The week ahead

- **Quiz 5:** mean is 81% and average completion time 5min 40sec!
- **Touch-point 1 deliverables due tonight at 11:59pm**
  - Three-min video + one-slide presentation → Piazza thread
- **Touch-point 1, Wed Sep 30th during class time**
  - Everyone should watch the pitch videos from the teams in their own cluster and be prepared to give feedback/ask questions
- **Quiz 6, Friday, Oct 2nd 6am until Oct 3rd 11:59am (noon)**
  - Density estimation
- **Project proposal due Oct 2nd 11:59pm (midnight)**
  - Link to GitHub page + pdf printout of your webpage → Gradescope
- **Assignment 2 due Oct 5th 11:59pm (midnight)**

CS4641B Machine Learning

Lecture 12: Density estimation

Rodrigo Borela ▸ rborelav@gatech.edu

Adopted from slides from Le Song and Mahdi Roozbahani

Outline

- Overview
- Parametric density estimation
- Nonparametric density estimation

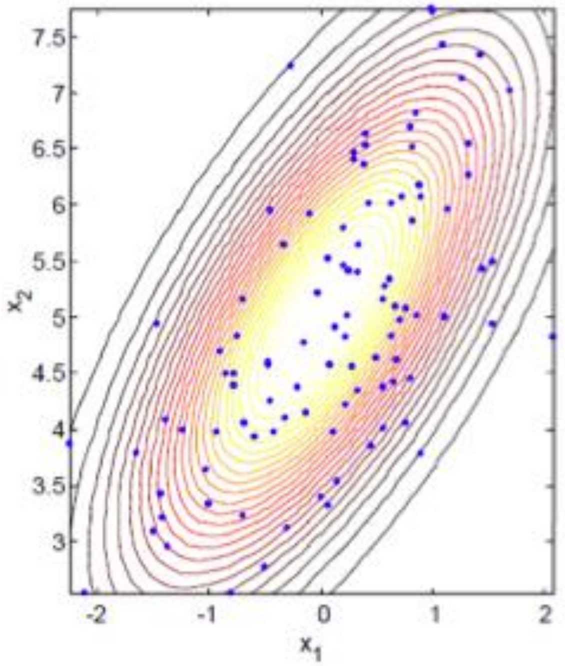
- *Complementary reading: Bishop PRML – Chapter 2, Parametric methods Sections 2.1 through 2.4 and Nonparametric methods Section 2.5 through 2.5.2*

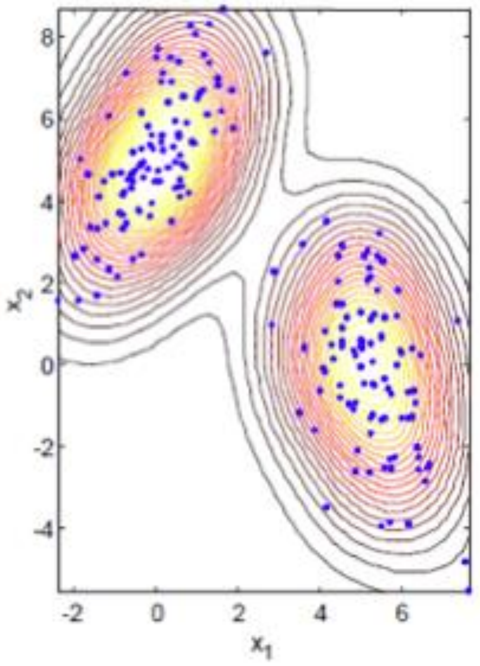
Outline

- Overview
- Parametric density estimation
- Nonparametric density estimation

Why density estimation?

- Learn more about the “shape” of the data cloud



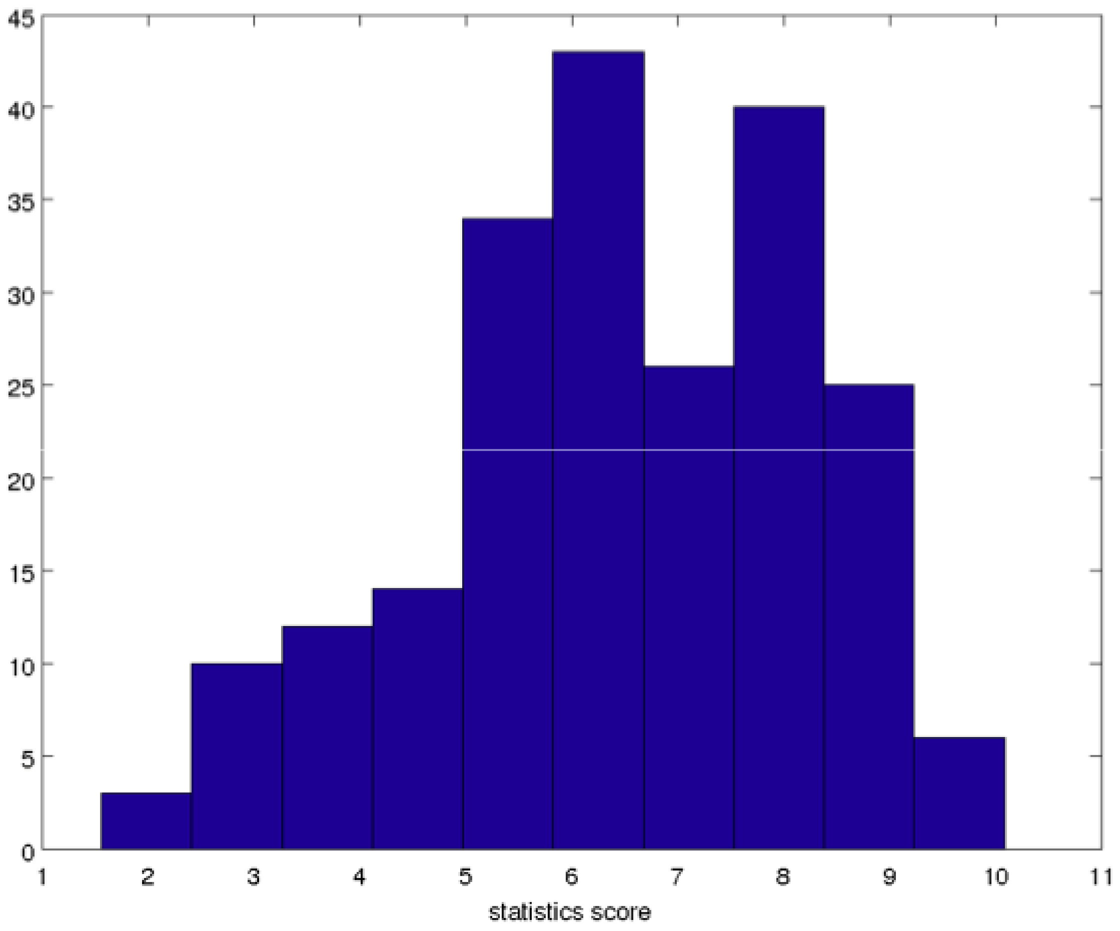


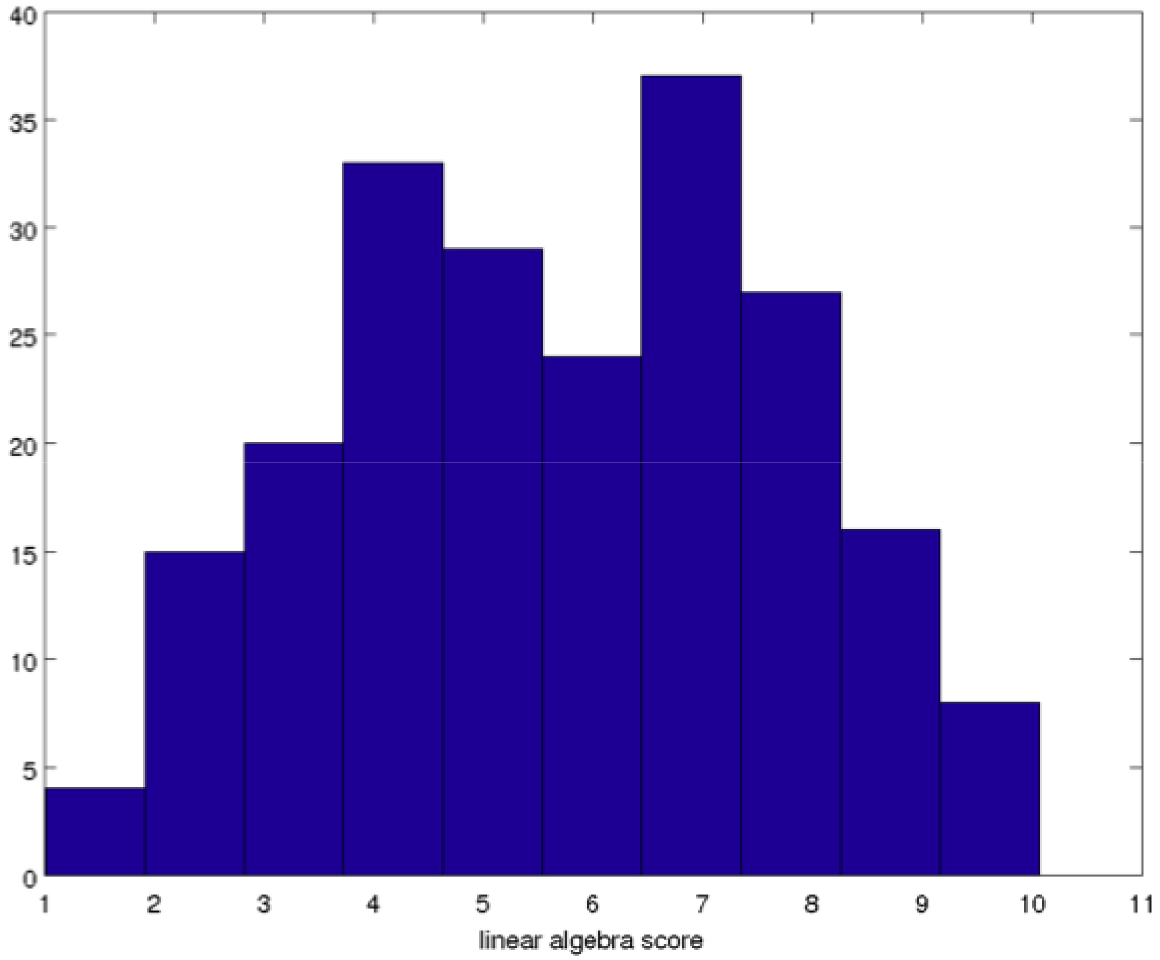
- Access the density of seeing a particular data point
  - Is this a typical data point? (high density value)
  - Is this an abnormal data point/outlier? (low density value)

- Building block for more sophisticated learning algorithms
  - Classification, regression, graphical models
  - A simple recommendation system

Why density estimation?

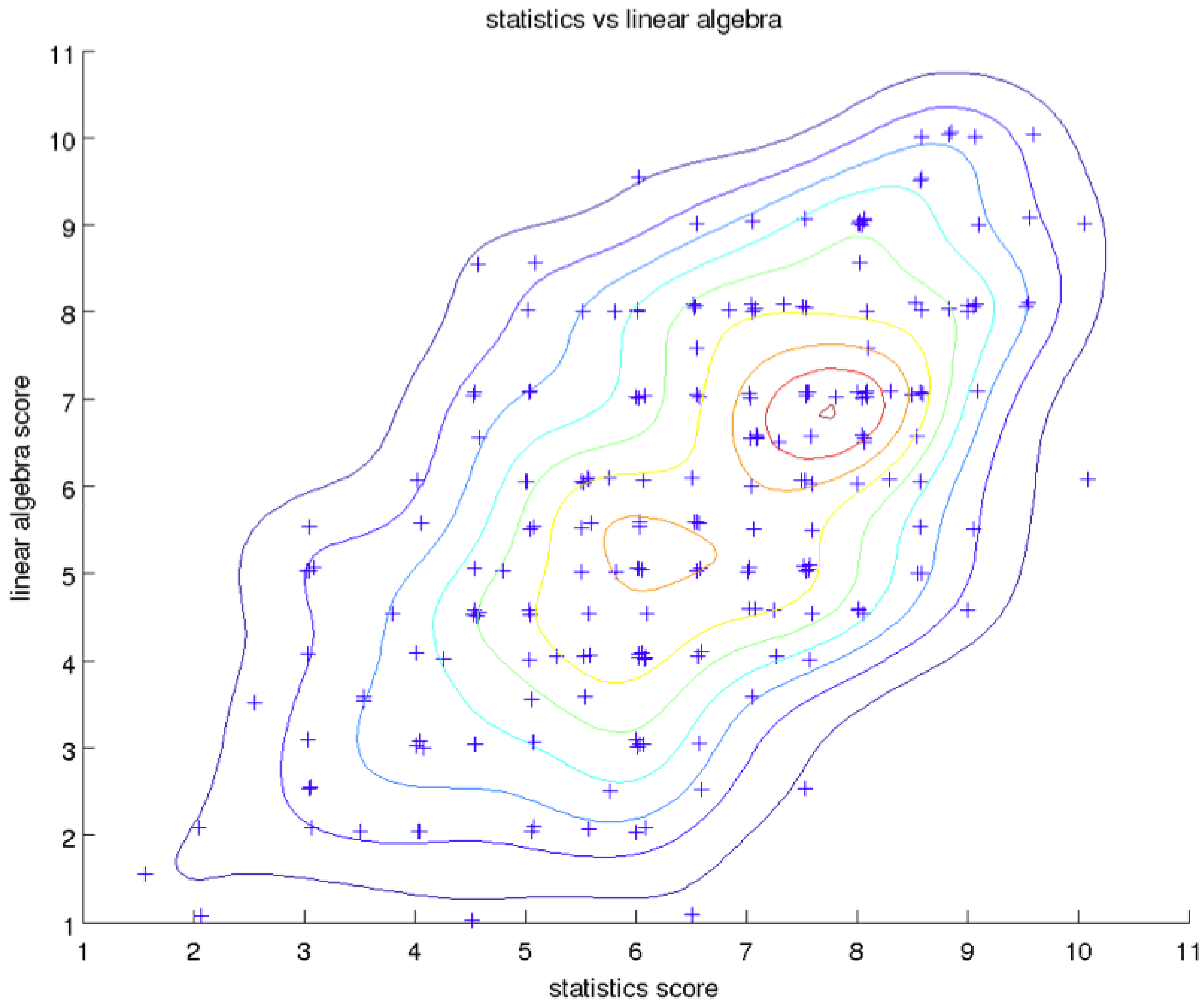
Example: test scores





Histogram is an estimate of the probability distribution of a continuous variable

Example: test scores



Parametric density estimation

- Model which can be described by a fixed number of parameters

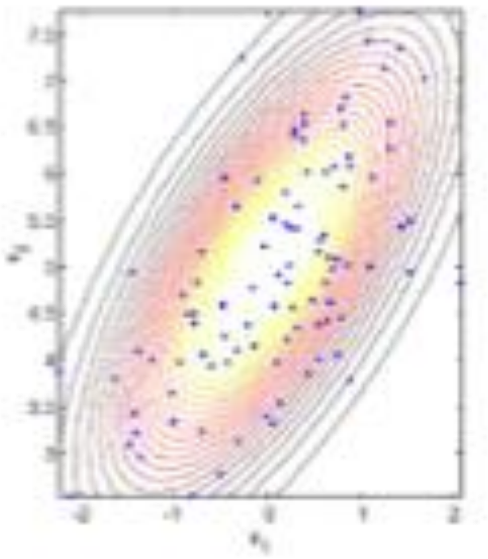
- **Discrete case:** e.g. Bernoulli distribution

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

one parameter θ (probability of possible outcome), $\theta \in [0,1]$, which generates a family of models $\mathcal{F} = \{p(x|\theta) |\ \theta \in [0,1]\}$

- **Continuous case:** e.g. Gaussian distribution in $\mathbb{R}^D$

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

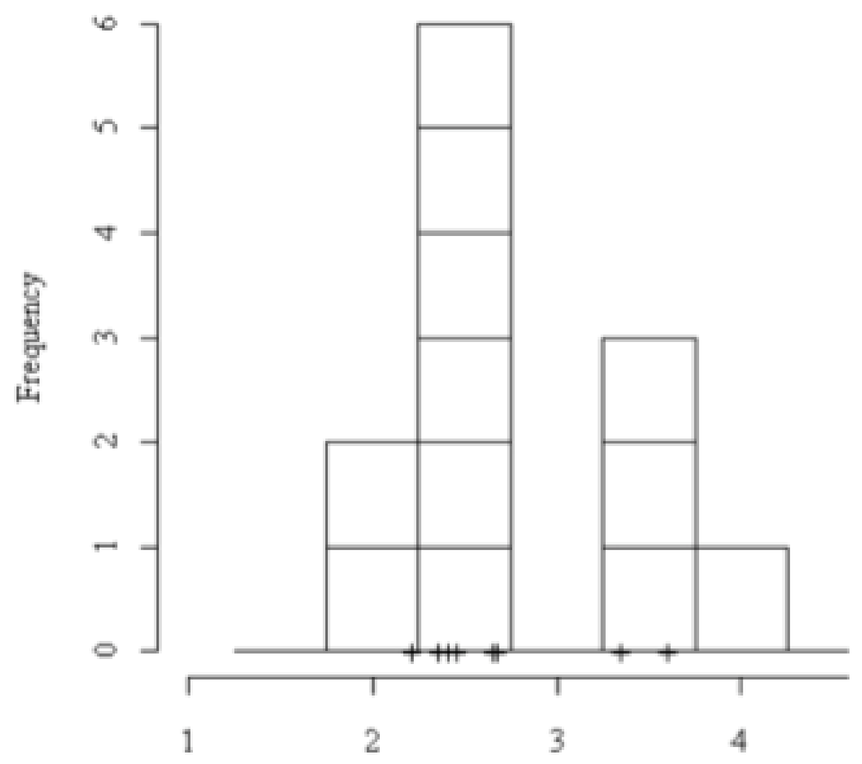


two sets of parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, which again generate a family of models $\mathcal{F} = \left\{p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) |\ \boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}, \{0,1\}\right\}$

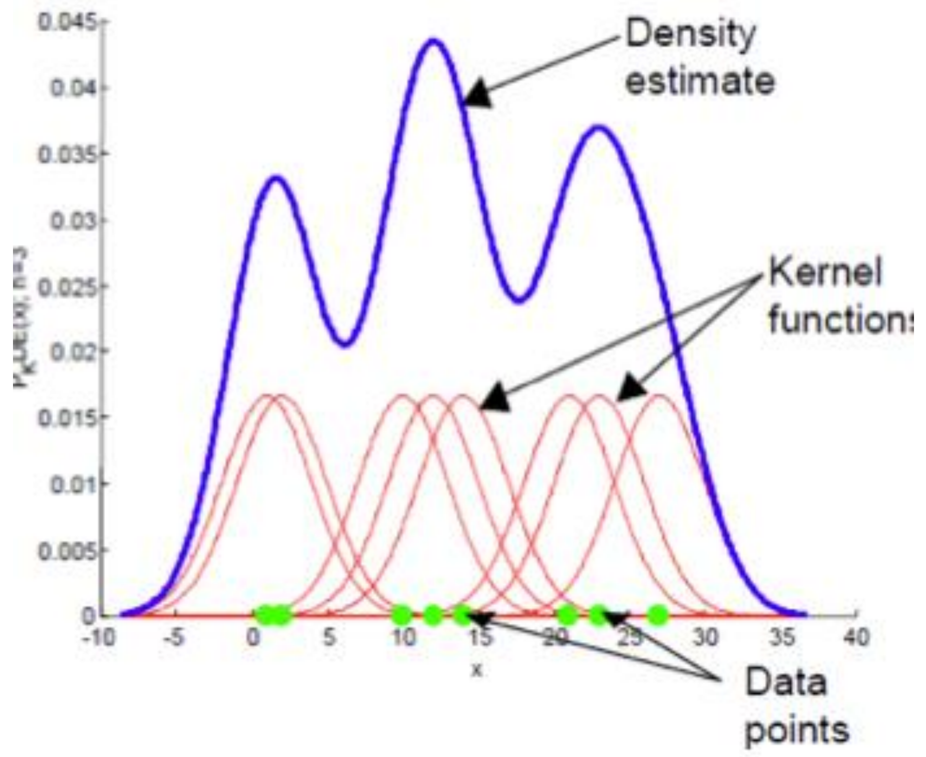
Nonparametric density estimation

- What are nonparametric models?
  - “Nonparametric” does not mean there are no parameters
  - Can not be described by a fixed number of parameters
  - One can think there are many parameters

- Examples: histogram and kernel density estimator

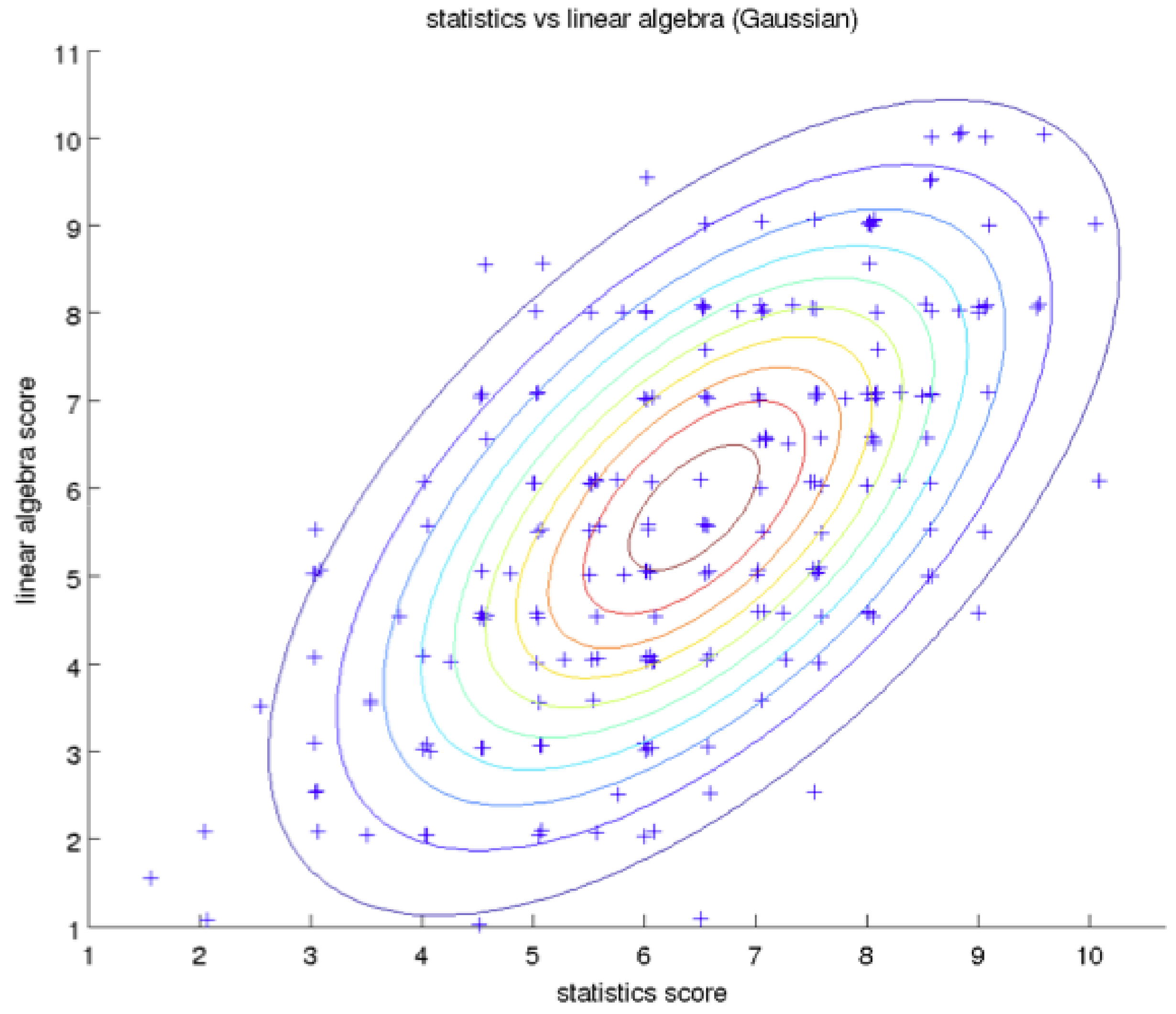


Histogram

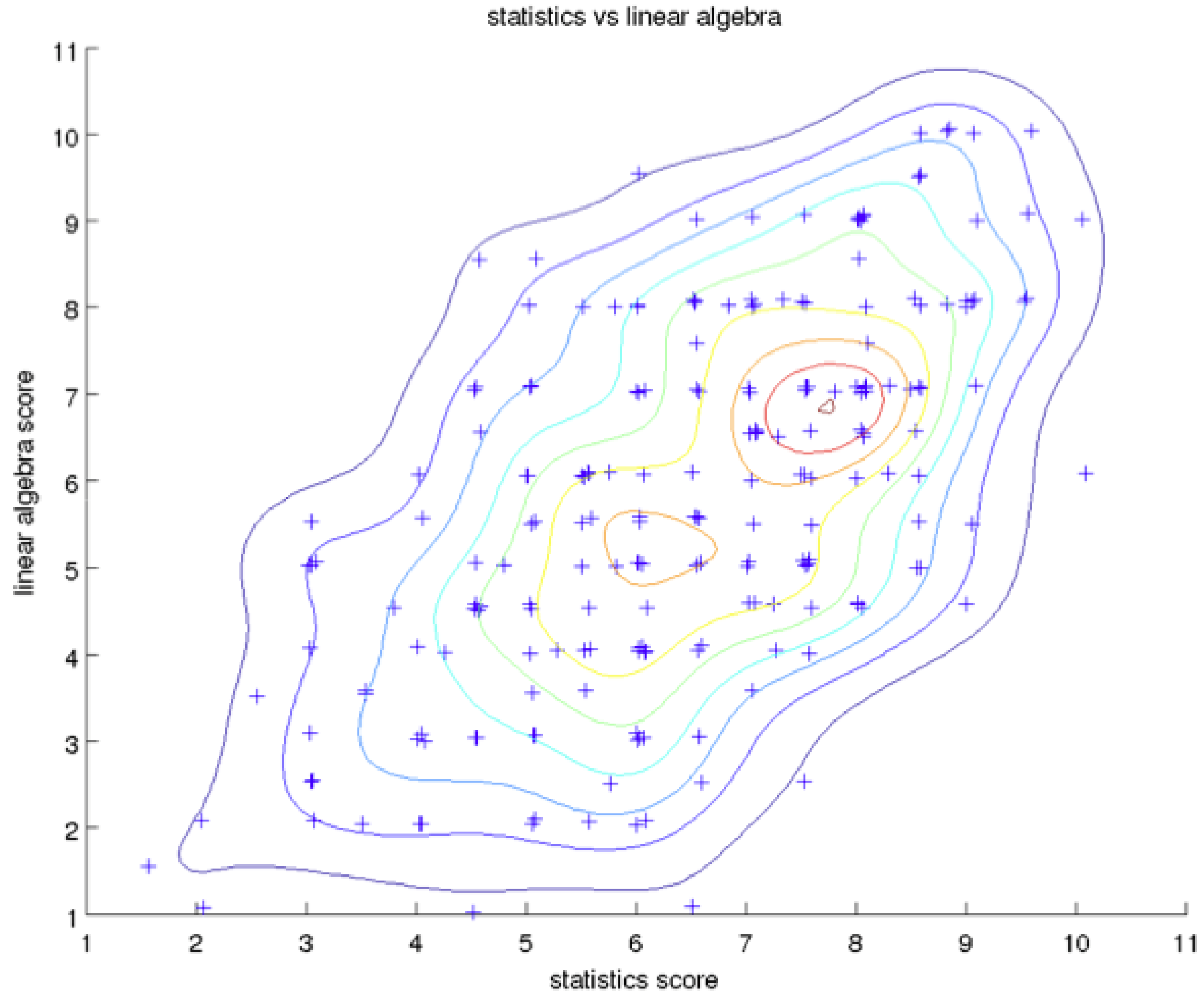


Kernel density estimator

Parametric vs. nonparametric density estimation

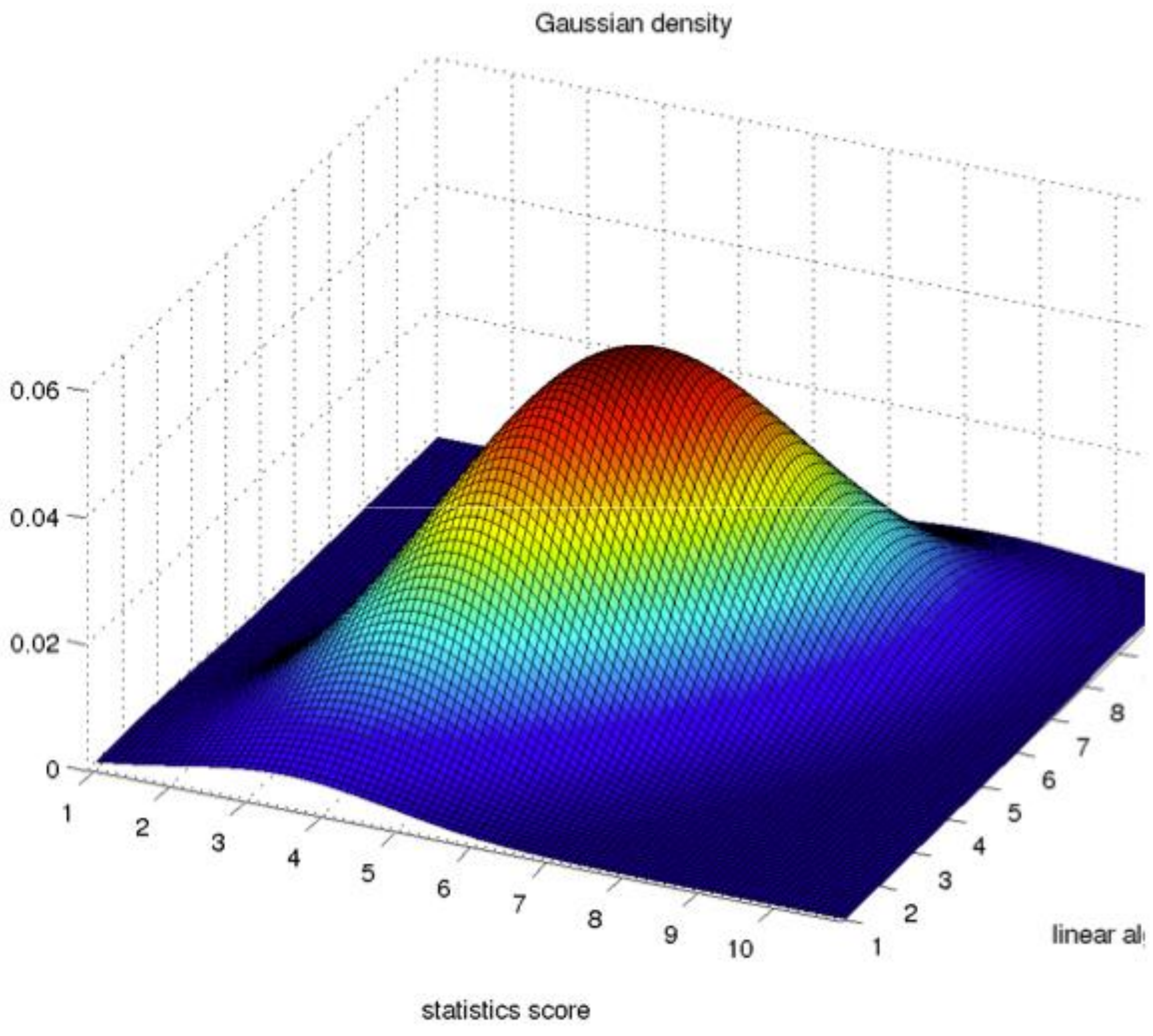


Parametric

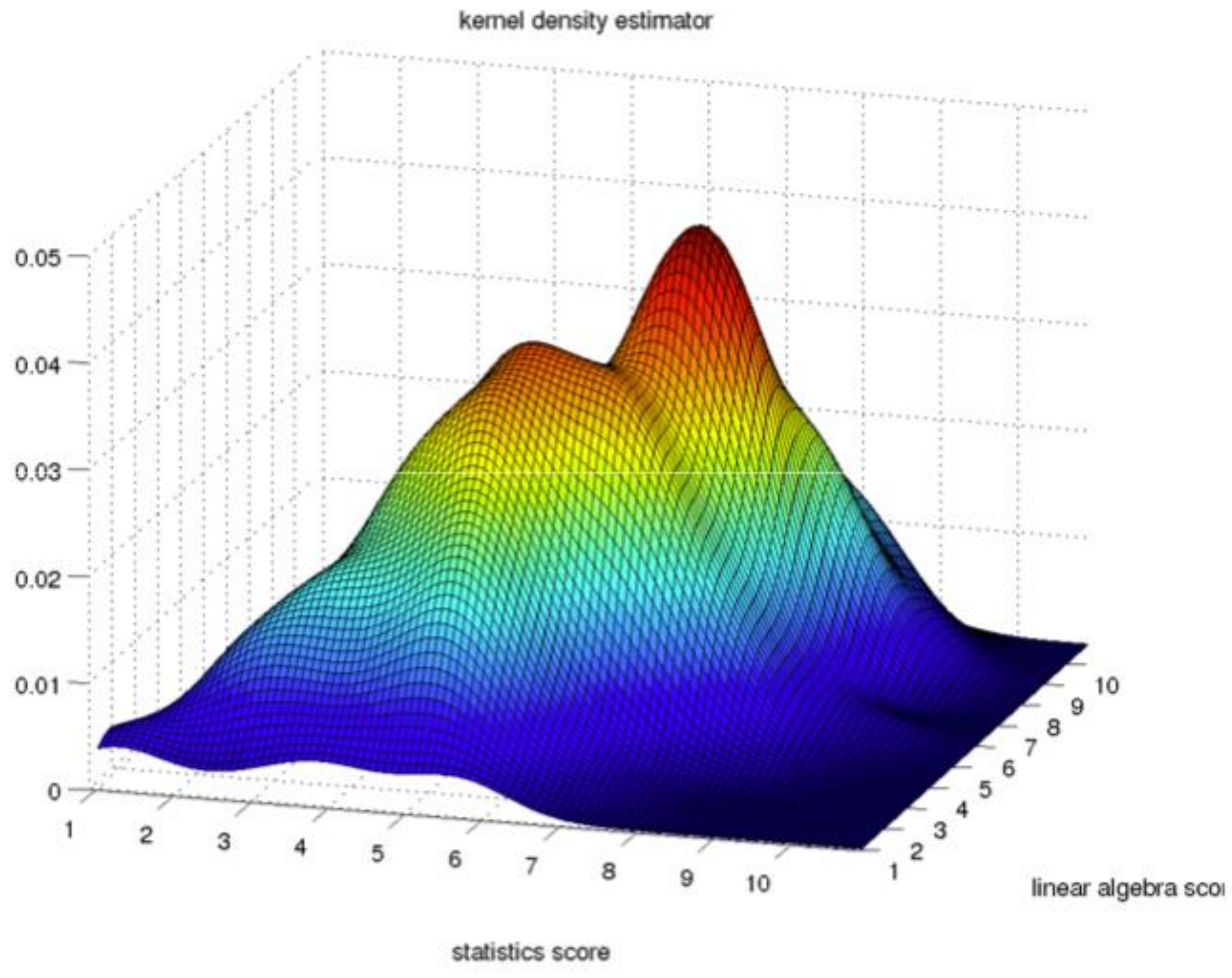


Nonparametric

Parametric vs. nonparametric density estimation



Parametric



Nonparametric

Outline

- Overview
- Parametric density estimation
- Nonparametric density estimation

Estimating parametric models

- A very popular estimator is the maximum likelihood estimator (MLE), which is simple and has good statistical properties

- Assume that we have N data points $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ drawn independently and identically (iid) from some distribution $p^*(\mathbf{x})$
- Want to fit the data with a model $p(\mathbf{x}|\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$, we want to maximize the log-likelihood of our dataset:

$$L(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{x}_1, \dots \mathbf{x}_n|\boldsymbol{\theta}) \overset{iid}{\Rightarrow} p(\mathbf{x}_1|\boldsymbol{\theta})p(\mathbf{x}_2|\boldsymbol{\theta}) \dots p(\mathbf{x}_N|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}}(\log p(\mathbf{X}|\boldsymbol{\theta})) = \arg\max_{\boldsymbol{\theta}}\left(\log \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta})\right) = \arg\max_{\boldsymbol{\theta}}\left(\sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta})\right)$$

MLE for a biased coin: example

- Estimate the probability θ of landing in heads using a biased coin
- Given a sequence of N independently and identically distributed (iid) flips
  - e.g. $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} = \{1,0,1,\ldots,0\}, x \in \{0,1\}$

- Model: $p(x|\theta) = \theta^x (1-\theta)^{1-x}$

$$p(x|\theta) = \begin{cases} 1-\theta, \text{for } x = 0 \\ \theta, \quad\quad\; \text{for } x = 1 \end{cases}$$

- Likelihood of a single observation x_n?

$$L(\theta|x_n) = p(x_n|\theta) = \theta^{x_n}(1-\theta)^{1-x_n}$$

MLE for a biased coin

- Objective function, log-likelihood

$$l(\theta|\mathbf{X}) = \log L(\theta|\mathbf{X}) = \log \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \log(\theta^{N_H}(1-\theta)^{N_T})$$

$$= N_H \times \log\theta + N_T \times \log(1-\theta)$$

$N_H =$ number of heads, $N_T =$ number of tails

- Maximize $l(\theta|\mathbf{X})$ w.r.t. $\theta \rightarrow$ take derivative w.r.t. θ and set it to zero

$$\frac{\partial l(\theta|\mathbf{X})}{\partial\theta} = \frac{N_H}{\theta} - \frac{N-N_H}{1-\theta} = 0 \rightarrow \theta_{MLE} = \frac{N_H}{N}$$

- Example: $N_H = 78$, $N_H = 22 \rightarrow \theta = 0.78$

Outline

- Overview
- Parametric density estimation
- Nonparametric density estimation

One-dimensional histogram

- One of the simplest nonparametric density estimator
- Given N iid samples $X = \{x_1, x_2, \dots, x_N\}, x_n \in [\min x, \max x)$
- Split the parameter space into M bins:

$$bin\ width = \Delta = \frac{(\max x - \min x)}{M}$$

$$bin_1 = [\min x, \min x + \Delta), \dots, bin_M = [\min x + (M-1)\Delta, \max x)$$

- Count the number of points x_n that belong in each bin_i = n_i
- For a new test point x

$$p_i = \frac{n_i}{N\Delta_i} = \frac{number\ of\ points\ in\ bin_i}{total\ number\ of\ data\ points\ \times bin_i\ width}$$

One-dimensional histogram

- The probability mass function is given by:

$$P_i = \frac{n_i}{N} = \frac{number\ of\ points\ in\ bin_i}{total\ number\ of\ data\ points}$$

- We know that the probability mass function is given by:

$$P = \int_{\mathcal{R}} p(x)\, dx$$

- Assuming that the probability is evenly distributed inside each bin region:

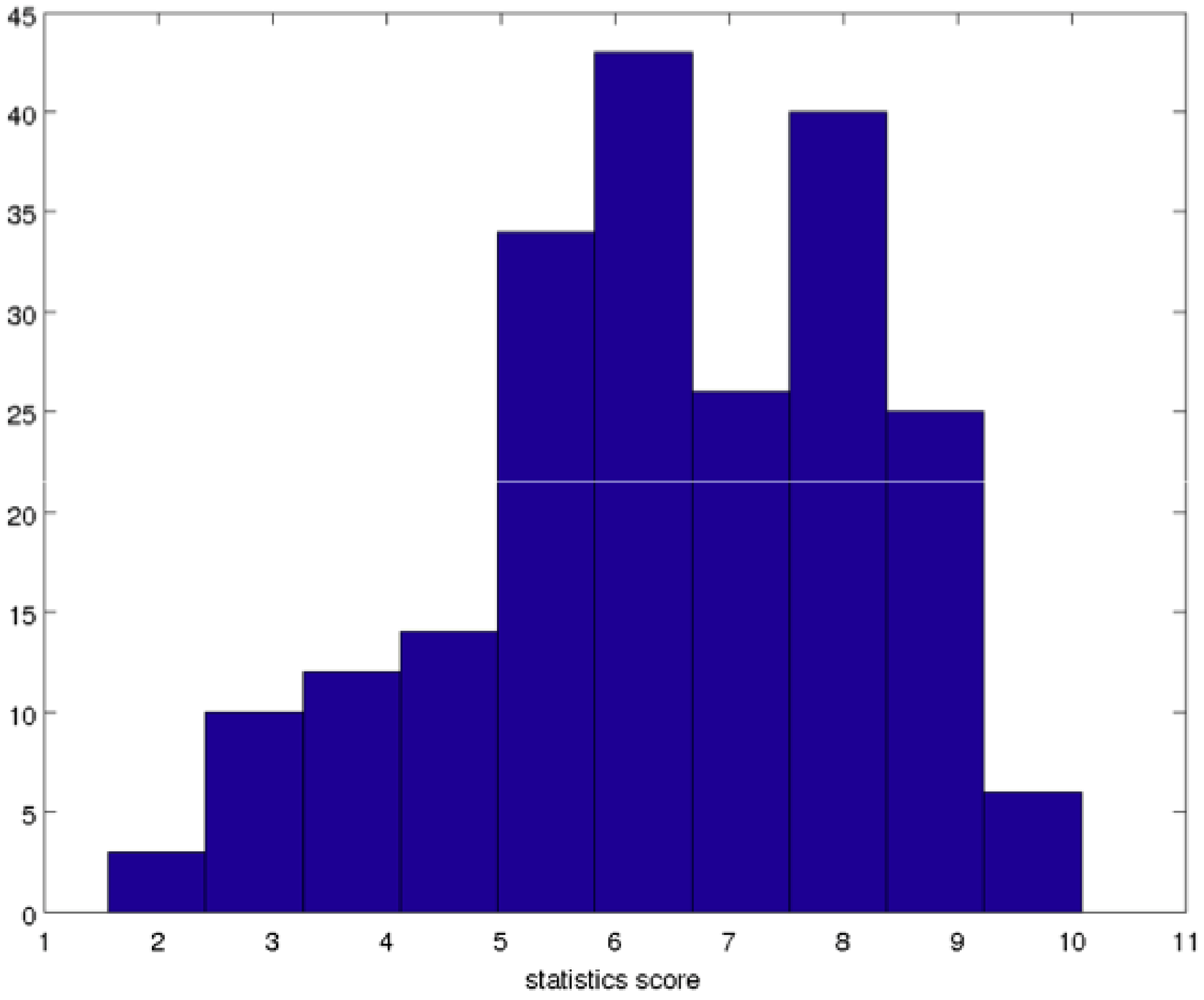
$$P_i = \int_{\mathcal{R}_i} p_i\, dx \rightarrow P_i = p_i \times \Delta_i$$

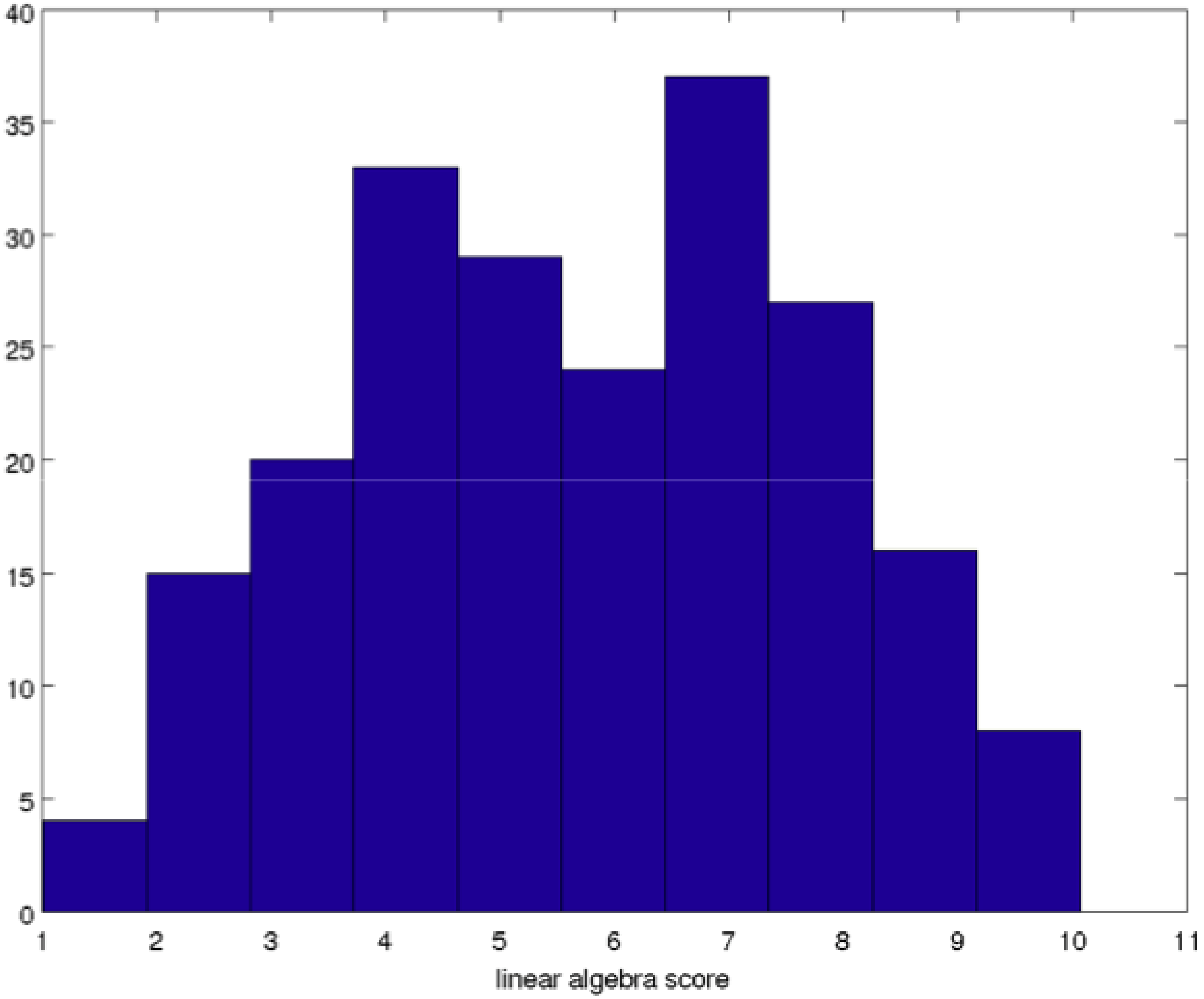
- Then, the probability density function is given by

$$p_i = \frac{n_i}{N\Delta_i} = \frac{number\ of\ points\ in\ bin_i}{total\ number\ of\ data\ points\ \times bin_i\ width}$$

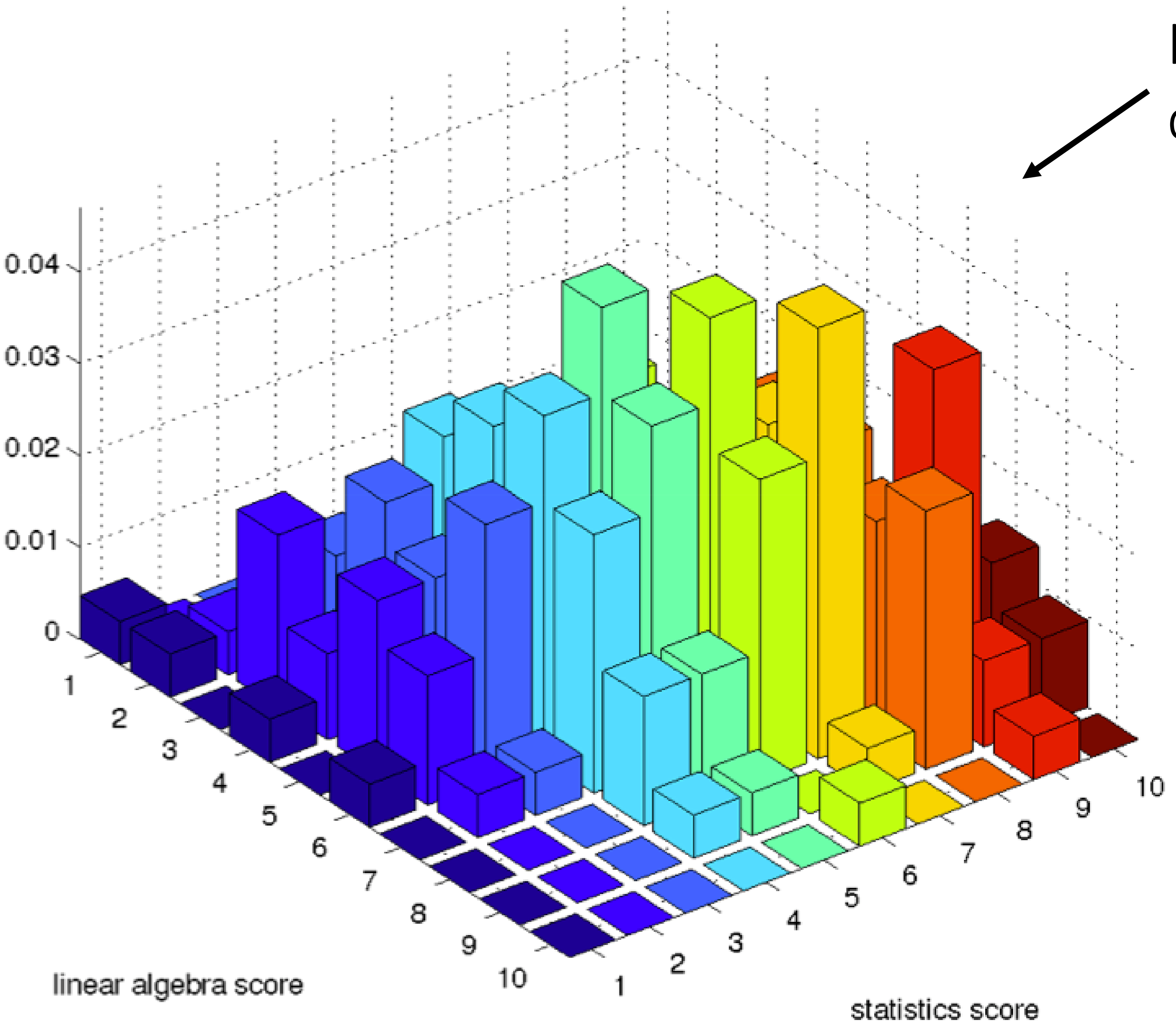
- Which satisfies $p(x) \geq 0$, $\int p(x)dx = 1$

Example: histogram prob. mass function



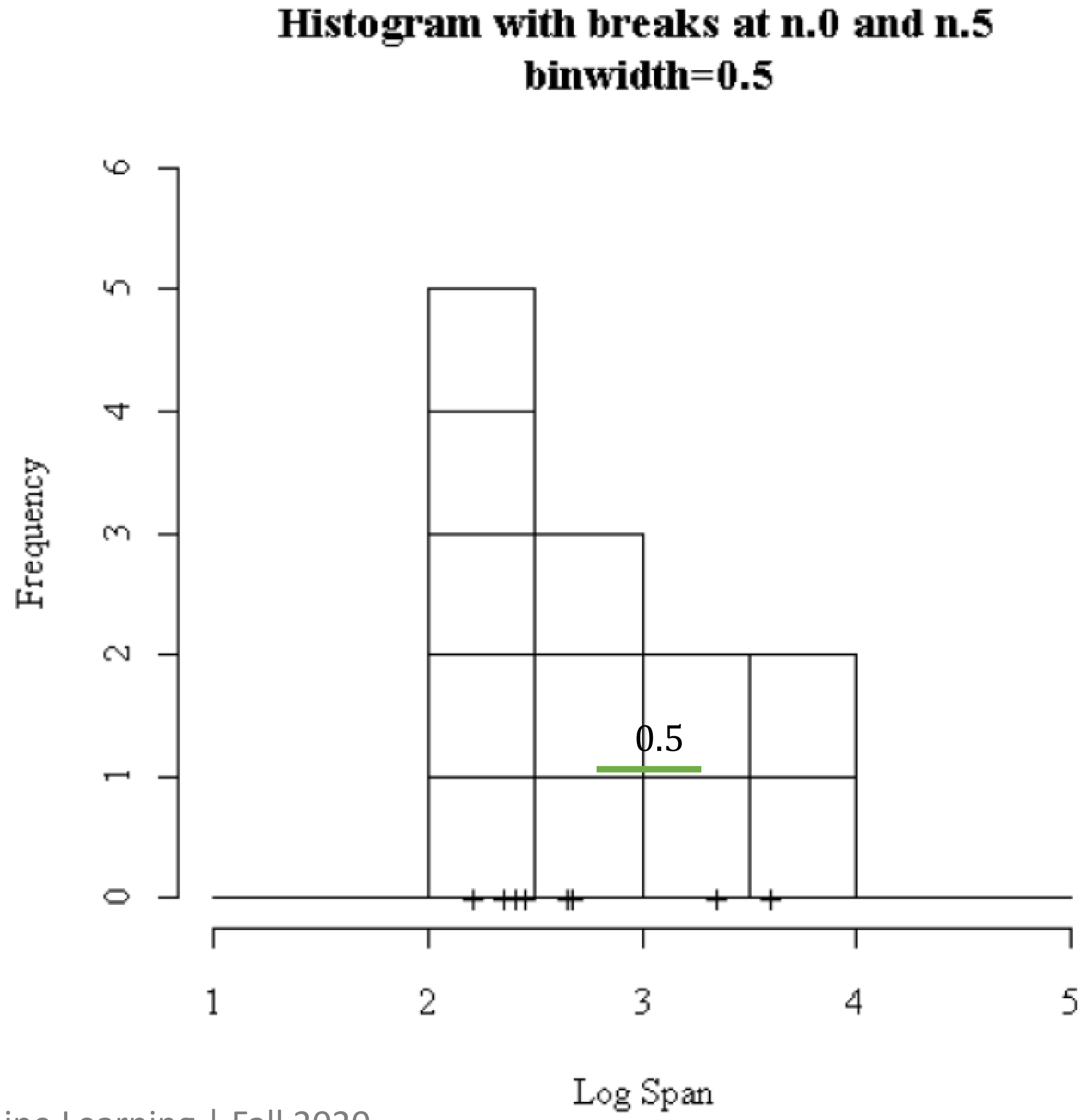


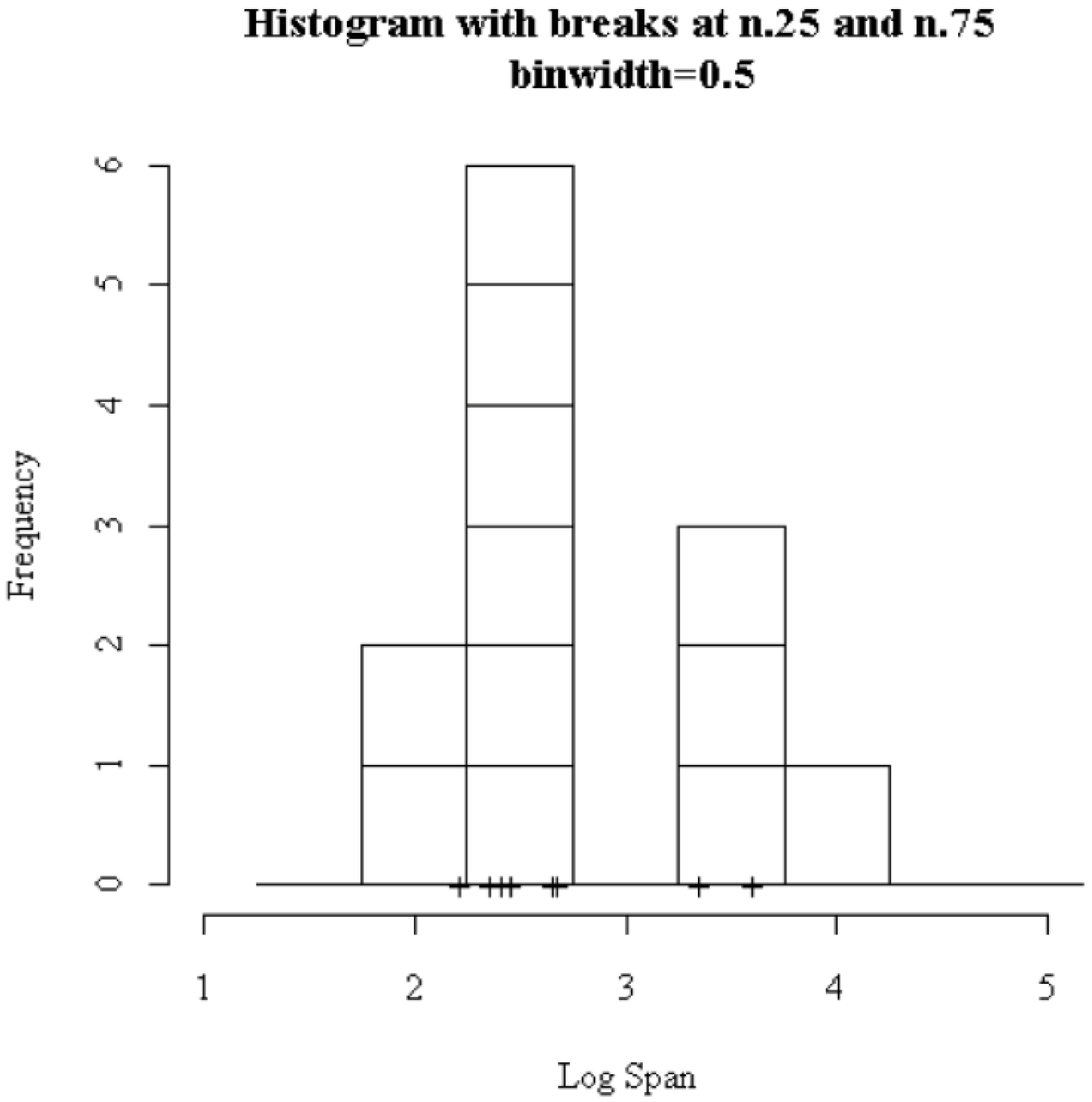
Higher-dimensional histogram



Horrible visualization, don't ever use it!

Histogram results depend on where you place the bins





Histogram results depend on the bin width

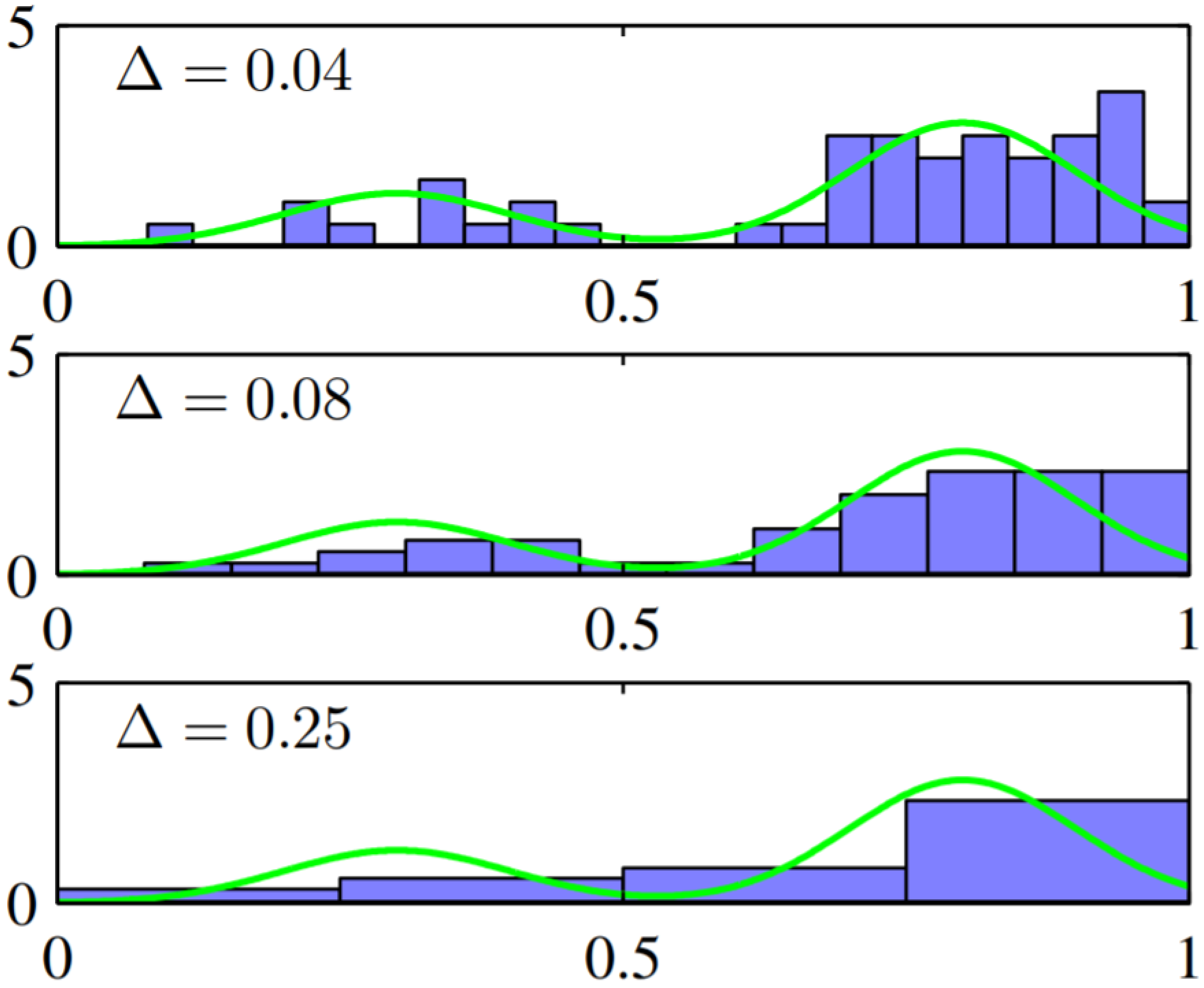


Image credit: Bishop (PRML), 2006

Limitations of histogram

- Scaling with dimensionality > curse of dimensionality
  - For a dataset, where each point is a D-dimensional vector, splitting each feature space in M bins, will lead to a total of M^D bins
- Discontinuities that are not associated with how the data is generated

How is it useful then?

- Visualization
- Provides us with the following intuitions:
  - Estimating the probability density at a particular location should consider the data points within a region
  - We should be careful about how we smooth the space (should not be too small neither too large)

Kernel density estimation

- Kernel function for a hyper-cube of size $\mathbf{u}$

$$k(\mathbf{u}) = \begin{cases} 1, |u_d| \le \frac{1}{2}, d = 1, \dots, D \\ 0, \qquad\qquad otherwise \end{cases}$$

- Total number of data points lying inside the cube centered on $\mathbf{x}_n$

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

- Estimated density at $\mathbf{x}$

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

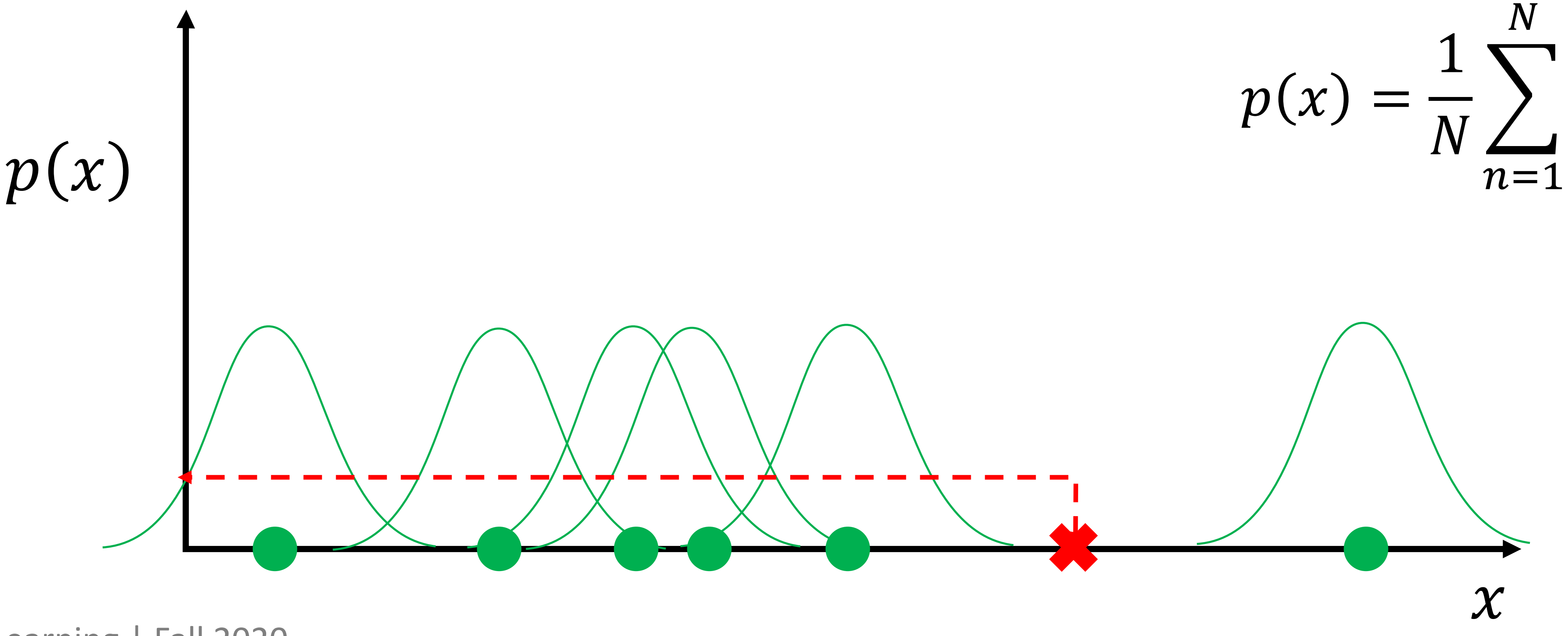
- Still suffering from discontinuities → need a smoother kernel

Kernel density estimation

- Gaussian smoothing kernel

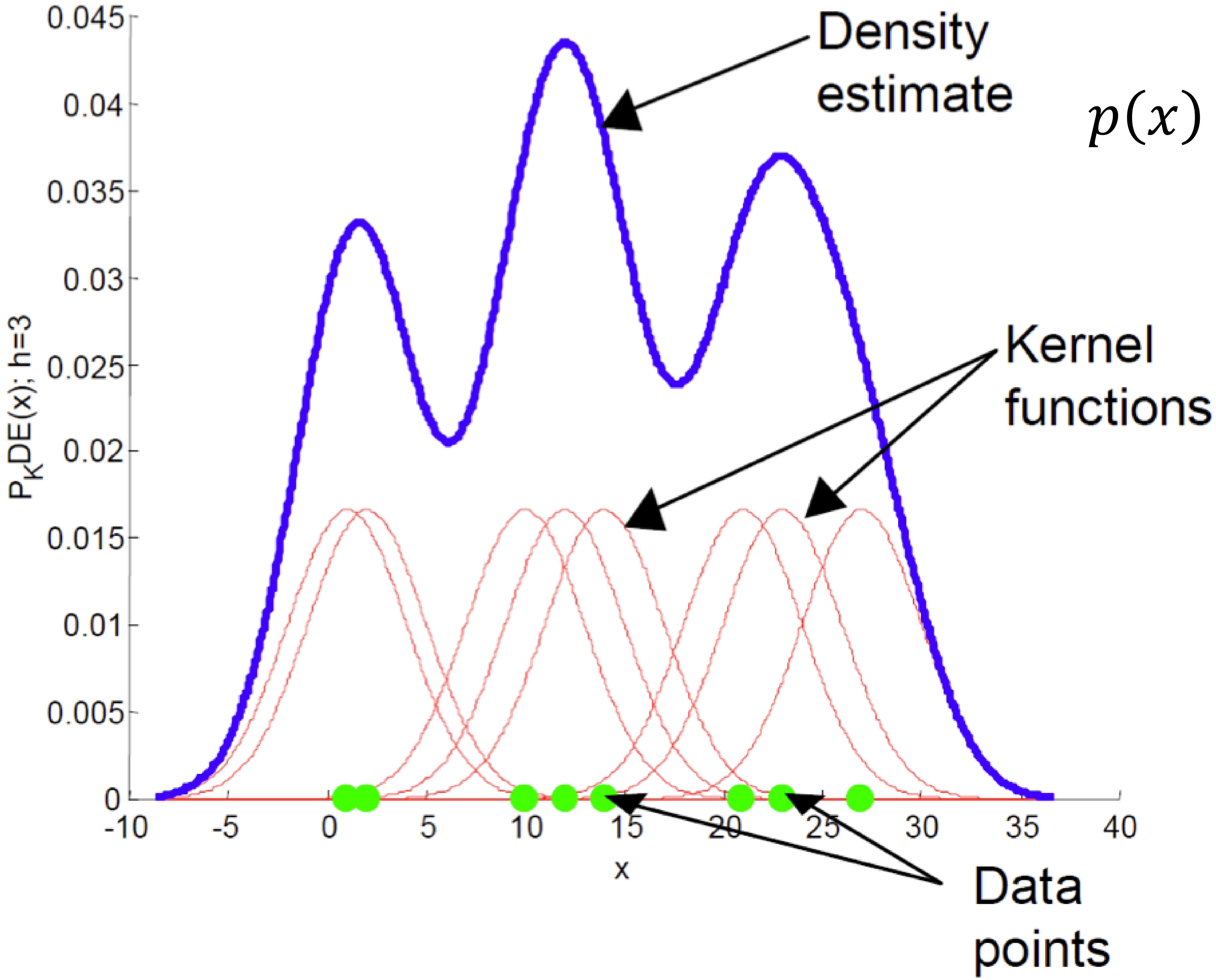
$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{\frac{D}{2}}}\exp\left\{\frac{\|\mathbf{x}-\mathbf{x}_n\|_2^2}{2h^2}\right\}$$

- **What does this mean?** Placing the Gaussian over each data point and summing up their contributions over the whole data set



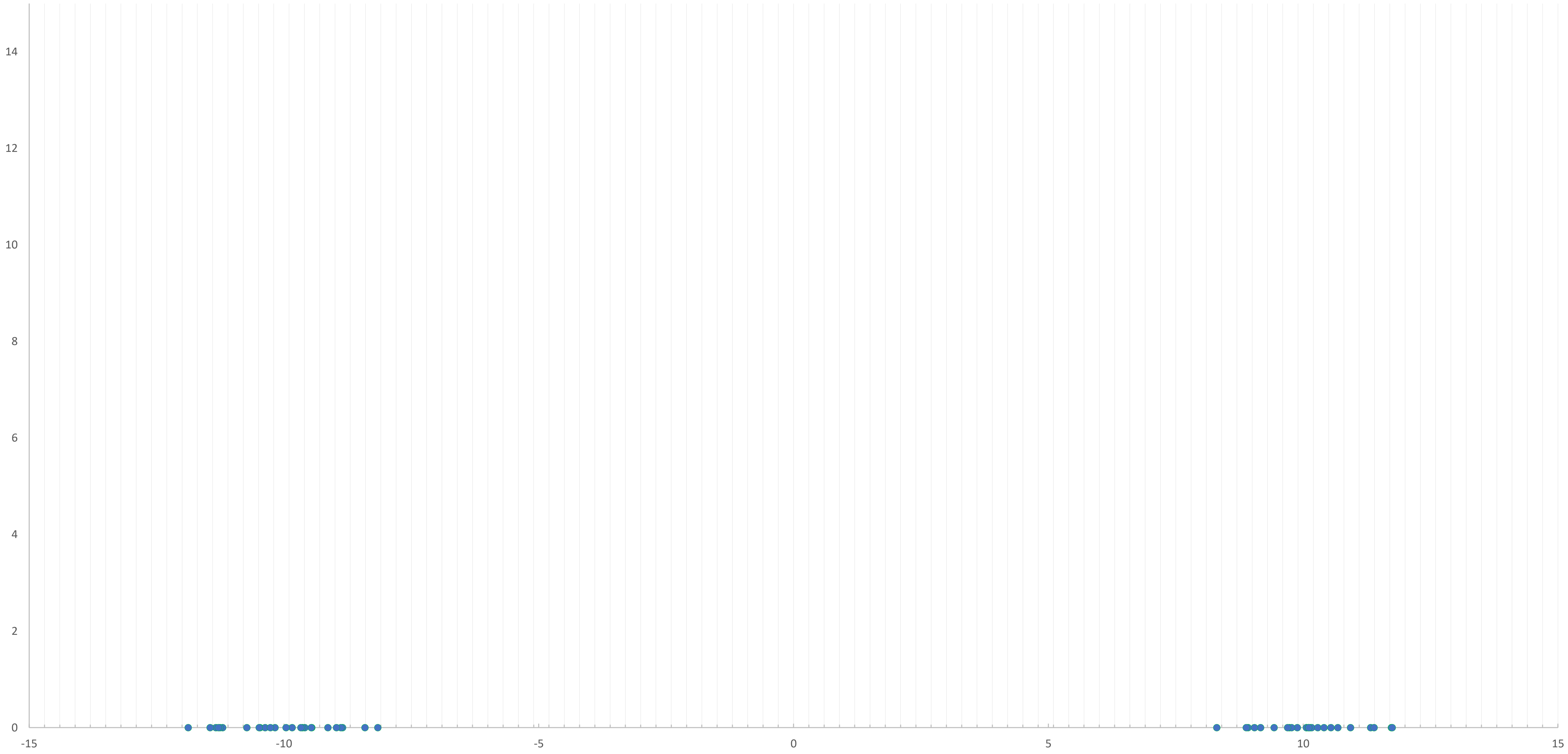
$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h\sqrt{(2\pi)}}\exp\left\{\frac{(x-x_n)^2}{2h^2}\right\}$$

Kernel density estimation: example



$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h\sqrt{(2\pi)}}\exp\left\{\frac{(x-x_n)^2}{2h^2}\right\}$$

Visual example with Gaussian kernel



Kernel Density Estimation

- We can choose any other kernel as long as it satisfies the following conditions:

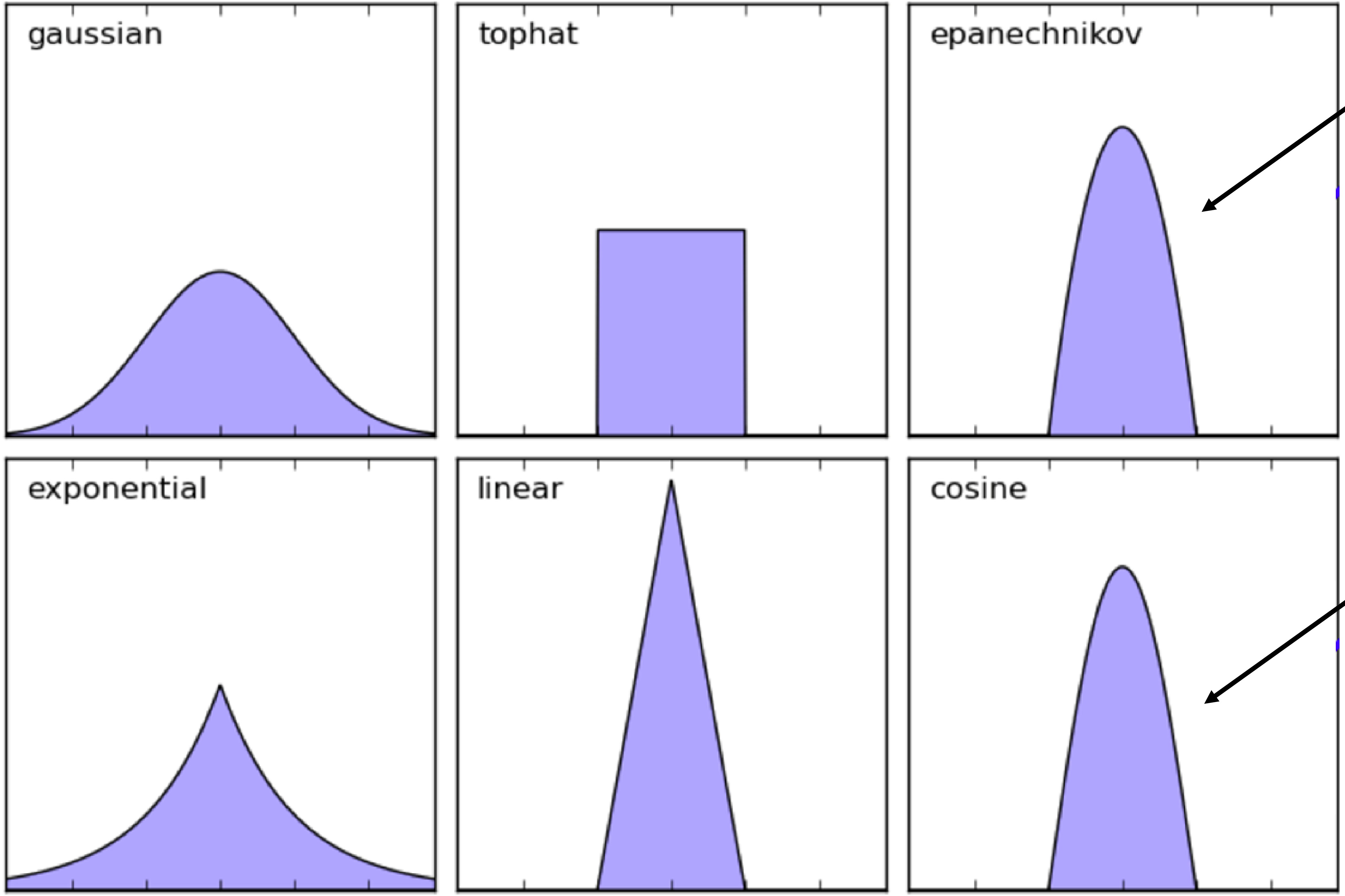
$$k(\mathbf{u}) \geq 0$$

$$\int k(\mathbf{u}) d\mathbf{u} = 1$$

$$k(-\mathbf{u}) = k(\mathbf{u})$$

- What about the training? Well, there isn't one. We have to store the entire dataset and compute the probability of x → large computational cost

Smoothing kernel functions (1D)



$$k(u) = \frac{3}{4}(1 - u^2)$$

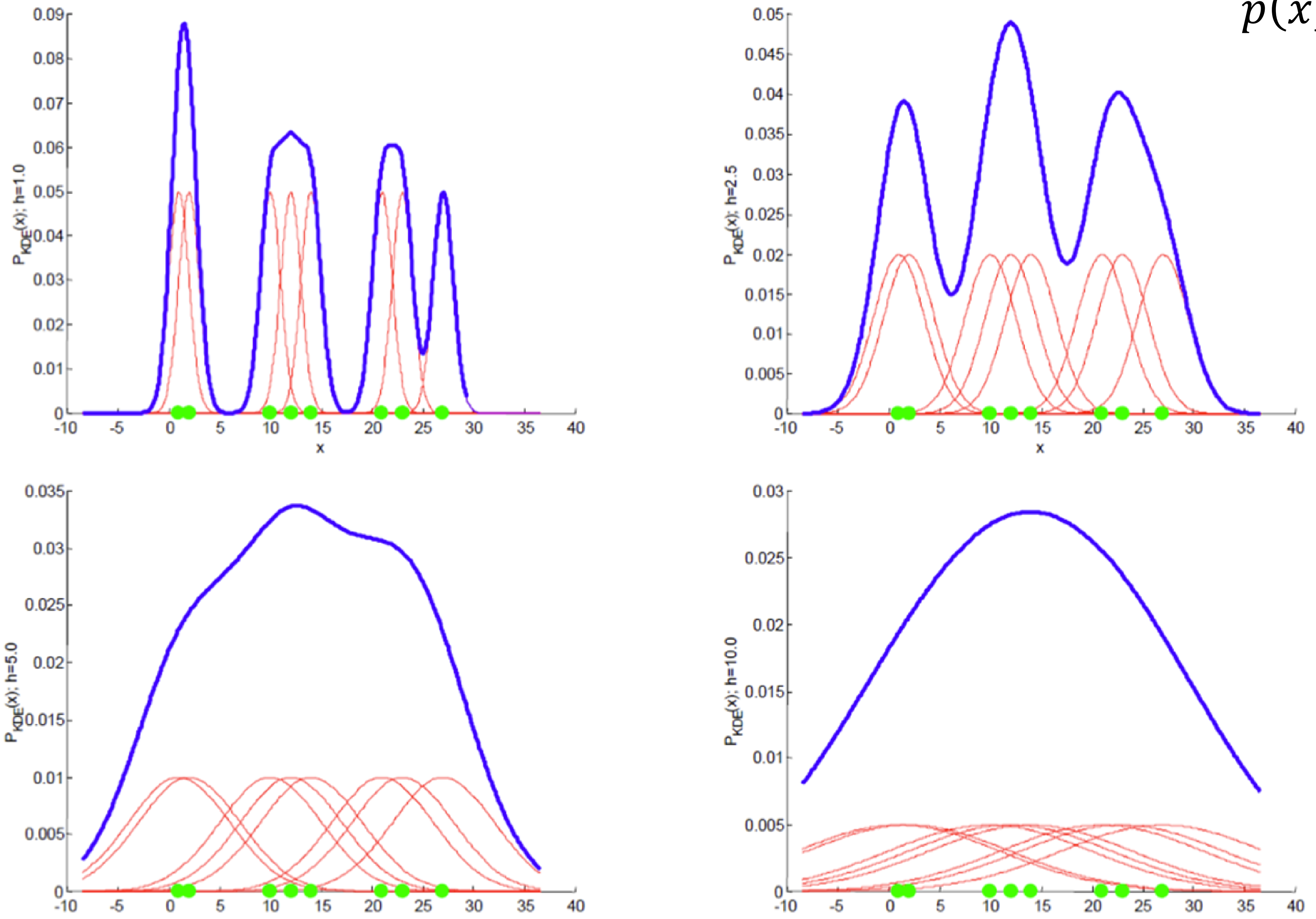
Support: $|u| \leq 1$

$$k(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$$

Support: $|u| \leq 1$

Effect of the Kernel Bandwidth

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h\sqrt{(2\pi)}}\exp\left\{\frac{(x-x_n)^2}{2h^2}\right\}$$



Choosing the kernel bandwidth

- Silverman's rule of thumb: if using the Gaussian kernel, a good choice for h is:

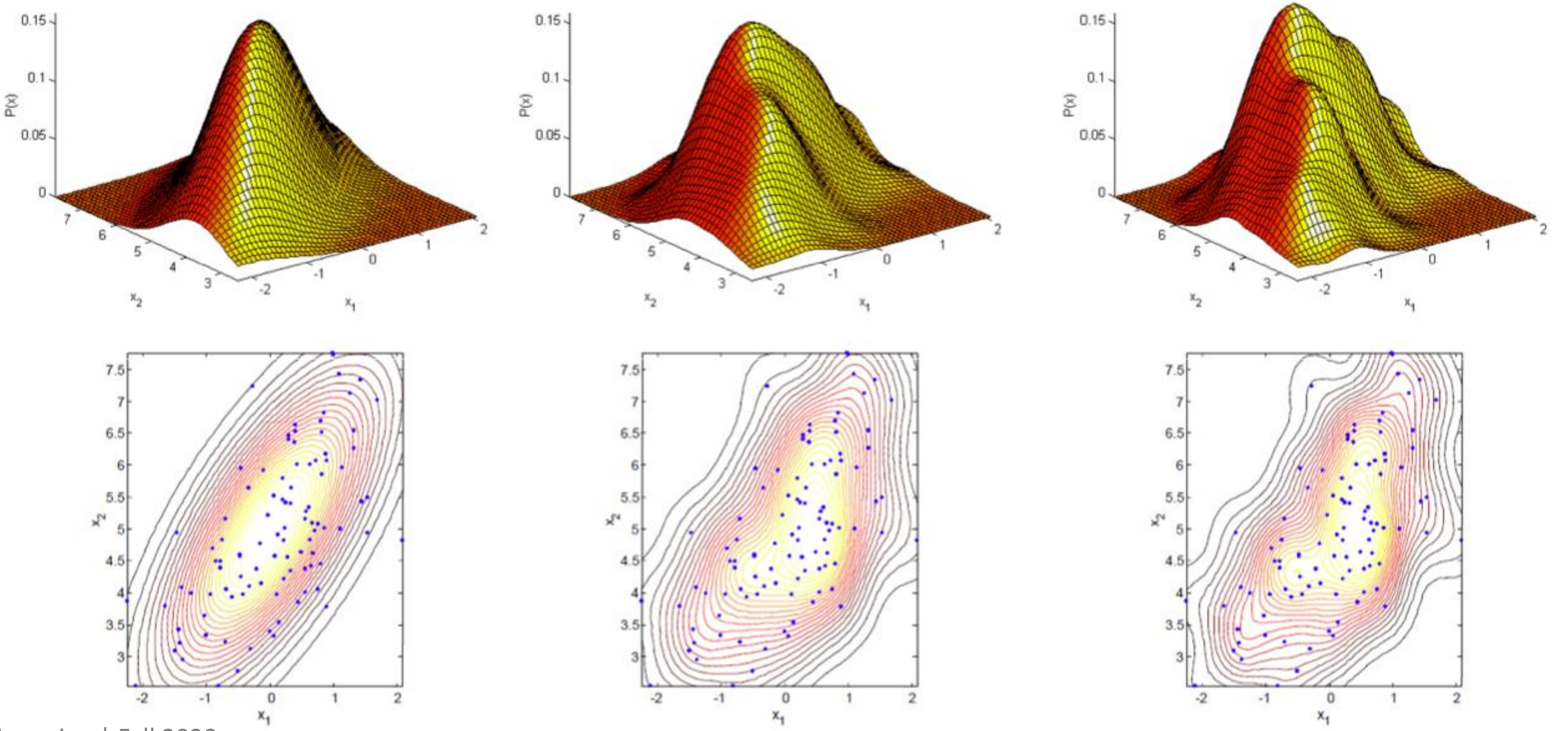
$$h = \left(\frac{4}{3N}\hat{\sigma}^5\right)^{\frac{1}{5}} = 1.06\hat{\sigma}N^{-\frac{1}{5}}$$

Where $\hat{\sigma}$ is the standard deviation and N is the number of datapoints

- Better (more computationally intensive approach)
  - Randomly split the data into two sets
  - Obtain a kernel density estimate for the first
  - Measure the likelihood of the second set
  - Repeat over many random splits and average

Two-dimensional examples

- From left to right: the true distribution from which 100 data points were sampled, the estimate using the Silverman's rule and using a modification with the [parameter A]



Parametric vs nonparametric