

# Week 3: Kernel Density Estimation

## MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-09-27

# Section 1

## Univariate Density Estimation

# The problem

**Setup:**  $X_1, \dots, X_n$  is a random sample from a distribution  $F$  with continuous density  $f(x)$

**Goal:** Estimate  $f$  non-parametrically, i.e., without assuming a particular form

The **histogram**

$$h(x) = \frac{1}{n} \sum_{k=1}^K \frac{\#\left\{\{X_j\}_{j=1}^n \cap \mathcal{J}_k\right\}}{|\mathcal{J}_k|} \mathbb{1}_{x \in \mathcal{J}_k}$$

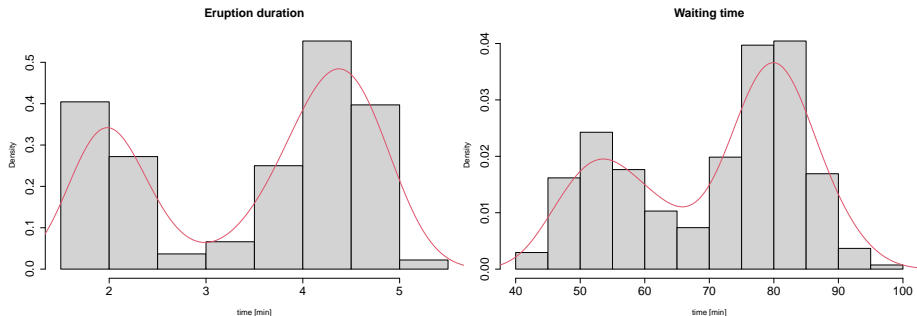
is the simplest form of density estimation. It requires a specification of

- *origin* and *binwidth*, or
- *breaks*: more general, but non-equidistant binning is bad anyway, so think only about origin and binwidth

**Running Ex.:** Yellowstone's Old Faithful geyser - faithful data:

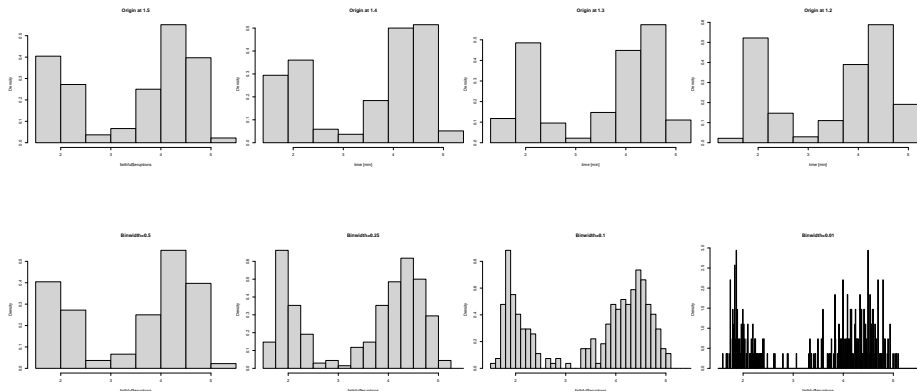
- waiting - time between eruptions
- eruptions - duration of the eruptions

# Basic estimator: Histogram



(equally spaced) breaks specified, so a rule of thumb used to choose origin and binwidth

# Histogram: Change in Origin and Binwidth

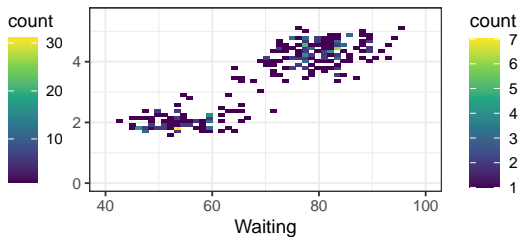
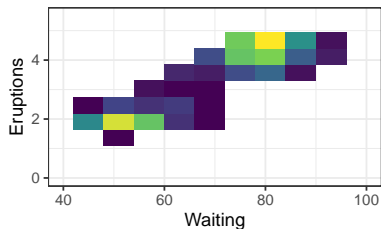


⇒ density estimate depends on the starting position and width of the bins

# Issues with Histogram

Histogram is great for visualization, but fails as a density estimator

- *origin* is completely arbitrary
- *binwidth* relates to smoothness of  $f$ , but histogram cannot be smooth. The discontinuities of the estimate are not due to the underlying density but to bins' locations and widths
- *curse of dimensionality*: number of bins grows exponentially with the number of dimensions

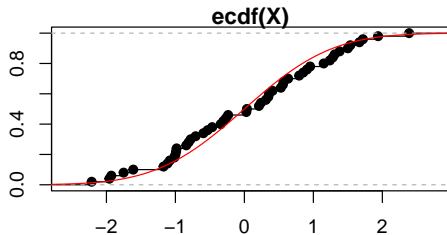
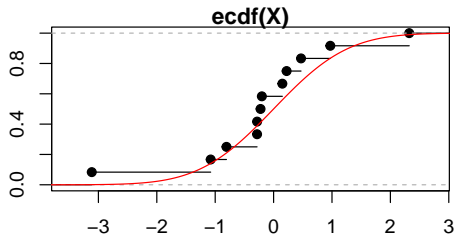


Let us now address these issues by a naive version of kernel density estimation

# ECDF

Let  $\widehat{F}$  denote the empirical (cumulative) distribution function (ECDF) of the data  $\{X_i\}_{i=1}^n$ , i.e.,

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[X_i \leq x]}$$



# Naive Density Estimator

- The ECDF  $\widehat{F}_n(x)$  is an estimator of  $F$ 
  - by [Glivenko-Cantelli theorem](#) uniformly almost surely consistent:

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

- Note that  $f$  is the derivative of  $F$ . However, plugging  $\widehat{F}_n(x)$  results in a sum of point masses at the observations as  $\widehat{F}_n$  is discrete
- But,

$$f(x) = \lim_{h \rightarrow 0_+} \frac{F(x+h) - F(x-h)}{2h}$$

and we can fix  $h = h_n$  small and depending on  $n$ , and plug it in:

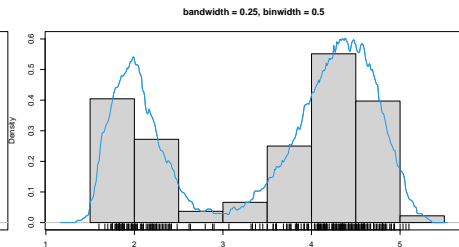
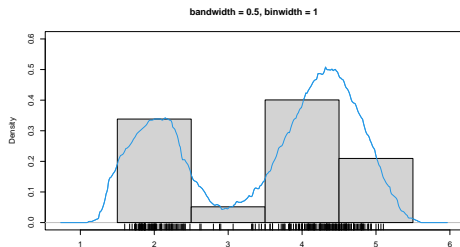
$$\widehat{f}_n(x) = \frac{\widehat{F}_n(x+h_n) - \widehat{F}_n(x-h_n)}{2h_n} = \frac{1}{2h_n} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[X_i \in (x-h_n, x+h_n] ]}$$

⇒ This is called the naive density estimator



# Naive Density Estimator

- The naive DE  $\hat{f}$  is a step function with jumps at the points  $X_i \pm h$ , and thus discontinuous
- $\hat{f}$  is the sum of boxcar functions centered at the observations with width  $2h$  and area  $1/n \Rightarrow$  this is equivalent to the notion of moving histogram with binwidth= $2h$ 
  - aggregate data in intervals of the form  $(x - h, x + h)$  and approximate the density at  $x$  by the relative frequency in  $(x - h, x + h)$
  - *origin* does not matter anymore



# Consistency

**Theorem** If  $h = h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$  as  $n \rightarrow \infty$ , then, for any continuity point  $t$ ,

$$\hat{f}_n(t) \xrightarrow{p} f(t),$$

as  $n \rightarrow \infty$ . Thus,  $\hat{f}_n$  is a (weakly) consistent estimator

For instance, since

$$\hat{f}(x) = \frac{1}{2nh_n} \sum_{i=1}^n \overbrace{\mathbb{1}_{[X_i \in (x-h_n, x+h_n)]}}^{\text{Ber}(F(x+h_n)-F(x-h_n))}$$

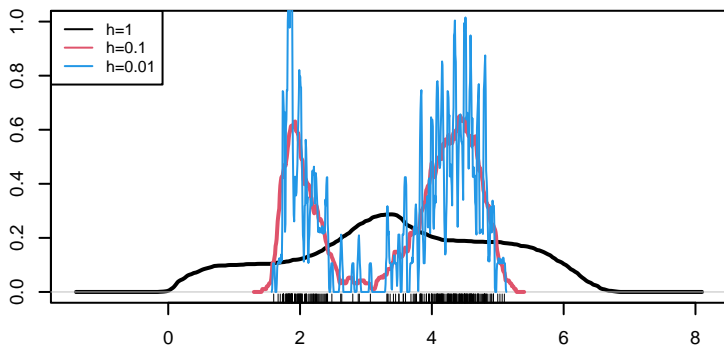
- $\mathbb{E}\hat{f}(x) = \frac{F(x+h_n)-F(x-h_n)}{2h_n} \rightarrow f(x)$  as  $h_n \rightarrow 0_+$ , when  $n \rightarrow \infty$
- $\text{var}\{\hat{f}(x)\} = \frac{1}{4nh_n^2} \{F(x+h_n) - F(x-h_n)\} \{1 - F(x+h_n) + F(x-h_n)\}$   
 $= \frac{F(x+h_n)-F(x-h_n)}{2h_n} \frac{1-F(x+h_n)+F(x-h_n)}{2nh_n} \rightarrow 0$

as  $h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$ , when  $n \rightarrow \infty$

# Smoothness of the Naive DE

Smoothness of  $\hat{f}$  depends on the *bandwidth*  $h$ , often called the *smoothing parameter*

- the bandwidth  $h$  is a *tuning parameter* and needs to be chosen somehow in practice
  - $h$  small  $\rightarrow$  wiggly estimator
  - $h$  large  $\rightarrow$  smooth estimator



# Naive DE Rewritten

The naive DE can be written as

$$\begin{aligned}\hat{f}(x) &= \frac{1}{2nh_n} \sum_{i=1}^n \mathbb{1}_{[X_i \in (x-h_n, x+h_n)]} = \frac{1}{2nh_n} \sum_{i=1}^n \mathbb{1}_{[-1 < \frac{X_i - x}{h_n} \leq 1]} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right)\end{aligned}$$

where  $K(t) = \frac{1}{2} \mathbb{1}_{\{-1 < t \leq 1\}}$  is the density of  $U[-1, 1]$ .

- Since  $\int_{-\infty}^{+\infty} K(t) dt = 1$ , we have that

$$\int_{-\infty}^{+\infty} \hat{f}(x) dx = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n} \int_{-\infty}^{+\infty} K\left(\frac{x - X_i}{h_n}\right) dx = 1$$

- Since  $K(x) \geq 0$ , then  $\hat{f}(x) \geq 0$  for all  $x$ .

$\Rightarrow \hat{f}(x)$  is a probability density function

**Next step:** replace  $K(x)$  by another probability density, maybe one giving more weight to points closer to  $x$  ?

# KDE - Definition and Properties

**Definition.** KDE of  $f$  based on  $X_1, \dots, X_N$  is

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right),$$

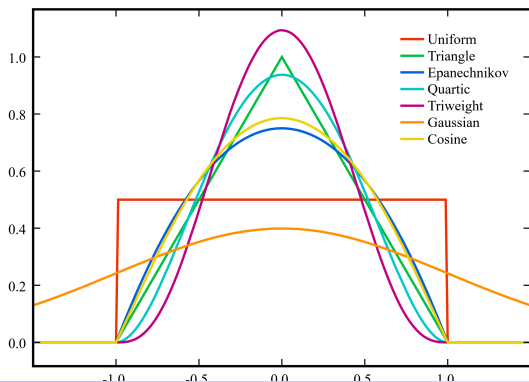
where the **kernel**  $K(\cdot)$  satisfies:

- ①  $K(x) \geq 0$  for all  $x \in \mathbb{R}$
- ②  $K(-x) = K(x)$  for all  $x \in \mathbb{R}$
- ③  $\int_{\mathbb{R}} K(x) dx = 1$
- ④  $\lim_{|x| \rightarrow \infty} |x|K(x) = 0$
- ⑤  $\sup_x |K(x)| < \infty$

- $K(\cdot)$  is usually taken to be a density, and the assumptions
  - 1-3 hold if it is symmetric
  - 4 holds if it has a finite absolute moment
  - 5 holds if it is uniformly bounded
- if  $h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$  (as  $n \rightarrow \infty$ ), we have pointwise consistency
  - we will show this in a bit
  - also uniform consistency, but tricky to show

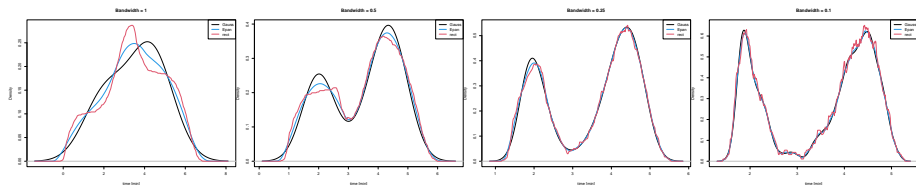
# Common Kernels

Kernel Name	Formula
Epanechnikov	$K(x) \propto (1 - x^2)\mathbb{1}_{\{ x  \leq 1\}}$
Tricube (a.k.a. Triweight)	$K(x) \propto (1 -  x ^3)^3\mathbb{1}_{\{ x  \leq 1\}}$
Gaussian	$K(x) \propto \exp(-x^2/2)$
...	...



# Bandwidth $>$ Kernel

- While there is improvement when using non-rectangular kernels, the choice of the bandwidth is more important than that of the kernel
- A good choice is one that makes the estimate asymptotically converge quite rapidly in some well-chosen norm



# Bias-Variance Trade-off

Goal: choose the tuning parameter  $h$  so that the mean squared error of the estimator is minimized:

$$\begin{aligned}\underbrace{\mathbb{E}[\{\hat{f}(x) - f(x)\}^2]}_{MSE\{\hat{f}(x)\}} &= \mathbb{E}[\{\hat{f}(x) \pm \mathbb{E}\hat{f}(x) - f(x)\}^2] \\ &= \underbrace{\{\mathbb{E}\hat{f}(x) - f(x)\}^2}_{bias^2} + \underbrace{\text{var}\{\hat{f}(x)\}}_{variance}\end{aligned}$$

**Blackboard calculations** (available in the lecture notes) give

$$\begin{aligned}\text{bias}\{\hat{f}(x)\} &= \frac{1}{2}h_n^2 f''(x) \int z^2 K(z) dz + o(h_n^2) \\ \text{var}\{\hat{f}(x)\} &= \frac{1}{nh_n} f(x) \int \{K(z)\}^2 dz + o\left(\frac{1}{nh_n}\right)\end{aligned}$$

This shows consistency for  $h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$  and the trade-off:

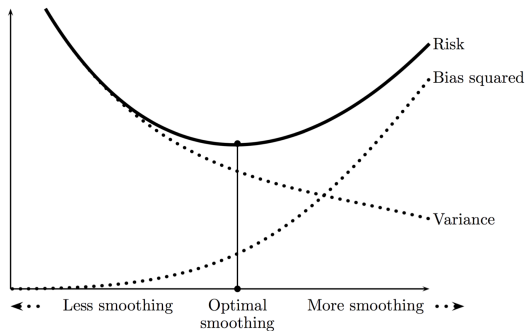
- small  $h \Rightarrow$  small bias but large variance
- large  $h \Rightarrow$  large bias but small variance



# Bias-Variance Trade-off

The bias-variance trade-off is common when it comes to smoothing:

- In KDE, the smoothing is determined by the bandwidth
- Smoother estimates result in smaller variance but higher bias



Source: Wassermann (2006)

# Optimal Bandwidth

Plugging this back in the MSE formula ignoring the *little-o* terms, differentiating the MSE w.r.t.  $h$  and setting it to zero leads to asymptotically optimal bandwidth choice:

$$h_{opt}(x) = n^{-1/5} \left( \frac{f(x) \int K(z)^2 dz}{[f''(x)]^2 [\int z^2 K(z) dz]^2} \right)^{1/5}$$

- $h_{opt}(x) \asymp n^{-1/5}$  ( $h_{opt}(x)$  is of the order  $n^{-1/5}$ ) and with this choice  $MSE \asymp variance = \mathcal{O}(n^{-4/5})$ 
  - optimal non-parametric convergence rate
  - slower than the MSE of a MLE ( $\mathcal{O}(n^{-1})$ ): price to pay for non-parametric approach
- $h_{opt}(x)$  is a local choice - depends on  $x$ . A global choice can be obtained by minimizing the MISE (to follow)

# Optimal Kernel

The optimal bandwidth results in

$$MSE_{h_{opt}} = c(n, f) \left[ \underbrace{\int x^2 K(z) dz \left\{ \int K^2(z) dz \right\}^2}_A \right]^{2/5},$$

where  $c(n, f)$  is constant and depends only on  $n$  and  $f$ .

⇒ The optimal kernel is the one minimizing the term  $A$ .

It can easily be shown to be the Epanechnikov kernel!

# Global Optimal Bandwidth

A common measure of performance of the estimator over all  $x$  is the Mean Integrated Squared Error (MISE):

$$\begin{aligned} MISE(\hat{f}) &= \mathbb{E} \int \{\hat{f}(x) - f(x)\}^2 dx \\ &= \int \mathbb{E}\{\hat{f}(x) - f(x)\}^2 dx = \int MSE\{\hat{f}(x)\} dx \end{aligned}$$

Minimizing the MISE yields the optimal bandwidth

$$\tilde{h}_{opt} = n^{-1/5} \left( \frac{\int K^2(z) dz}{\int \{f''(x)\}^2 dx [\int z^2 K(z) dz]^2} \right)^{1/5}$$

The resulting MISE is also of order  $n^{-4/5}$

# The Chicken and Egg problem

The theoretically optimal bandwidth  $\tilde{h}_{opt} = n^{-1/5} (C(k) / \int \{f''(x)\}^2 dx)^{1/5}$  cannot be directly used as it depends on the unknown  $f$ . There are different approaches for the practical choice of  $h$ .

- **Reference method:** choose a parametric family for this formula
  - assume that  $f$  is the density of a  $\mathcal{N}(\mu, \sigma^2)$  and then plug in its curvature  $\frac{3}{8\sqrt{\pi}\sigma^5}$  into the formula of  $\tilde{h}_{opt}$ . This yields

$$\tilde{h}_{opt} = n^{-1/5} \sigma C(k)^{1/5} (8\sqrt{\pi}/3)^{1/5}$$

which when combined with a normal kernel gives the famous rule of thumb  $\hat{h}_{opt} = (4/3)^{1/5} n^{-1/5} \hat{\sigma}$

- **Two-step method:**  $f$  in the formula is estimated non-parametrically by a pilot fit
  - estimate  $f''$  by kernel estimate with pilot bandwidth
  - plug this estimate into  $\tilde{h}_{opt}$  to estimate the optimal bandwidth in the kernel estimation of  $f$

## Section 2

# Multivariate Density Estimation

# Multivariate Density estimator

In practice, data are often multivariate

Consider  $n$  i.i.d. realizations of a  $d$ -dimensional random vector  $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,d})^\top$  from unknown  $F$ . We wish to estimate  $f$ , the density of  $F$

The multivariate kernel density estimator is defined as

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right),$$

where the kernel  $K(\cdot)$  is a  $d$ -dimensional density

# Multivariate Kernel

In practice,  $K$  is often chosen as

- the product of univariate kernels:  $K(\mathbf{x}) = \prod_{i=1}^d K_0(x_i)$
  - ellipsoidal kernel
    - multivariate normal density:  $(2\pi)^{-d/2} \exp(-\mathbf{x}\mathbf{x}^\top/2)$
- ⇒ the matrix of bandwidths plays the role of the covariance-variance matrix
- multivariate Epanechnikov:  $\frac{d+2}{2c_d}(1 - \mathbf{x}\mathbf{x}^\top)\mathbb{1}_{[-1,1]}(\mathbf{x}\mathbf{x}^\top)$ , with  $c_d$  the volume of a  $d$ -dimensional unit ball ( $c_1 = 1$ ,  $c_2 = \pi$ ,  $c_3 = 4\pi/3$ )

Degrees of smoothing are controlled by  $h$  and can be set different along the directions, i.e., under a product kernel, the KDE is

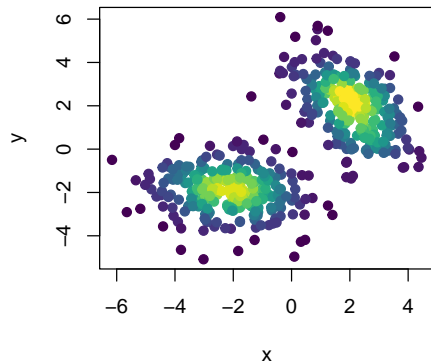
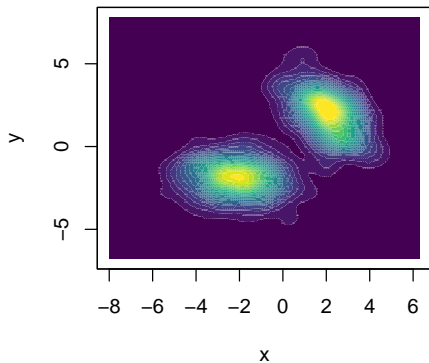
$$\hat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1} K_0\left(\frac{x_1 - X_{i,1}}{h_1}\right) \times \dots \times \frac{1}{h_d} K_0\left(\frac{x_d - X_{i,d}}{h_d}\right)$$

⇒ if margins are standardized (on the same scale), set  $h = h_1 = \dots = h_d$



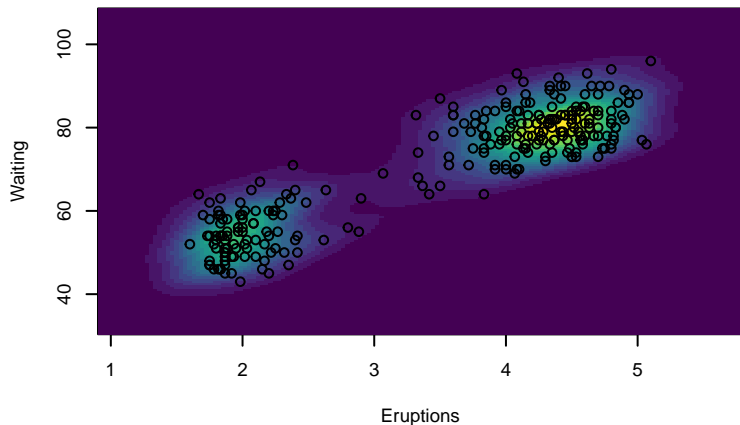
# Multivariate KDE

- Mixture of bivariate normal



# Multivariate KDE

- Faithful dataset



# Curse of dimensionality

- KD estimation is typically restricted to  $d = 2$
- Unless sample size is very large, neighbourhoods will be sparsely populated with data points (in higher dimensions)

For instance,

- If you have  $n$  data points uniformly distributed on the interval  $[0, 1]$ , how many data points are there in the interval  $[0, 0.1]$ ?

Around  $n/10$

- If you have  $n$  data points uniformly distributed on the 10-dimensional unit cube  $[0, 1]^{10}$ , how many are there in the cube  $[0, 0.1]^{10}$ ?

Around  $0.1^{10}n$

⇒ estimation gets harder very quickly as dimension increases

# Curse of dimensionality

Under some smoothing conditions on  $f$ , the best possible MSE rate (the one obtained with optimal choice of bandwidth) is  $O(n^{-4/(d+4)})$ . That is,  $MSE_{h_{opt}} \approx cn^{-4/(d+4)}$  and  $n \approx (c/MSE_{h_{opt}})^{d/4}$

$\Rightarrow$  sample size grows exponentially with dimension

$n^{-4/(d+4)}$	$d = 1$	$d = 2$	$d = 5$
$n = 100$	0.025	0.046	0.129
$n = 1000$	0.004	0.010	0.046
$n = 10000$	$6.3 \times 10^{-4}$	$2.1 \times 10^{-3}$	$1.6 \times 10^{-2}$

Thus, for  $d = 5$ , the rate with  $n = 10000$  is the same than for  $d = 2$  with 10 times less data ...

# Summary - Overall

## Motivation:

- 1 On Week 2, we introduced the histogram as a data exploratory tool and noticed its limitations
- 2 Histogram is a poor estimator of density, because it
  - is never smooth and requires a choice of *origin*
- 3 Today, we introduced naive KDE by generalizing histogram to its *origin-free* version
- 4 Then, we generalized naive KDE by allowing for better kernels
- 5 Now we have a decent nonparametric density estimation tool: KDE
  - in exploratory analysis, histograms often overlaid with KDEs

## Main takeaways:

- 7 Asymptotic properties analyzed using Taylor expansions
  - suggest a way to choose *bandwidth*
  - the bias-variance trade-off made explicit
- 8 Multivariate extension works well in low dimensions

# Assignment 2 and Exercise

Go to [Assignment 2](#) for details.

Go to [Exercise 2](#) for details.