frac{3}{8}\pi^{-1/2}\sigma^{-5}.
\end{aligned}
$$

Appendix (cont'd)

If in addition K is the Gaussian kernel

$$
\begin{aligned}
R(K) &= \int_{-\infty}^{\infty}\left(1/(\sqrt{2\pi})\exp\left(-\frac{x^2}{2}\right)\right)^2 \mathrm{d}x = \int_{-\infty}^{\infty} 1/(2\pi)\exp\left(-x^2\right)\mathrm{d}x \\
&= 1/(\sqrt{2\pi})\int_{-\infty}^{\infty} 1/(\sqrt{2\pi})\exp\left(-x^2\right)\mathrm{d}x \\
&\overset{x=y/\sqrt{2}}{=} 1/(\sqrt{2\pi})\int_{-\infty}^{\infty}\phi(y)1/\sqrt{2}\mathrm{d}y = \frac{1}{2}\pi^{-1/2} = (4\pi)^{-1/2}.
\end{aligned}
$$

Thus, for the kernel approach, under normality

$$
\begin{aligned}
h_{\mathrm{opt}} &= \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5} \\
&\underset{\text{Gaussian}}{\overset{\text{for}}{=}} (4\pi)^{-1/10}\left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5}\sigma n^{-1/5} \approx 1.059\sigma n^{-1/5}.
\end{aligned}
$$

Kernels - Simulation

- Choose one observation from the sample at random, denoted by X, with probability $1/n$.
- $Y \sim K$ (easy if K Gaussian) (*)
- $Z = X + hY \sim \hat{f}(x)$

* $K \rightarrow$ Epanechnikov, i.e. $K(x) = 3/4(1 - x^2),\ |x| < 1$

$V_1, V_2, V_3 \sim U[-1, 1]$

If $|V_3| \geq |V_2|$ and $|V_3| \geq |V_1|$ then $Y = V_2$ else $Y = V_3$.

Kernels - Other criteria for selecting h_{opt}

- Mean integrated absolute error

$$\mathrm{MIAE}(\hat{f}) = \int \mathbb{E}\left[\left|\hat{f}(x) - f(x)\right|\right] \mathrm{d}x$$

- Cross-validation

  Idea (Leave-one-out)

  We leave one observation out, we fit our model using the rest and afterwards we test how well our model can predict the observation we left out. We repeat the procedure n times (leaving each time a different observation) and at the end get an overall score.

Kernels - Other criteria for selecting h_{opt} (cont'd)

- Cross-validation (cont'd)

  Models $\rightarrow$ Kernel estimators with different h's.

  Given observations $x_1, x_2, \ldots x_n$ and h, the likelihood is $L(h) = \prod_{i=1}^{n} \hat{f}_h(x_i)$ and is maximized for $h = 0$!

  Let $\hat{f}_{h,-i}(x) = (n-1)^{-1} h^{-1} \sum_{j=1, j \neq i}^{n} K\left(\frac{x - x_j}{h}\right)$ be the estimate without observation i.

  The cross-validated likelihood is given by:

$$L(h, i) = \prod_{i=1}^{n} \hat{f}_{h,-i}(x_i)$$

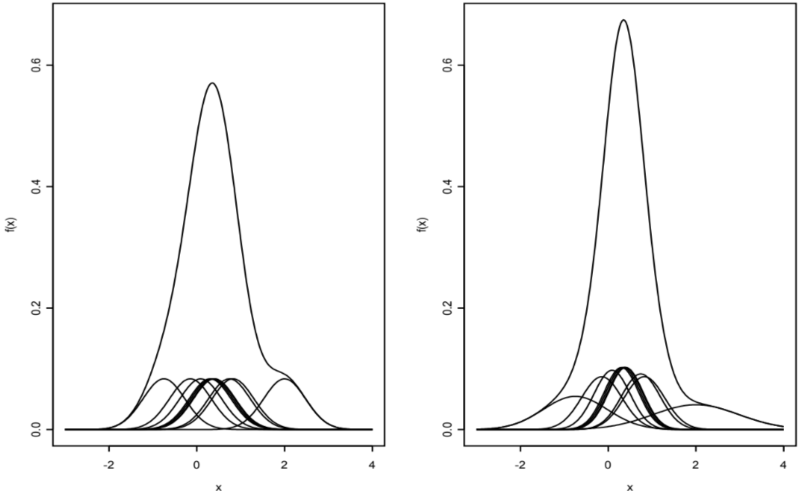
  Find h that maximizes it!

Kernels with variable width

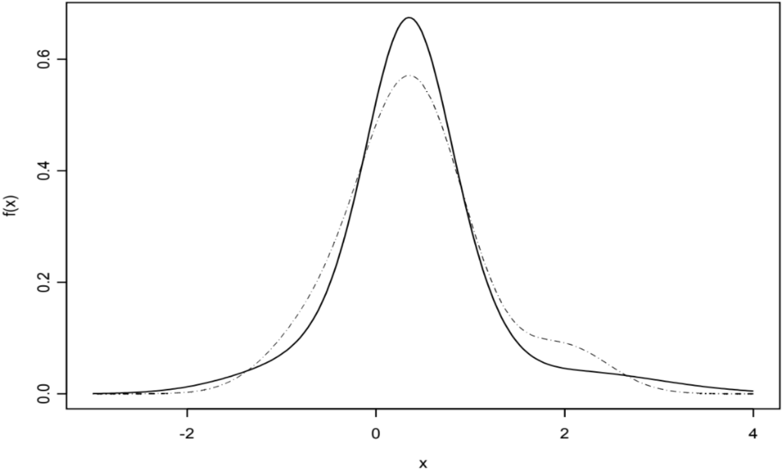
Useful for data with outliers

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

Kernels with variable width (cont'd)



Kernels with variable width (cont'd)



Kernels with variable width (cont'd)

Options:

1.
   - Find optimal fixed h
   - Compute $\hat{f}(x)$ based on h
   - $h(x_i) = \frac{h}{\sqrt{\hat{f}(x_i)}}$
   - Compute new $\hat{f}(x)$ based on $h(x_i)$
2.
   - Compute $\hat{f}(x)$ for some fixed h
   - Geometric mean $G = \left[\prod_{i=1}^{n} \hat{f}(x_i)\right]^{(1/n)}$
   - $\lambda_i = \sqrt{\frac{G}{\hat{f}(x_i)}}$
   - $h_i = h\lambda_i$

Kernels - Multivariate data

Let $\mathbf{X} = (X_1, X_2 \ldots X_d)$ r.v.'s with joint pdf $f(x_1, x_2 \ldots x_d) \to$?
$\mathbf{x}_i$: n observations (d-dimensional each) for each $X_1, X_2 \ldots X_d$. Then

$$\hat{f}(x_1, x_2 \ldots x_d) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K_d \left[\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right], \quad \forall \mathbf{x} = (x_1, x_2 \ldots x_d)$$

$\mathbf{H}$ is a $d \times d$ matrix. E.g.

$$\mathbf{H} = \begin{bmatrix} h & \ldots & 0 \\ \vdots & \ddots & \\ 0 & & h \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} h_1 & \ldots & 0 \\ \vdots & \ddots & \\ 0 & & h_d \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} h_{11} & h_{21} & \ldots & h_{d1} \\ \vdots & \ddots & & \\ & & & \end{bmatrix}$$

(diagonal) (same width $\forall d$) — (diagonal) (different width for each d) — (symmetric) (correlation)

The kernels could either be the product of independent kernels in each dimension or d-dimensional kernels, e.g. Epanechnikov:

$$K_d(\mathbf{x}) = \begin{cases} \frac{d(d+2)}{4}\Gamma(d/2)n^{-d/2}(1 - \mathbf{x}^T\mathbf{x}) & \mathbf{x}^T\mathbf{x} \leq 1 \\ 0 & \text{else} \end{cases}$$

4. Categorical data

Let X be a categorical r.v. (with k categories) with pmf f. Suppose we have a sample of size n. Denote by n_j the frequency of each category $j = 1, \dots k$.

$$f \to \hat{p}_j = \frac{n_j}{n}$$

This might create issues in case of zero frequencies in small samples.

Correction/Smoothing $\to \hat{p}_j = \frac{n_j + a}{n + ak}$

$$a = \begin{cases} z^{-1} & z \geq 1 \\ 1 & z < 1 \end{cases}$$

where $z = \frac{1}{k} \sum_{j=1}^{k} \frac{(n_j - n/k)^2}{n/k}$, i.e. z corresponds to the value of Pearson's χ^2 test statistic for testing the null hypothesis that all categories are of equal probability divided by the number of categories k.

Big value of $z \Rightarrow$ we reject the null hypothesis (of equal probability)

$$\Downarrow$$

$$a \to 0 \Rightarrow \hat{p}_j = \frac{n_j}{n}$$

4. Categorical data (cont'd)

It can be shown that

$$\hat{p}_j = \frac{\epsilon}{k} + (1-\epsilon)\frac{n_j}{n}, \quad \epsilon = \frac{ak}{n+ak}$$

$$\left(\begin{aligned} &= \frac{ak}{nk+ak^2} + \frac{n+ak-ak}{n+ak}\frac{n_j}{n} \\ &= \frac{ak}{nk+ak^2} + \frac{nn_j}{n^2+nak} \\ &= \frac{a}{n+ak} + \frac{n_j}{n+ak} = \frac{a+n_j}{n+ak} \end{aligned} \right)$$

- $\epsilon \approx 1 \rightarrow \hat{p}_j = \frac{1}{k}$
- $\epsilon \approx 0 \rightarrow \hat{p}_j = \frac{n_j}{n}$

Weighted average between the relative frequencies $\frac{n_j}{n}$ and the case of equal probability events ($\frac{1}{k}$).

- small a ($\iff$ not of equal probability) $\rightarrow \epsilon$ small $\rightarrow 1-\epsilon$ large.

4. Categorical data (cont'd)

Further, the estimator $\hat{p}_i$ is related to kernels!

$$\hat{p}_i = \sum_{j=1}^{k} \frac{n_j}{n} W_j(i, \lambda), \text{ where}$$

$$W_j(i, \lambda) = \begin{cases} \lambda & j = i \\ \frac{1-\lambda}{k-1} & j \neq i \end{cases}$$

is a kernel.

Thus, $\hat{p}_i = \frac{n_i}{n}\lambda + \frac{1-\lambda}{k-1}\frac{n-n_i}{n}$

For $\lambda = 1 \rightarrow \hat{p}_i = \frac{n_i}{n}$, $\quad \lambda = 1/k \rightarrow \hat{p}_i = \frac{1}{k}$

The kernel is telling us that once observing the value j, then the probability for it to be correct (in the sense that we have all the information about that value) is λ while all other categories have probability $\frac{1-\lambda}{k-1}$ each.
It holds true that:

$$\lambda = 1 - \frac{\epsilon(k-1)}{k}, \quad \epsilon = \frac{k(1-\lambda)}{k-1}.$$

Categorical data - Example

$n = 50$ individuals

5 parties: A, B, C, D, E

$$\begin{array}{ccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 20, & 18, & 7, & 5, & 0 \end{array}$$

$$z = \frac{1}{5}\left[\frac{(20 - 50/5)^2}{10} + \ldots + \frac{(0 - 50/5)^2}{10}\right] = 5.96$$

$$\Rightarrow a = 1/5.96 = 0.167$$

	n_j	n_j/n	$\hat{p}_j$
A	20	0.4	0.39
B	18	0.36	0.35
C	7	0.14	0.14
D	5	0.10	0.10
E	0	0	0.003

($\rightarrow$ small changes because categories not of equal probability)

$\epsilon = \left(\frac{ak}{n+ak} =\right) 0.0165,\ \lambda = 0.9868$

5. Ordinal data

$$W_j(i,\lambda) = \begin{cases} \lambda & j = i \\ \frac{1-\lambda}{2^{|i-j|+1}} & 0 < |i-j| \leq j \\ \frac{1-\lambda}{2^{|i-j|}} & |i-j| > j \end{cases}$$

λ is chosen using different criteria

6. Non-parametric regression

Let Y, X be two r.v. We have data $(y_i, x_i), i = 1, \ldots, n$.
Simple linear model: $m(x) = \mathbb{E}[Y|X = x] = \alpha + \beta x$ (why linearity?)

$$\begin{aligned} m(x) &= \mathbb{E}[Y|X = x] = \int y f(y|x) \mathrm{d}y \\ &= \int y \frac{f(x,y)}{f_X(x)} \mathrm{d}y = ? \quad (f(x,y) = ?, f_X(x) = ?) \end{aligned}$$

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

(i.e. we use a product of independent kernels)

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

6. Non-parametric regression (cont'd)

Thus,

$$\begin{aligned}
\hat{m}(x) &= \int y \frac{\hat{f}(x,y)}{\hat{f}_X(x)} \mathrm{d}y \\
&= \int \frac{y}{\hat{f}_X(x)} \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) K_y\left(\frac{y-y_i}{h_y}\right) \mathrm{d}y \\
&= \frac{1}{\hat{f}_X(x)} \sum_{i=1}^{n} \frac{1}{nh_x} K_x\left(\frac{x-x_i}{h_x}\right) \int \frac{y}{h_y} K_y\left(\frac{y-y_i}{h_y}\right) \mathrm{d}y \\
&\overset{u=(y-y_i)/h_y}{=} \frac{1}{\hat{f}_X(x)} \sum_{i=1}^{n} \frac{1}{nh_x} K_x\left(\frac{x-x_i}{h_x}\right) \int [uh_y + y_i] K_y(u) \mathrm{d}u \\
&\underset{\int uK(u)\mathrm{d}u=0}{\overset{\int K(u)\mathrm{d}u=1}{=}} \frac{1}{\hat{f}_X(x)} \sum_{i=1}^{n} \frac{1}{nh_x} K_x\left(\frac{x-x_i}{h_x}\right) y_i \\
&= \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) y_i}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)} = \sum_{i=1}^{n} w_i y_i = \hat{m}_{NW}(x)
\end{aligned}$$

→ Nadaraya-Watson

- no assumption!
- h_x related to smoothing
  - $h_x \to 0$: we just connect the observed points $\to$ non-smooth $\to$ estimator is 0 for all other points (see bottom right plot next slide) - **overfitting** - **high variance**
  - $h_x \uparrow$: $\hat{m}(x) = \bar{y}$!! (see top left plot next slide) - **underfitting** - **high bias**
  - choice of h_x? Usually cross-validation
- Generalization in more dimensions

Non-parametric regression - Illustration