Lecture 11
Density Estimation

Mahdi Roozbahani
Georgia Tech

Adapted from slides from Le Song.

Outline

- Overview
- Parametric Density Estimation
- Nonparametric Density Estimation

Continuous variable
Continuous probability distribution
Probability density function
Density value
Temperature (real number)
Gaussian Distribution

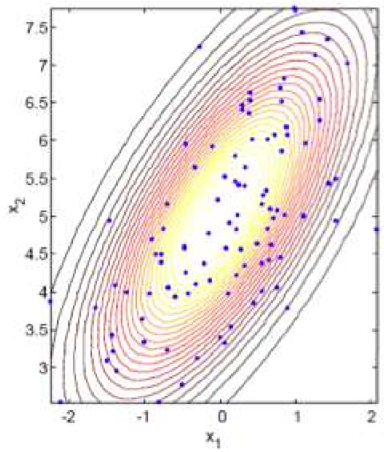
$$\int f_X(x)dx = 1$$

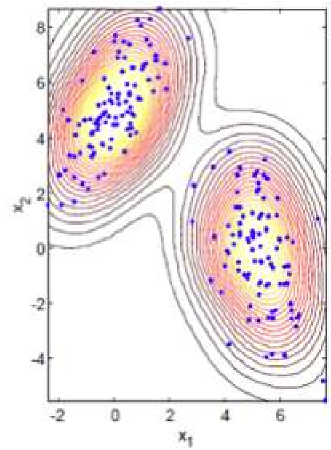
Discrete variable
Discrete probability distribution
Probability mass function
Probability value
Coin flip (integer)
Bernoulli distribution

$$\sum_{x \epsilon A} f_X(x) = 1$$

Why Density Estimation?

- Learn more about the "shape" of the data cloud

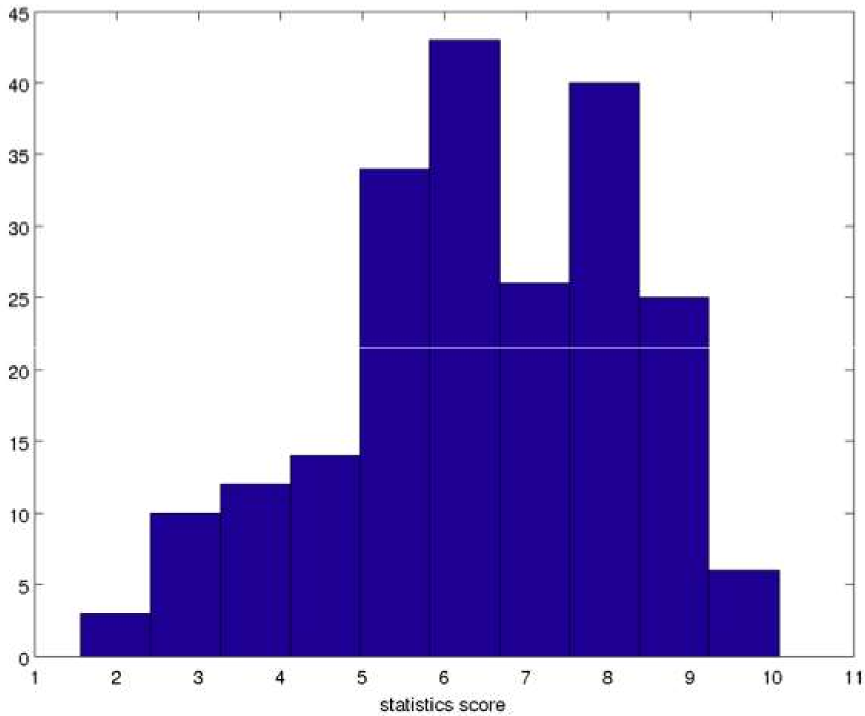


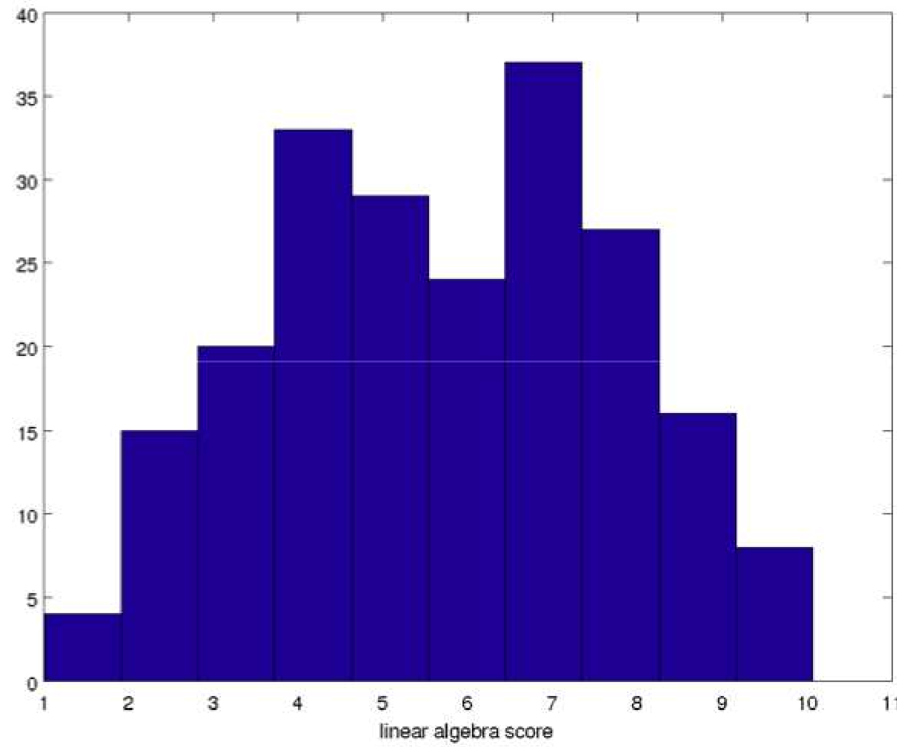


- Access the density of seeing a particular data point
  - Is this a typical data point? (high density value)
  - Is this an abnormal data point / outlier? (low density value)

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

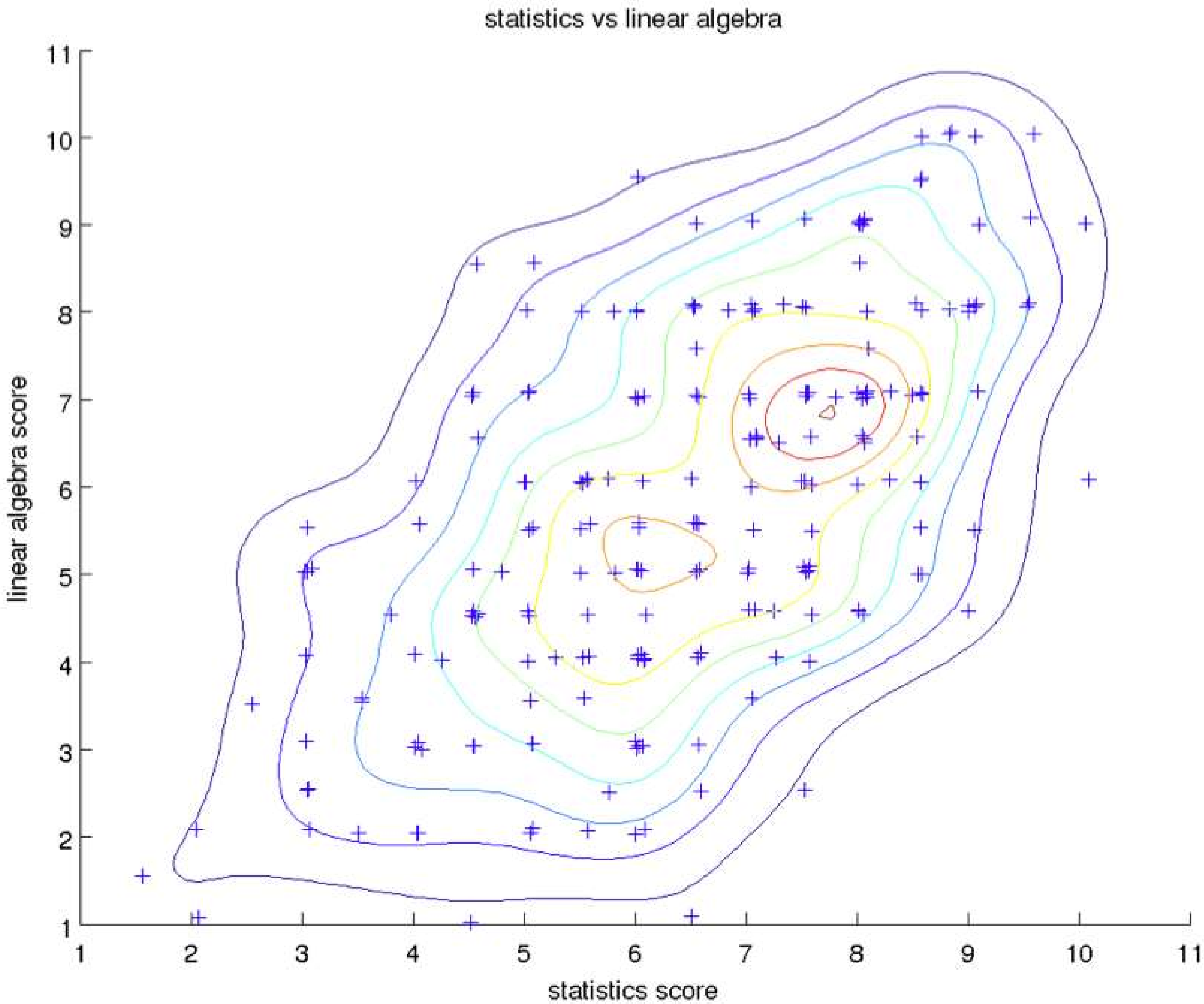
Example: Test Scores





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

Example: Test Scores



Parametric Density Estimation

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

- Discrete case: eg. Bernoulli distribution

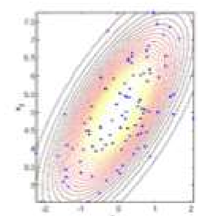
$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

$1 \rightarrow Head$
$0 \rightarrow Tails$

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

- Continuous case: eg. Gaussian distribution in R^n

$$p(x|\mu,\Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right)$$



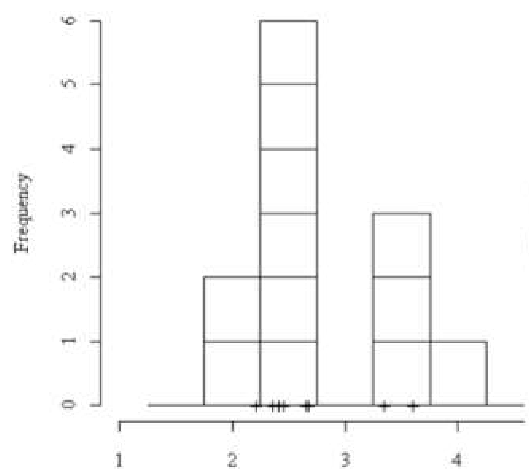
Two sets of parameters $\{\mu,\Sigma\}$, which again generate a family of models, $\mathcal{F} = \{p(x|\mu,\Sigma) \mid \mu \in R^n, \Sigma \in R^{n\times n} \; and \; PSD\}$,

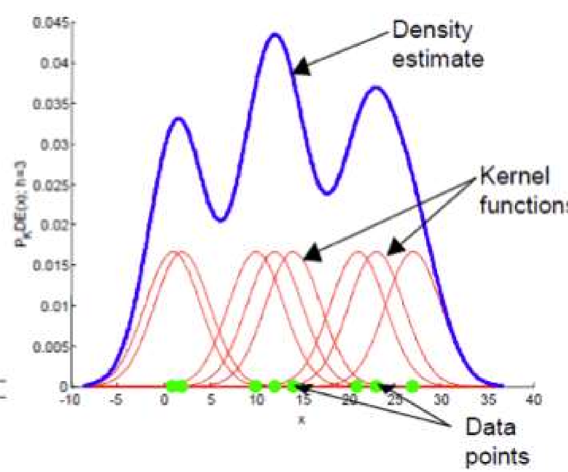
Nonparametric Density Estimation

- 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 of there are many parameters

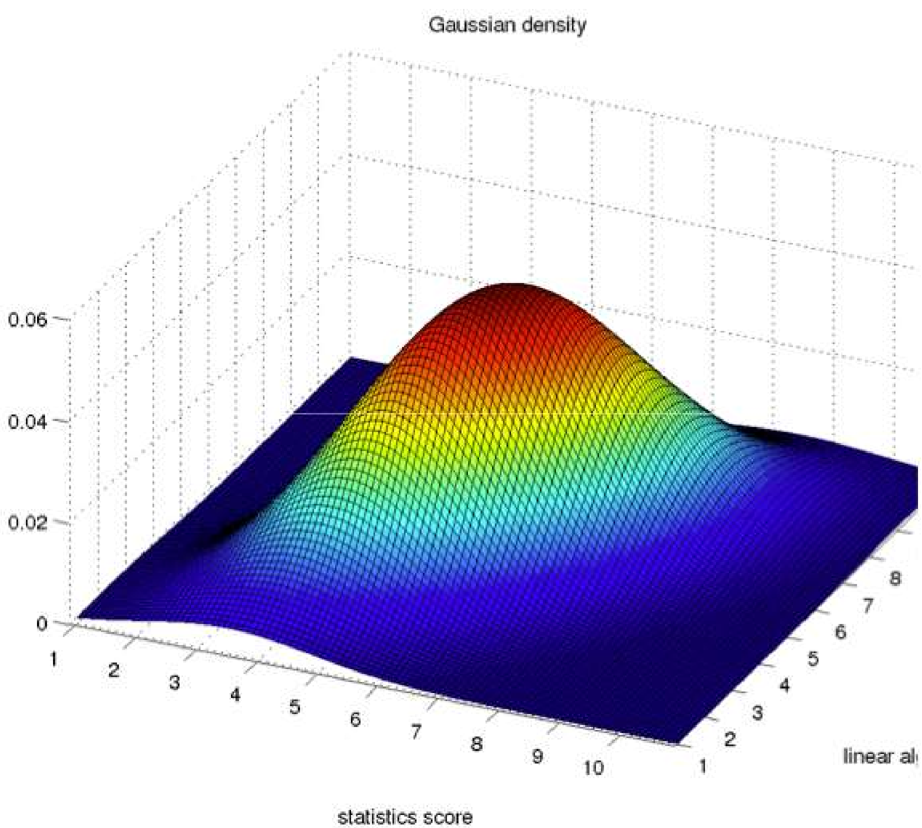
- Eg. Histogram

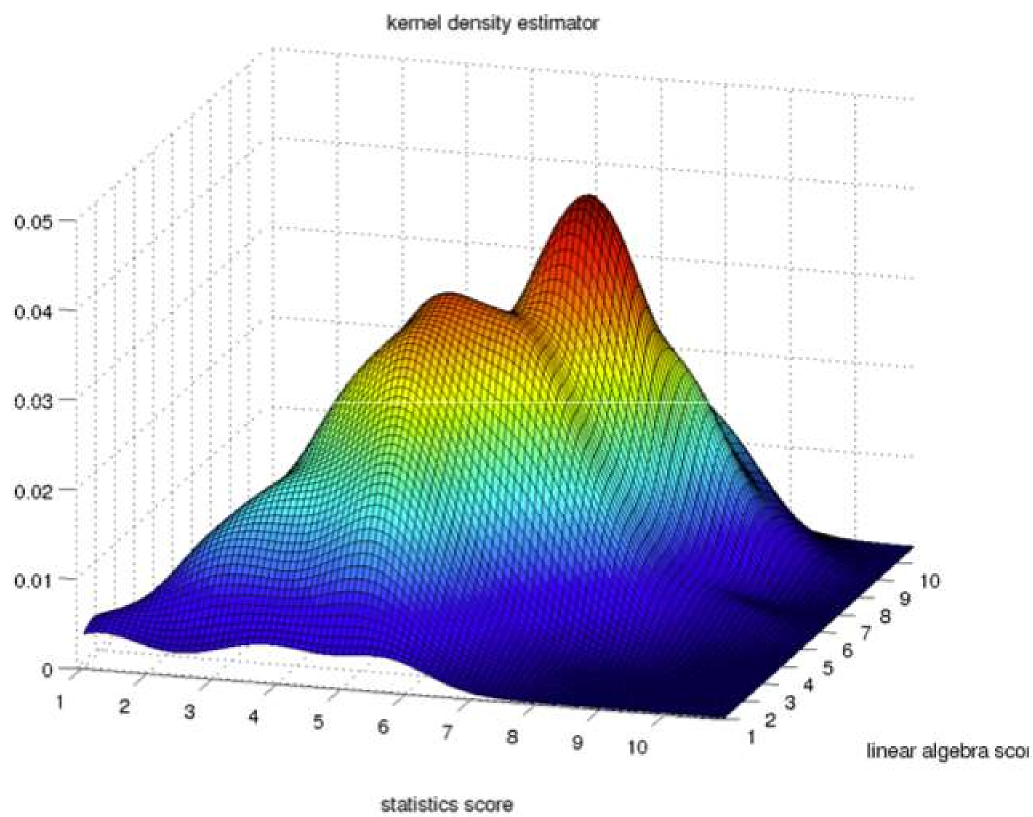
- Eg. Kernel density estimator



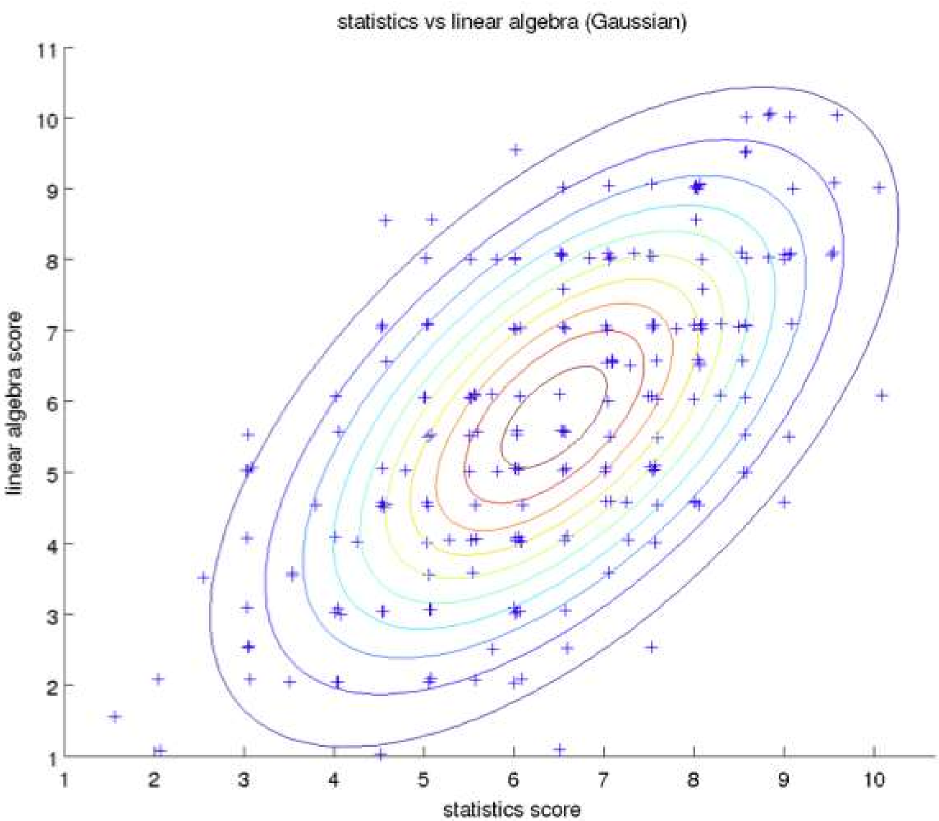


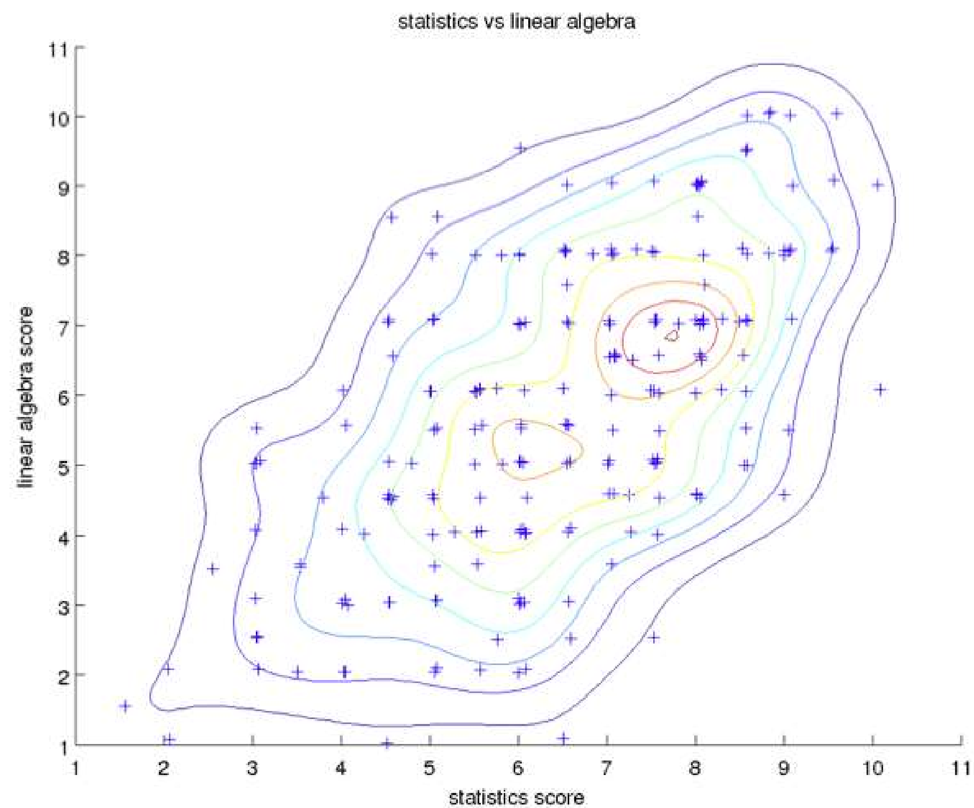
Parametric v.s. Nonparametric Density Estimation





Parametric v.s. Nonparametric Density Estimation





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 m data points $\mathcal{D} = \{x^1, x^2, \ldots x^n\}$ drawn independently and identically (iid) from some distribution $P^*(x)$

Using the parameters, we can estimate each data point

- Want to fit the data with a model $P(x|\theta)$ with parameter θ

$$L_\theta = \theta = argmax_\theta \ \log P(\mathcal{D}|\theta) = argmax_\theta \ \log \prod_{i=1}^{n} P(x^i|\theta)$$

Example Problem

- Estimate the probability θ of landing in heads using a biased coin
- Given a sequence of m independently and identically distributed (iid) flips
  - Eg., $\mathcal{D} = \{x^1, x^2, \dots x^n\} = \{1,0,1, \dots, 0\}, x^i \in \{0,1\}$
- Model: $P(x|\theta) = \theta^x(1-\theta)^{1-x}$
  - $P(x|\theta) = \begin{cases} 1-\theta, for\ x=0 \\ \theta, \quad for\ x=1 \end{cases}$
- Likelihood of a single observation x_i ?
  - $P(x^i|\theta) = \theta^{x^i}(1-\theta)^{1-x^i}$

MLE for Biased Coin

- Objective function, log likelihood

$n_t = n - n_h$

$$l(\theta;\mathcal{D}) = \log P(\mathcal{D}|\theta) = \log \theta^{n_h}(1-\theta)^{n_t}$$
$$= n_h \log \theta + (\text{n} - n_h)\log(1-\theta)$$

n_h: number of heads, n_t: number of tails

- Maximize $l(\theta;\mathcal{D})$ w.r.t. θ

- Take derivatives w.r.t. θ

$$\frac{\partial l}{\partial \theta} = \frac{n_h}{\theta} - \frac{(\text{n} - n_h)}{1-\theta} = 0$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{n_h}{\text{n}} \text{ or } \hat{\theta}_{MLE} = \frac{1}{\text{n}}\sum_i x^i$$

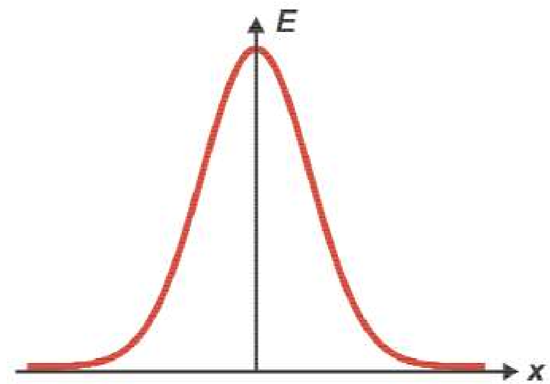
$\text{n} = 100(fliping\ a\ coing)$
$n_h = 50, n_t = 50$
$\theta = 0.5$

Estimating Gaussian Distributions

- Gaussian distribution in R

$$p(x|\mu,\sigma) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Need to estimate two sets of parameters μ, σ



- Given n iid samples

$$\mathcal{D} = \{x^1, x^2, \dots x^n\}, x^i \in R$$

- Density of a data point:

$$p(x^i|\mu,\sigma) \propto exp\left(-\frac{1}{2\sigma^2}(x^i-\mu)^2\right)$$

Estimating Gaussian Distributions

- Gaussian distribution in R

$$p(x|\mu,\sigma) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^i$$

- Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x^i-\mu\right)^2$$

MLE for Gaussian Distribution

- Objective function, log likelihood

$$l(\mu,\sigma;\mathcal{D}) = \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\left(-\frac{1}{2\sigma^2}(x^i-\mu)^2\right)$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(x^i-\mu)^2}{2\sigma^2}$$

- Maximize $l(\mu,\sigma;\mathcal{D})$ with respect to μ, σ

- Take derivatives w.r.t. μ, σ^2

$$\frac{\partial l}{\partial \mu} = 0$$

$$\frac{\partial l}{\partial \sigma^2} = 0$$

MLE for Gaussian Distribution

$\sigma^2 = U$ $\qquad \frac{n}{2}\log U \qquad = \frac{n}{2}\frac{1}{U}$ $\quad \sigma^2$

$$l(\mu, \sigma; \mathcal{D}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{(x^i - \mu)^2}{2\sigma^2}$$

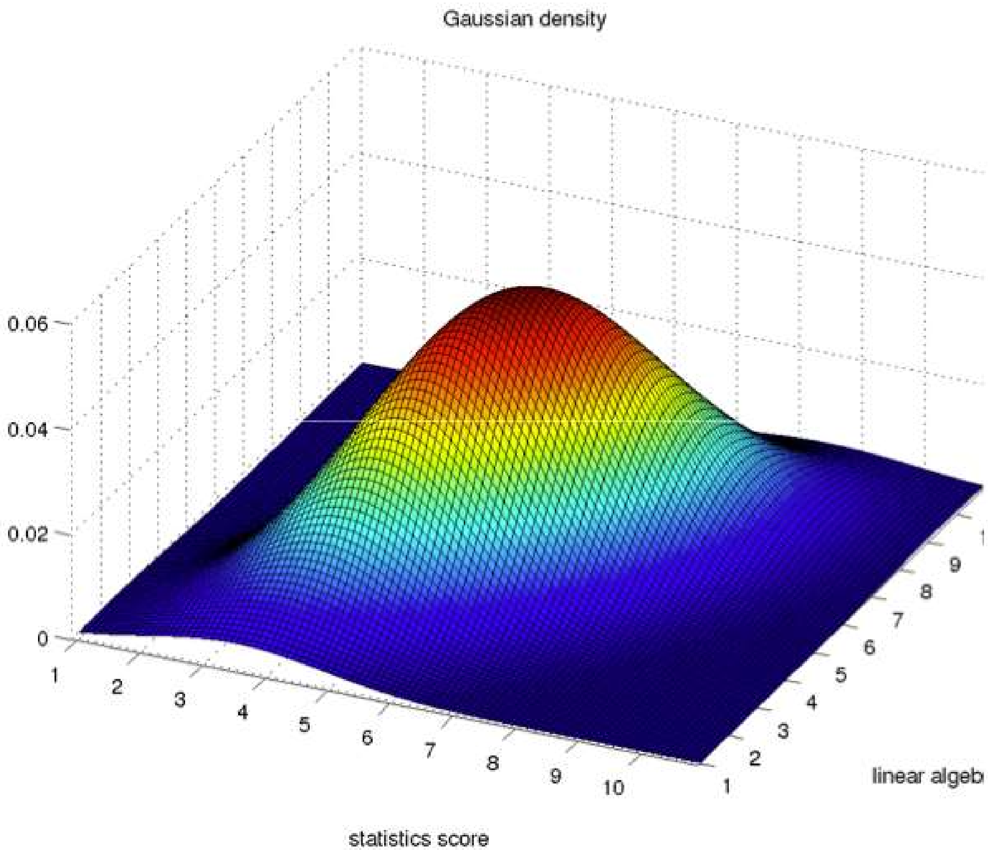
$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x^i - \mu) = 0$$

$$\Rightarrow \sum_{i}^{n} x^i = n\,\mu \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x^i$$

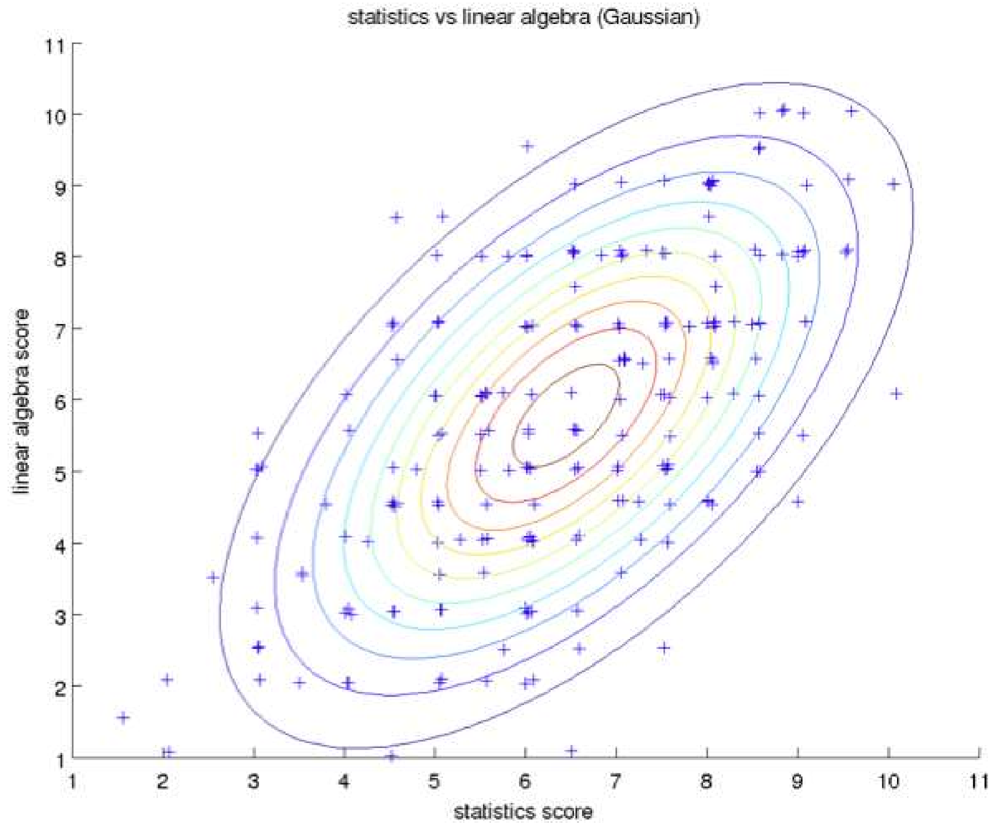
$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i}^{n}(x^i - \mu)^2 = 0$$

$$\Rightarrow \sum_{i}^{n}(x^i - \mu)^2 = n\,\sigma^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)^2$$

Example



Gaussian density
0.06
0.04
0.02
0
1 2 3 4 5 6 7 8 9 10
statistics score
1 2 3 4 5 6 7 8 9
linear algeb



statistics vs linear algebra (Gaussian)
linear algebra score
1 2 3 4 5 6 7 8 9 10 11
statistics score

Outline

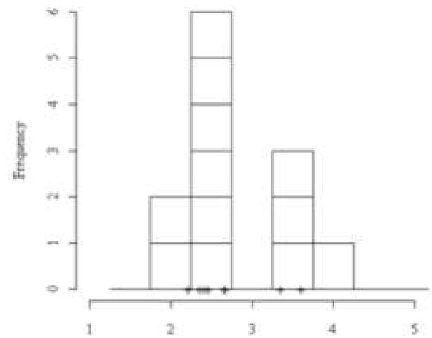
- Overview
- Parametric Density Estimation
- Nonparametric Density Estimation

Can be used for:

- Visualization
- Classification
- Regression

1-D Histogram

- One the simplest nonparametric density estimator
- Given ~~m~~ n iid samples $\mathcal{D} = \{x^1, x^2, \dots x^n\}, x^i \in [0,1)$



1 bin wich

$\frac{1}{m}$ $2 \times \frac{1}{m}$

- Split [0,1) into m bins

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \dots B_m = \left[\frac{m-1}{m}, 1\right)$$

c_1 1 bin width

- Count the number of points, c_1 within B_1, c_2 within B_2...

$$\frac{1}{m}$$

- For a new test point x

Identity matrix

$$p(x) = \sum_{j=1}^{m} \frac{m c_j}{n} I(x \in B_j) = \frac{number\ of\ points\ in\ bin\ c}{total\ number\ of\ data\ points\ \times bin\ width}$$

$\frac{1}{m}$

$$P = \int p(x)dx$$

The probability that point x is drawn from a distribution p(x)

Why is Histogram Valid?

- Requirement for density $p(x)$
- $p(x) \geq 0, \int_\Omega p(x)dx = 1$

$$\sum_{j=1}^{m} c_j = n$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- For histogram,

$$\int_{[0,1)} p(x)dx = \int_{[0,1)} \sum_{j=1}^{m} \frac{\mathrm{m}c_j}{\mathrm{n}} I(x \in B_j)dx$$

$B_1 =$

$$= \int_0^{\frac{1}{m}} \sum_{j=1}^{m} \frac{mc_j}{n} I\, dx + \int_{\frac{1}{m}}^{\frac{2}{m}} \sum_{j=1}^{m} \frac{mc_j}{n} I\, dx + \ldots + \int_{\frac{j-1}{m}}^{\frac{j}{m}} \sum_{j=1}^{m} \frac{mc_j}{n} I\, dx = \sum_{j=1}^{m} \int_{[\frac{j-1}{\mathrm{m}}, \frac{j}{\mathrm{m}})} \frac{\mathrm{m}c_j}{\mathrm{n}} dx$$

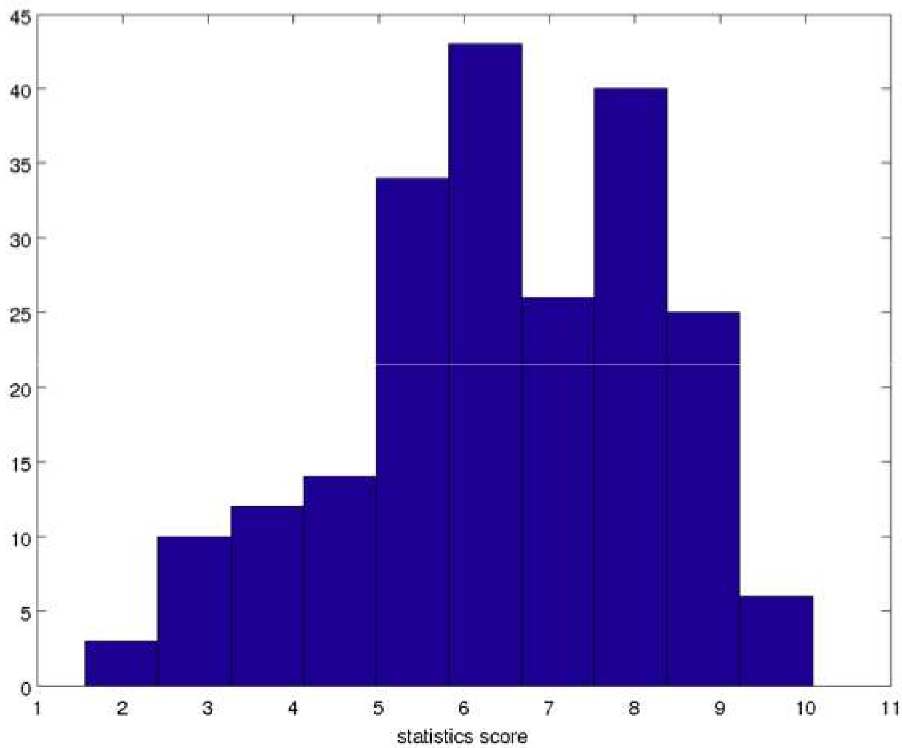
$0, \frac{1}{m}$

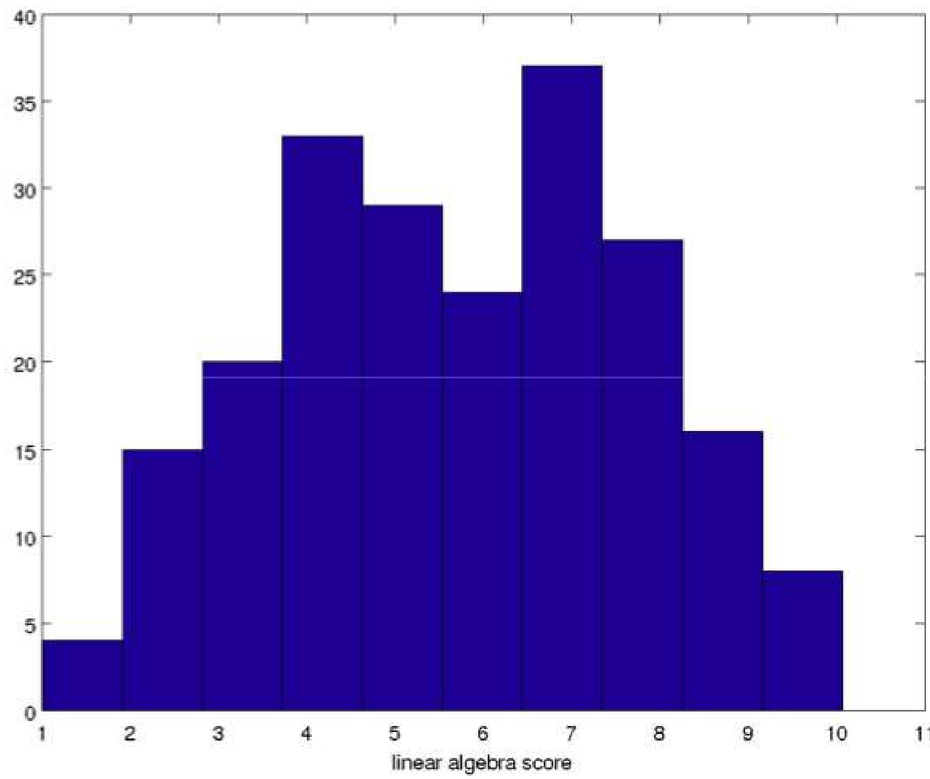
$\frac{1}{m}\ \frac{2}{m}$

$$= \sum_{j=1}^{m} \frac{\cancel{m}c_j}{n} I \underbrace{\left[\frac{j}{m} - \frac{j-1}{m}\right]}_{\frac{1}{m}} = \sum_{j=1}^{m} \frac{c_j}{\mathrm{n}} = 1$$

Example: Test Scores

- What is missing if we want density?





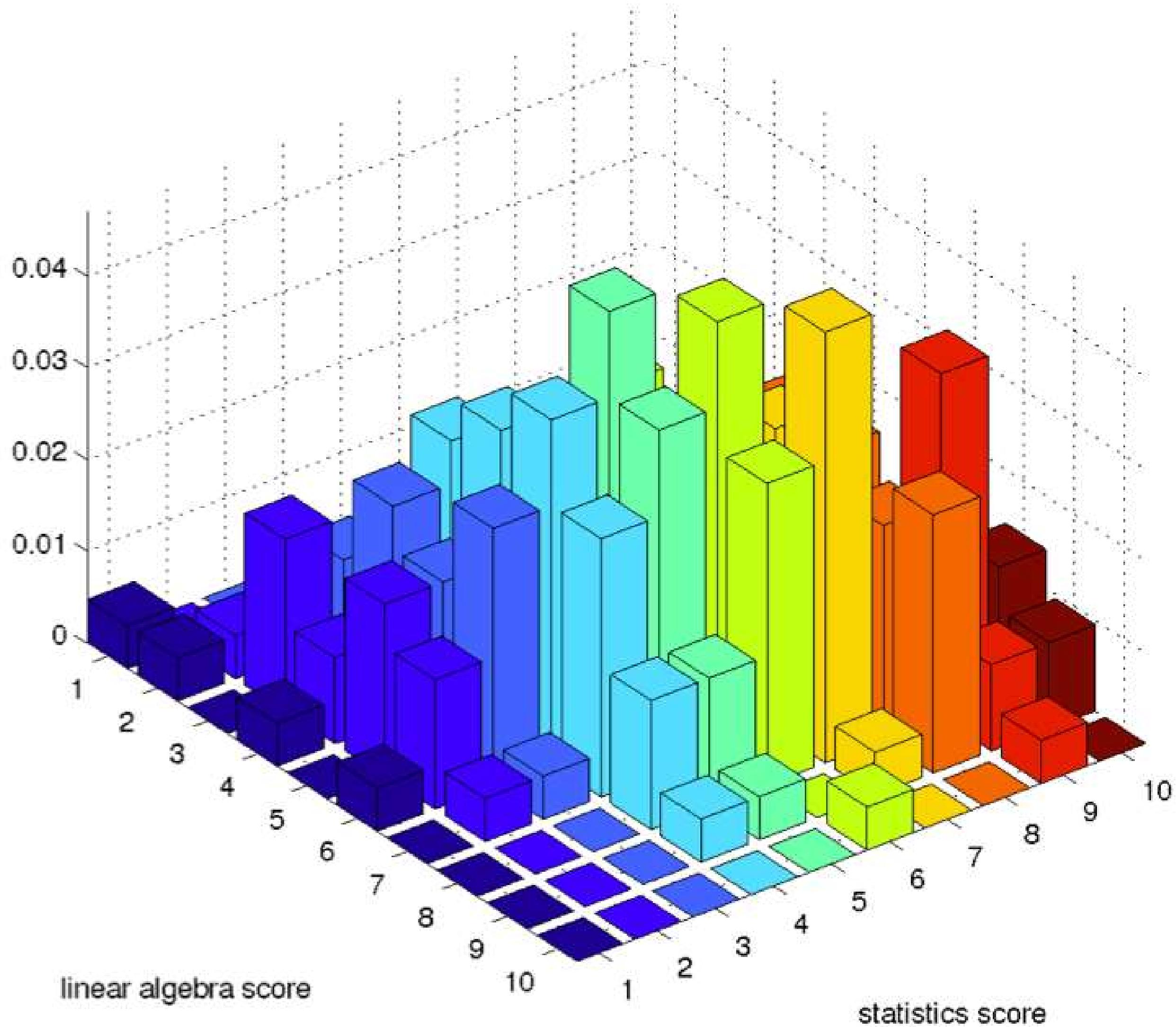
Higher-Dimensional Histogram

- Given n iid samples $\mathcal{D} = \{x^1, x^2, \dots x^n\}, x^i \in [0,1)^d$
- Split $[0,1)^d$ evenly into m^d bins
- Bin size is $h = \frac{1}{m}$

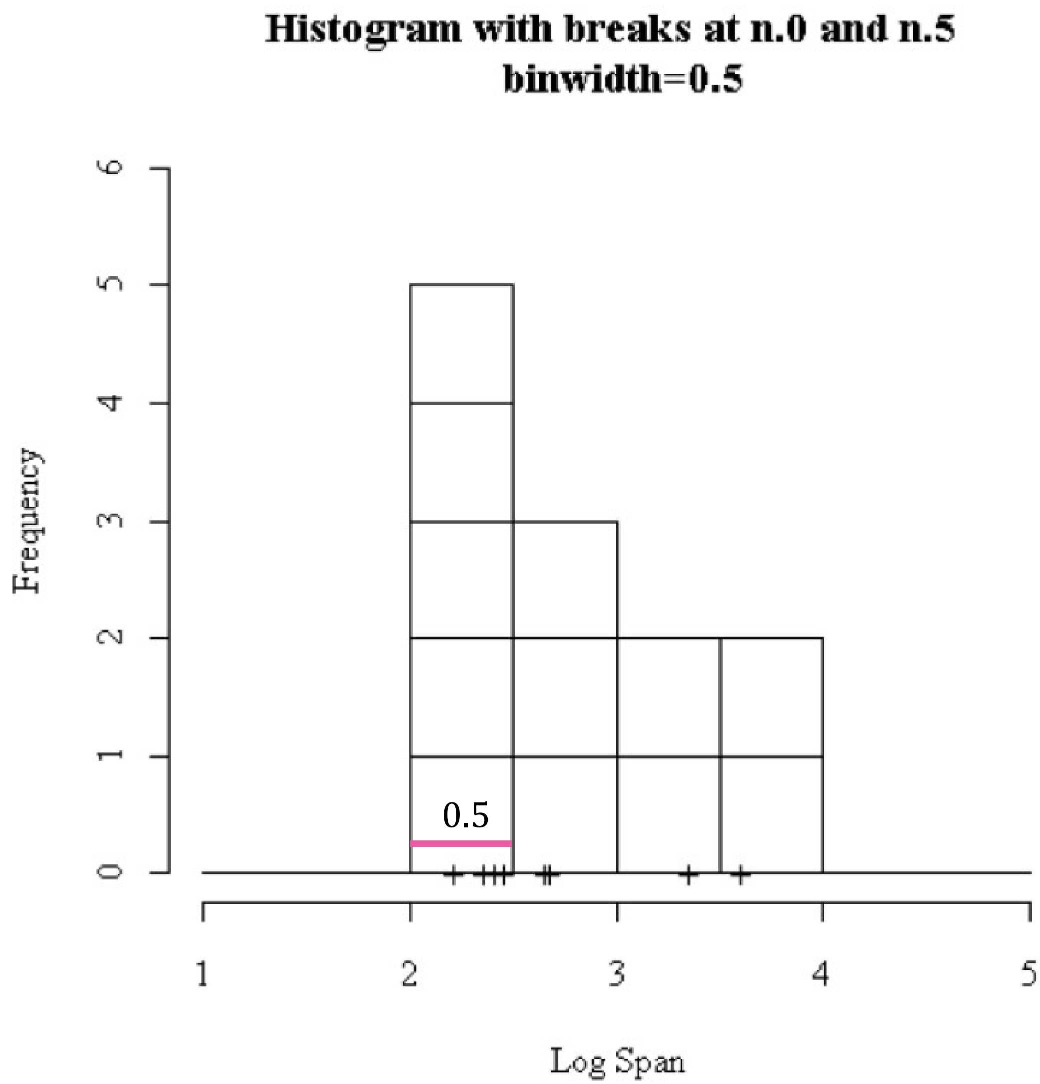
$Two\ Dimensional\ data$:

$m = 10$ (number of bins in each dimension)

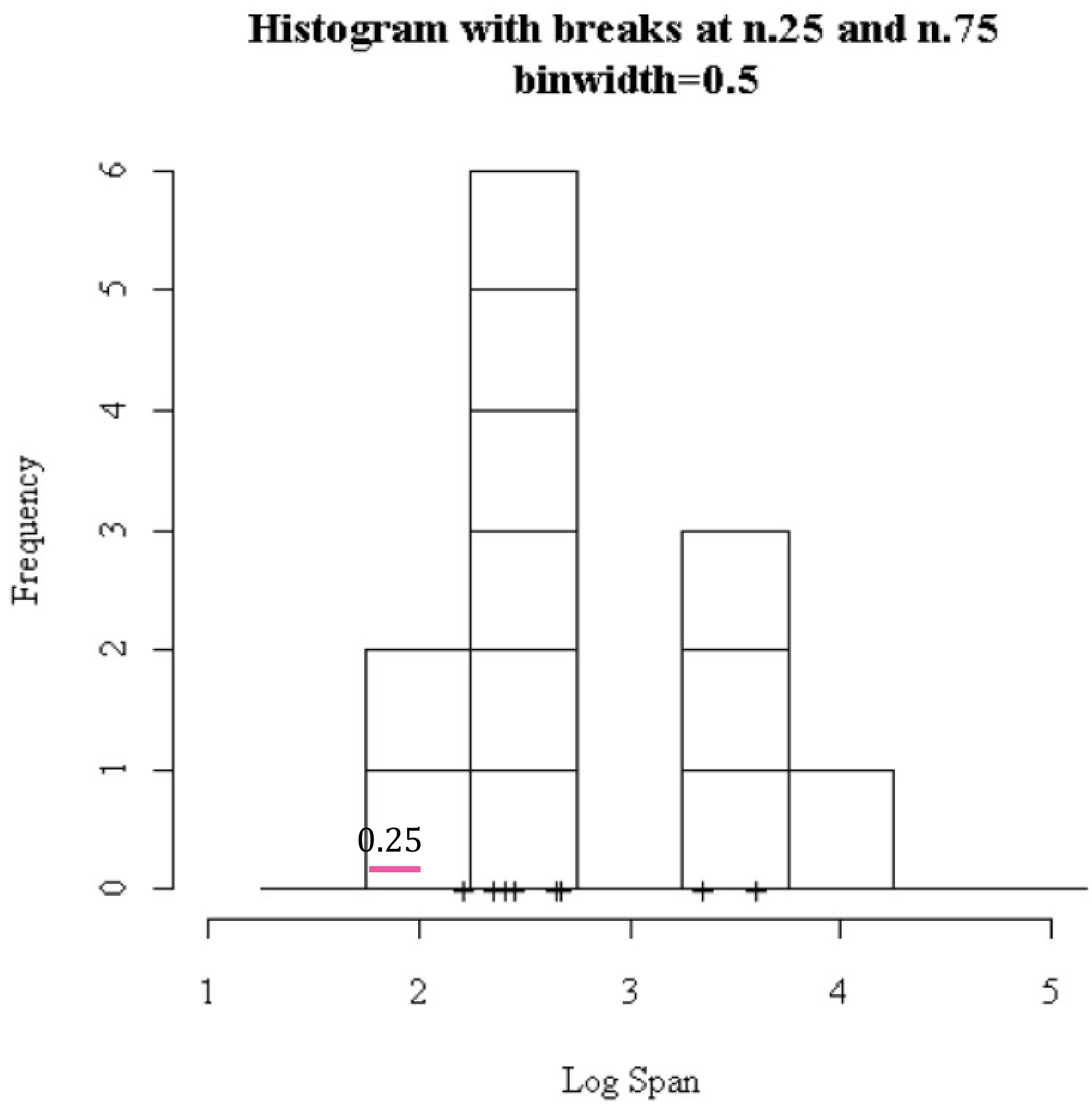
$m^2 = 100$ (total number of bins for two dimensional data)


0.04
0.03
0.02
0.01
0
1
2
3
4
5
6
7
8
9
10
linear algebra score
1
2
3
4
5
6
7
8
9
10
statistics score

Output Depends on Where You Put the Bins



Output Depends on Where You Put the Bins



Kernel Density Estimation

- Kernel density estimator

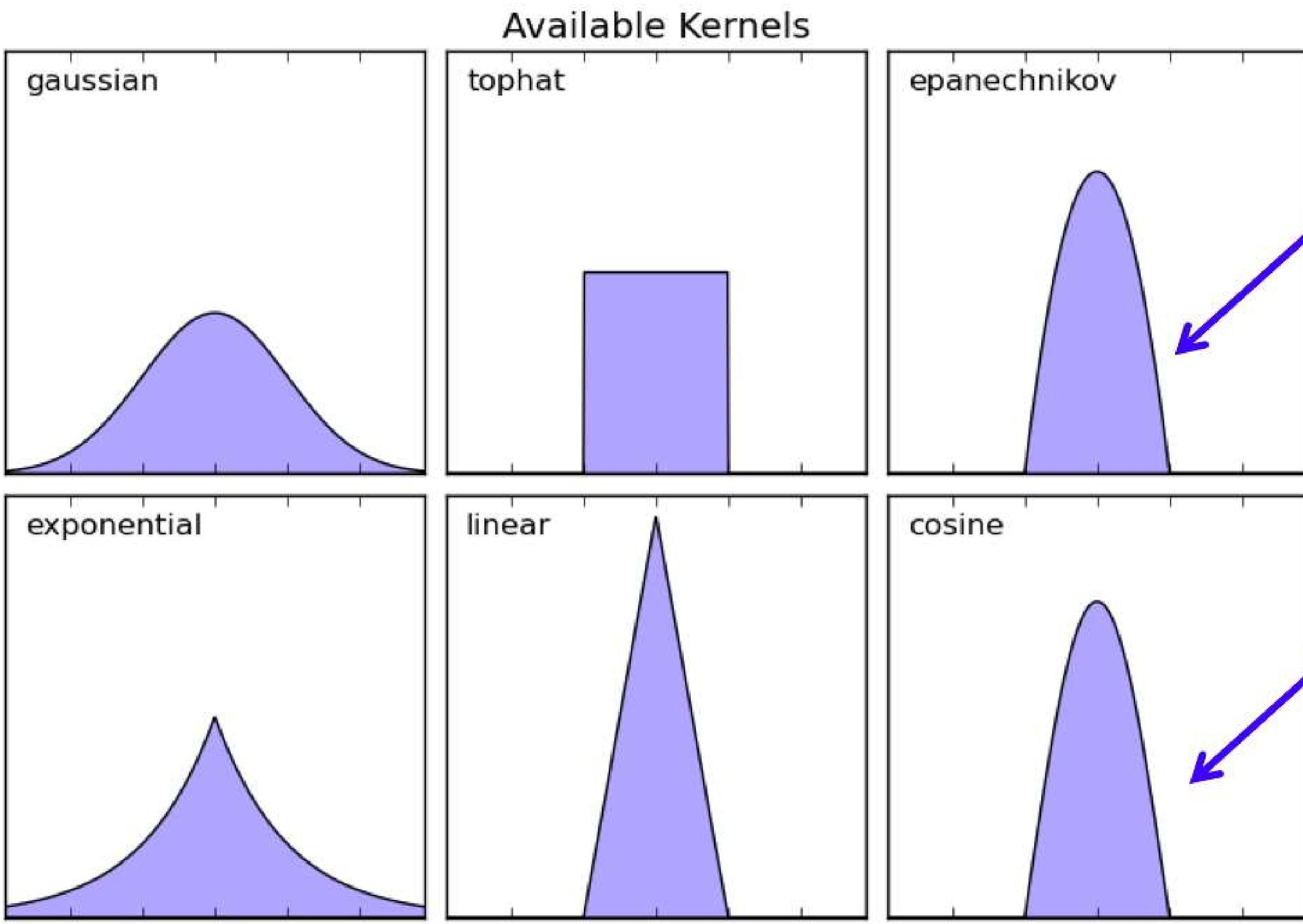
$$p(x) = \frac{1}{n}\sum_{i}^{n}\frac{1}{h}K\left(\frac{x^i - x}{h}\right)$$

- Smoothing kernel function
  - $K(u) \geq 0,$
  - $\int K(u)du = 1,$
  - $\int uK(u) = 0,$
  - $\int u^2K(u)du \leq \infty$

- An example: Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$

Smoothing Kernel Functions

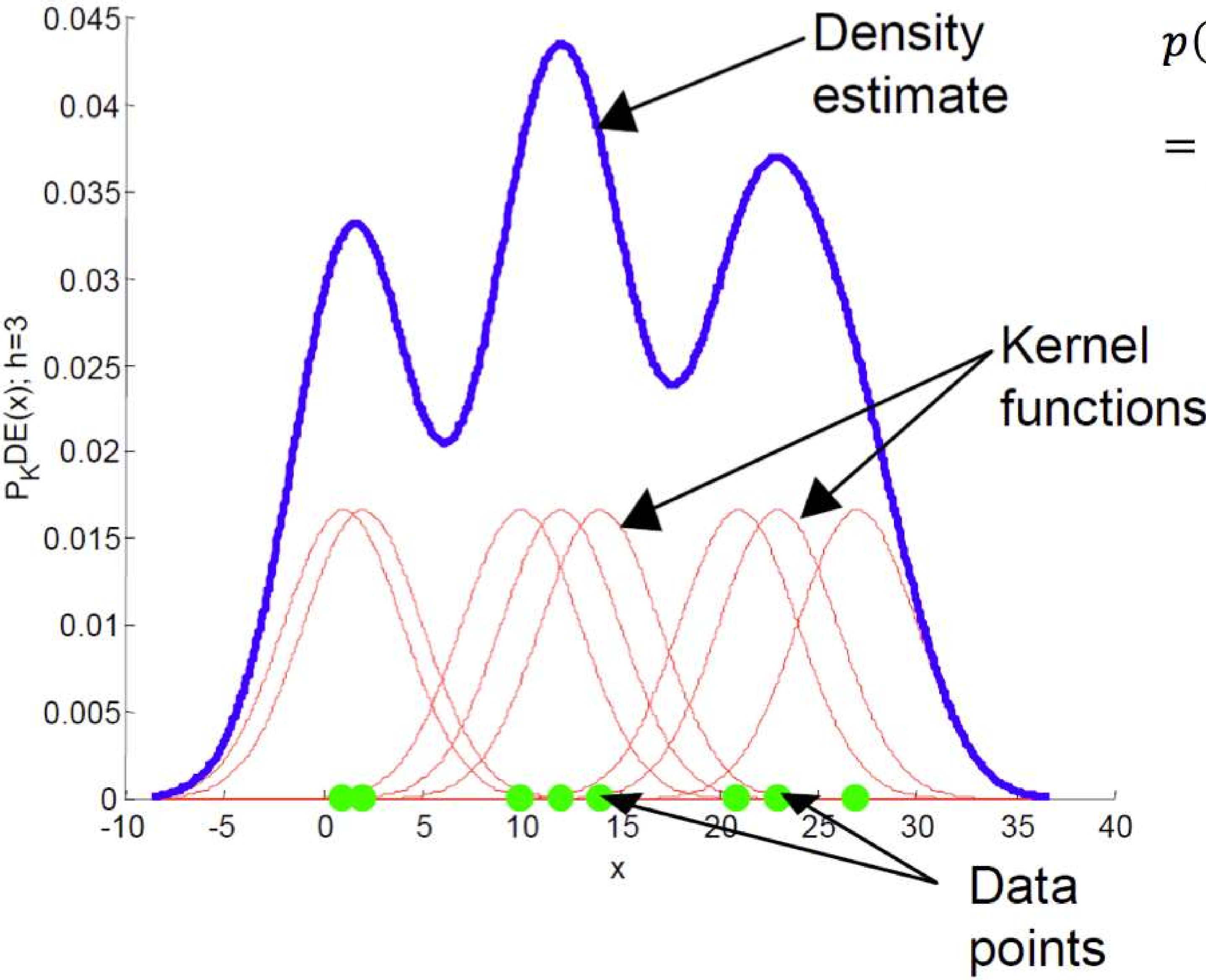
- An example: Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$



$K(u) =$
$\frac{3}{4}(1 - u^2) I(|u| \leq 1)$

$K(u) =$
$\frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right) I(|u| \leq 1)$

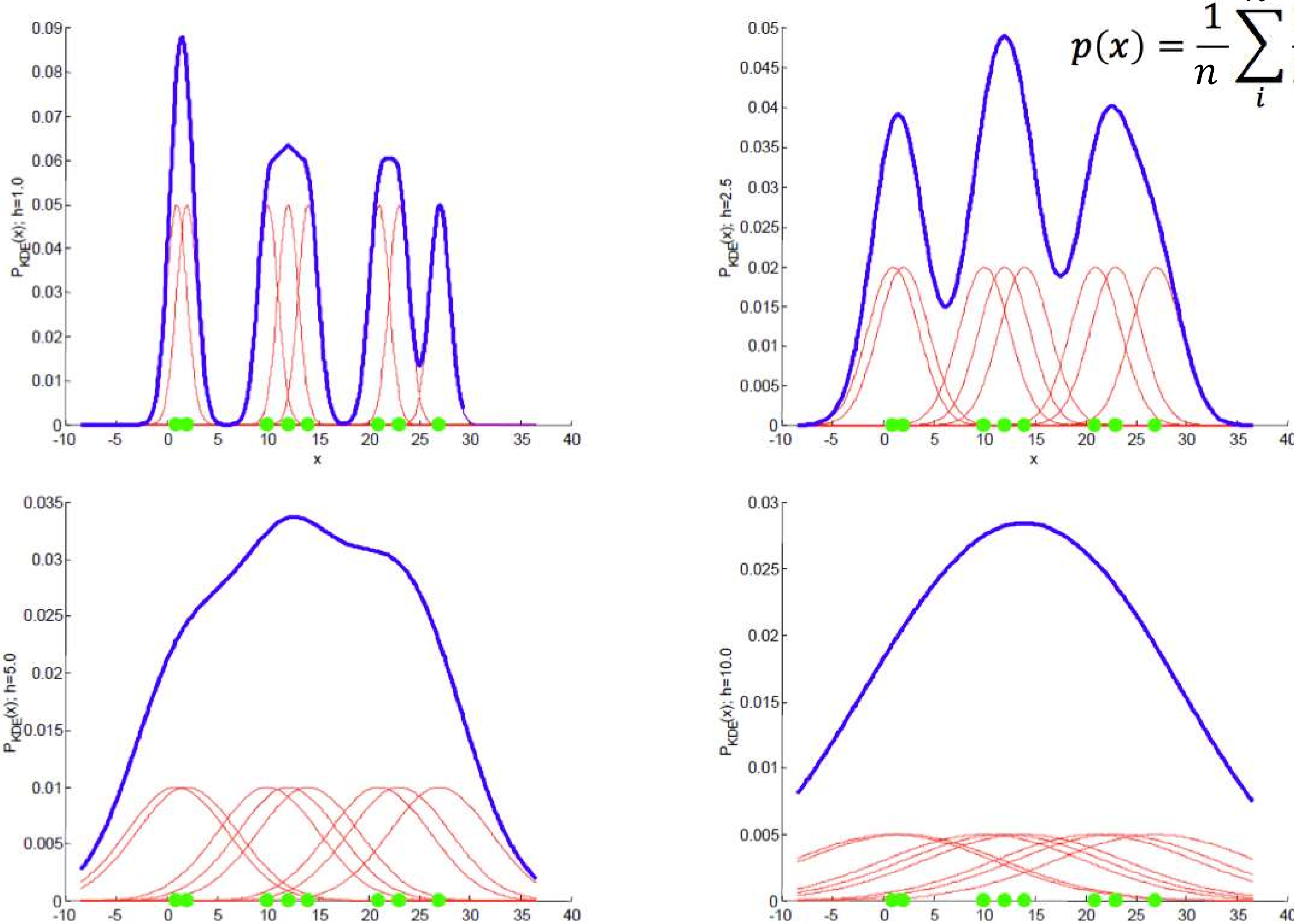
Example



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

Effect of the Kernel Bandwidth

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

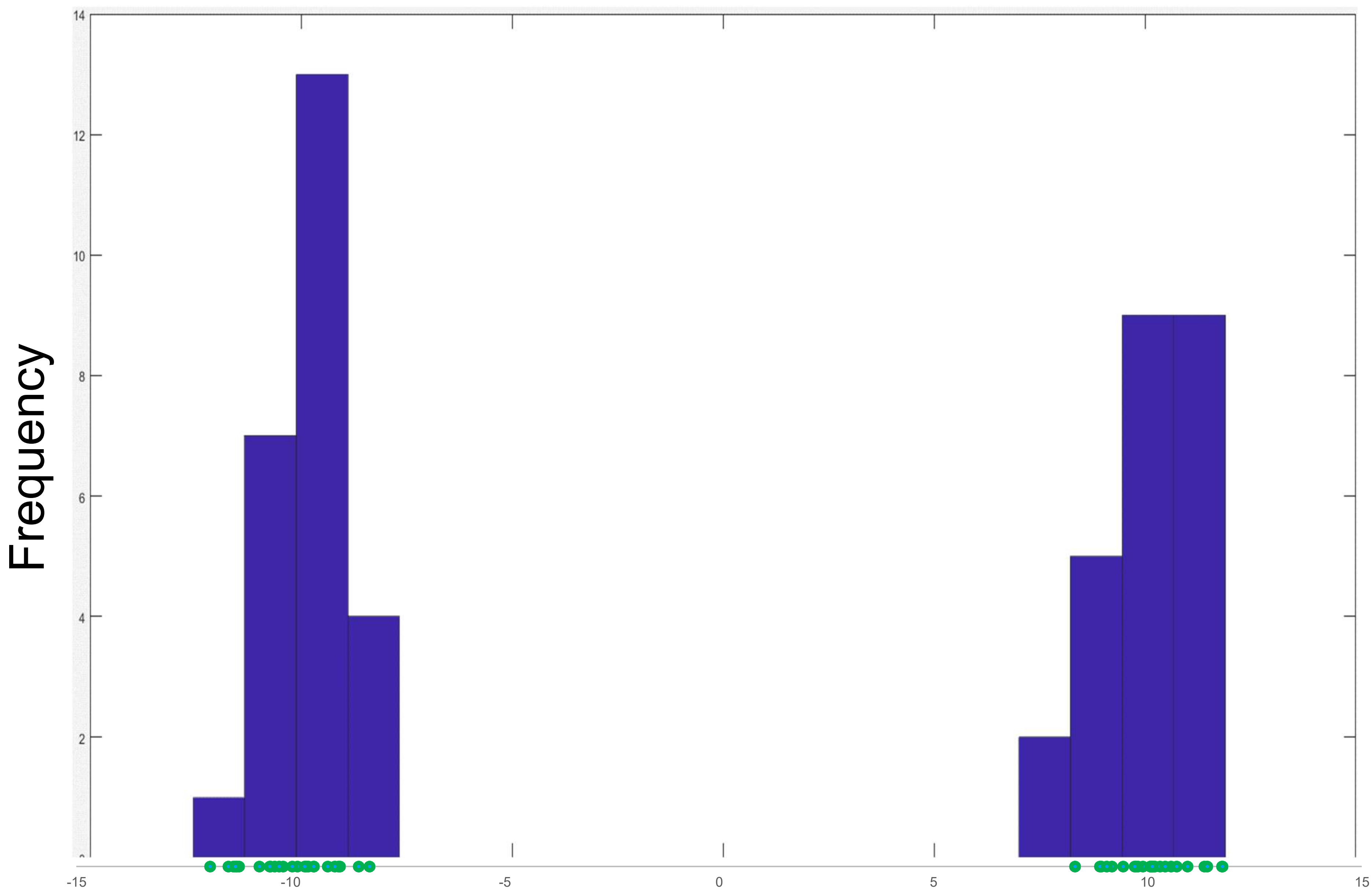


Visual Example

50 datapoints are given to us

Visual Example

Let's implement 20 bins histogram

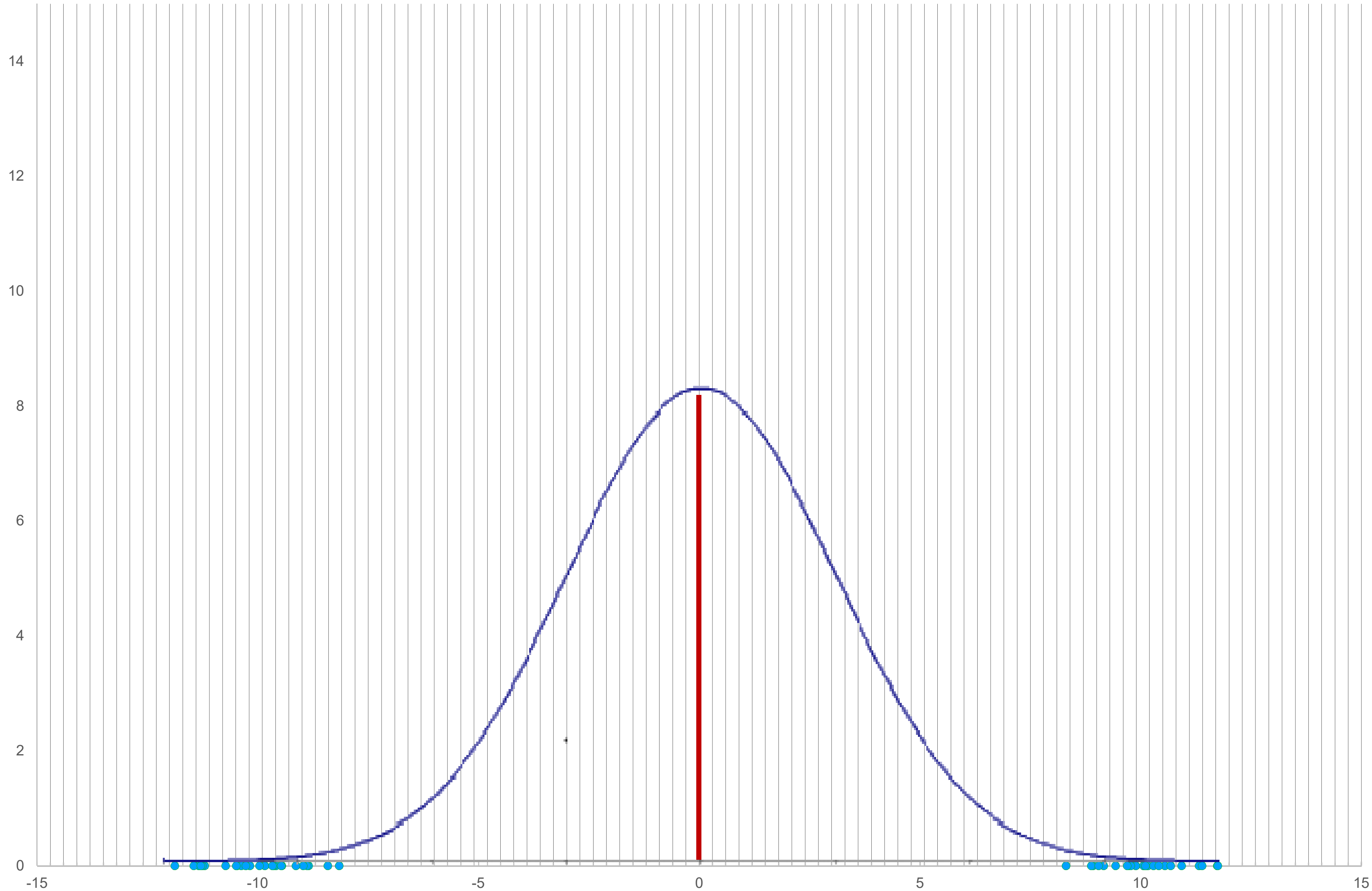


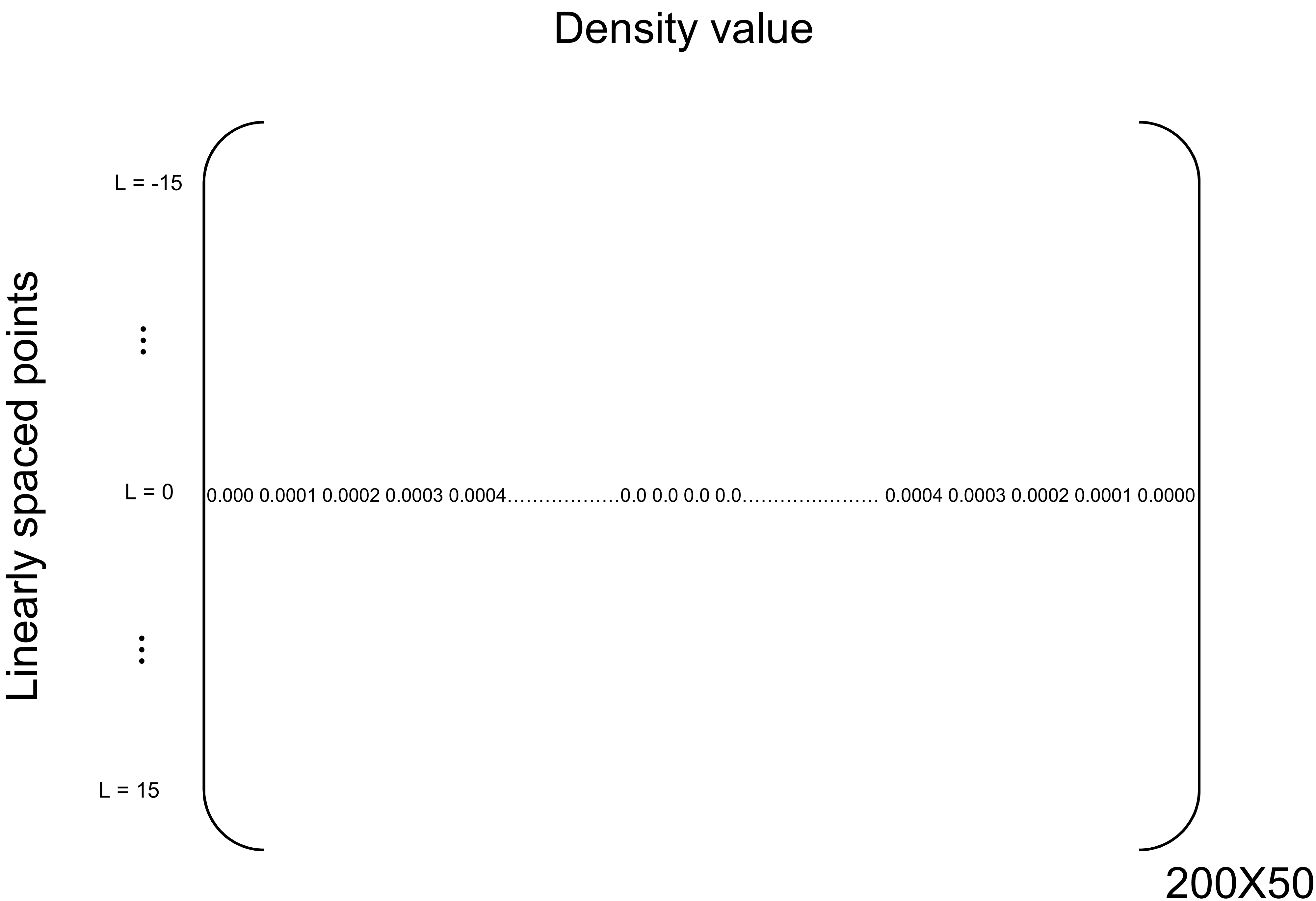
Visual Example

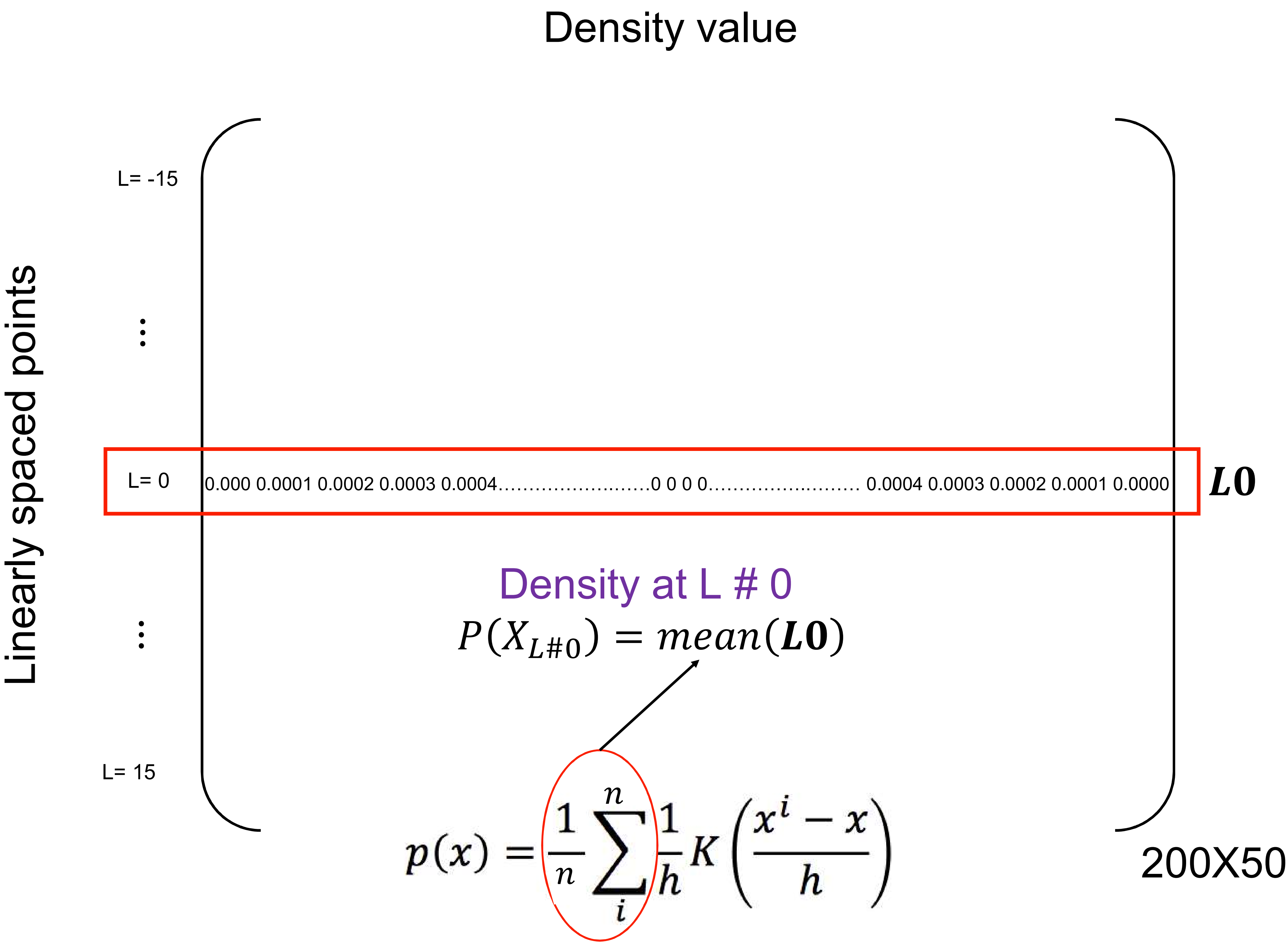
Let’s create 200 uniform points to have a smoother density function
OR simply you can just implement this on each datapoint

Visual Example

For **each** linearly spaced point, let’s calculate the Gaussian kernel value over the given 50 points
As an example, let’s do it for the 0

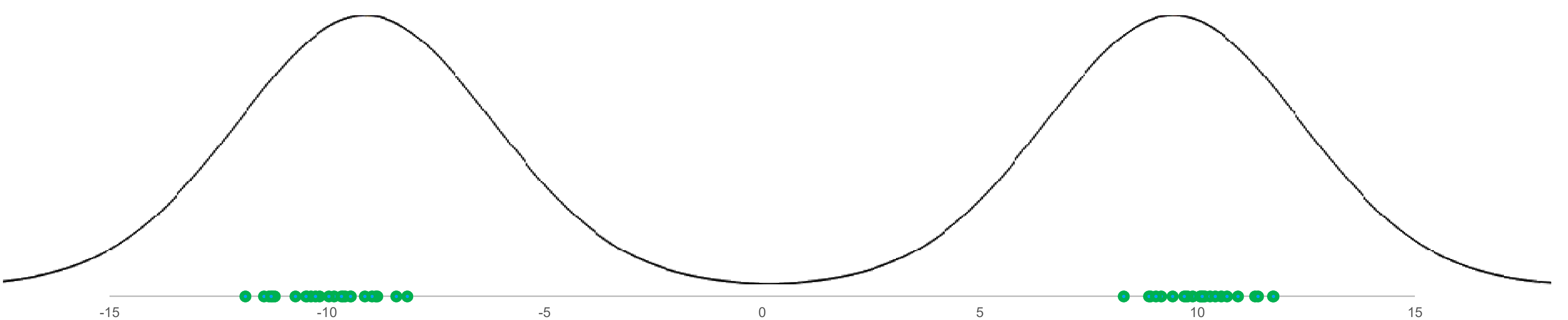



Density value
Linearly spaced points
L = -15
L = 0
L = 15
0.000 0.0001 0.0002 0.0003 0.0004……………..0.0 0.0 0.0 0.0………………… 0.0004 0.0003 0.0002 0.0001 0.0000
200X50


Density value
Linearly spaced points
L= -15
L= 0
L= 15
0.000 0.0001 0.0002 0.0003 0.0004……………………0 0 0 0…………………… 0.0004 0.0003 0.0002 0.0001 0.0000
$\boldsymbol{L0}$
Density at L # 0
$P(X_{L\#0}) = mean(\boldsymbol{L0})$
$p(x) = \frac{1}{n}\sum_{i}^{n}\frac{1}{h}K\left(\frac{x^i - x}{h}\right)$
200X50

Visual Example

Based on Gaussian kernel estimator



Numerical Example

$For\ \sigma = 1;$

```
% Data ; There are 200 data points (-13~<data<~13)
randn('seed',1)                                  % Used for reproducibility
data = [randn(100,1)-10; randn(100,1)+10]; % Two Normals mixed (GROUND TRUTH)
```

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

$$h = \left(\frac{4\hat{\sigma}^2}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

```
h = std(data)*(4/3/numel(data))^(1/5);      % Bandwidth estimated by Silverman's Rule of Thumb
```

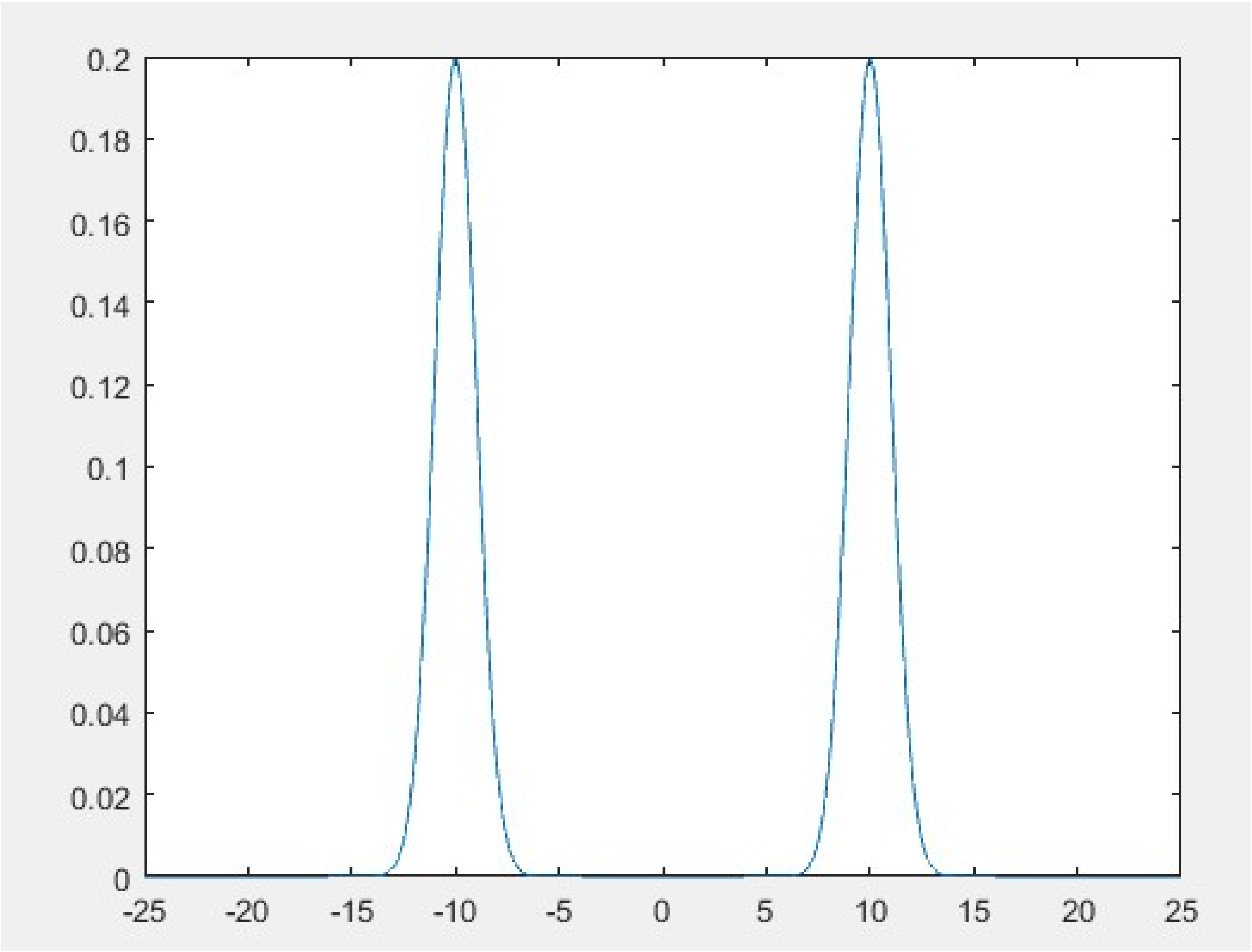
```
% Let's create apply density estimation over 1000 linearly spaced points

x = linspace(-25,+25,1000);

% Let's generate a "TRUE" density over all the bins given the "Ground Truth" information.

truepdf_firstnormal = exp(-.5*(x-10).^2)/sqrt(2*pi);
truepdf_secondnormal = exp(-.5*(x+10).^2)/sqrt(2*pi);
truepdf = truepdf_firstnormal/2 + truepdf_secondnormal/2;
% divided down by 2, because we are adding density value two times
```

```
plot(x,truepdf)                                  % Plot True Density
```



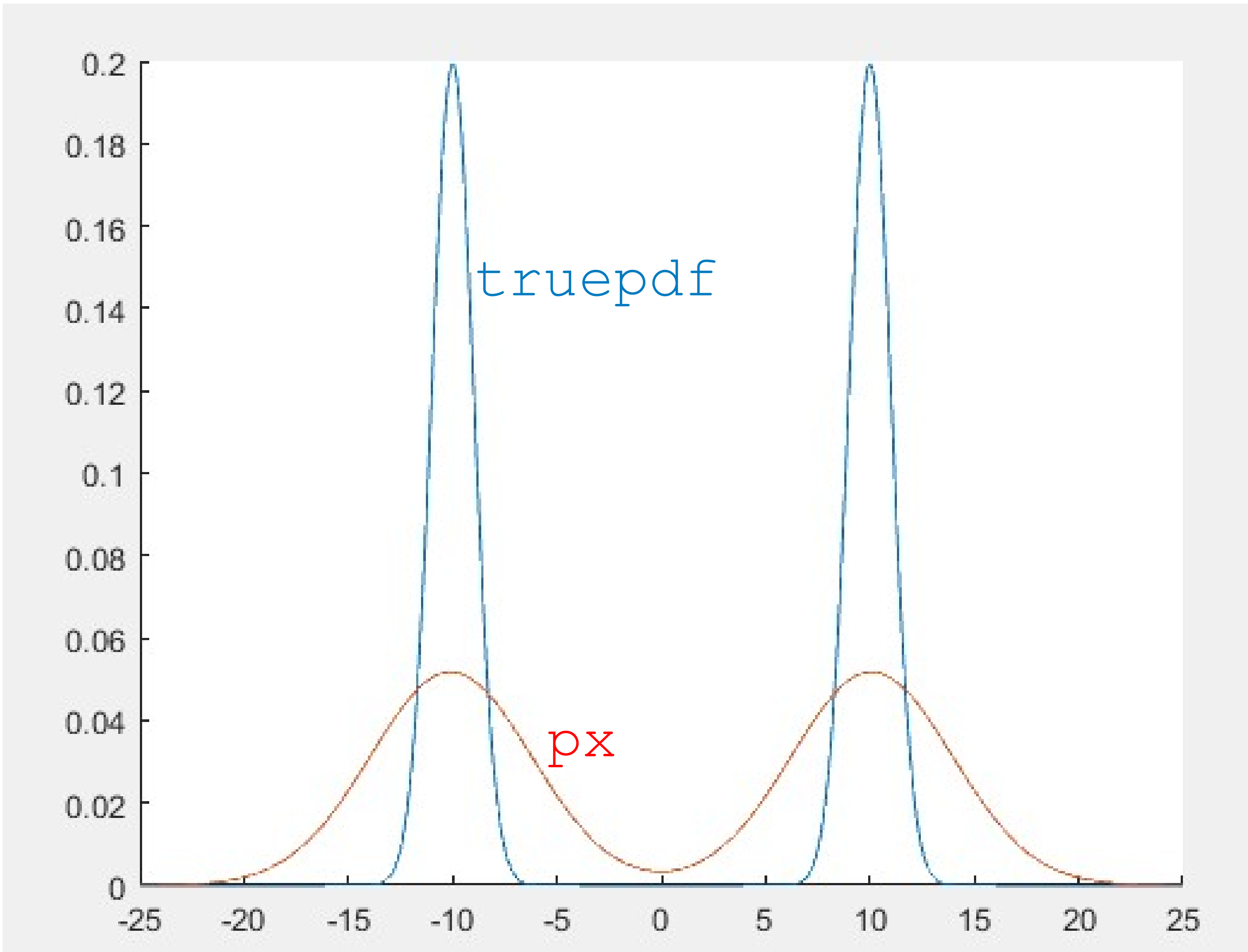
```
% Let's calculate Gaussian kernel density for each linearly spaced
point over 200 Given data points
```

$$p(x) = \frac{1}{n}\sum_{i}^{n}\frac{1}{h}K\left(\frac{x^i - x}{h}\right) \qquad u = \frac{x^i - x}{h}$$

Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$

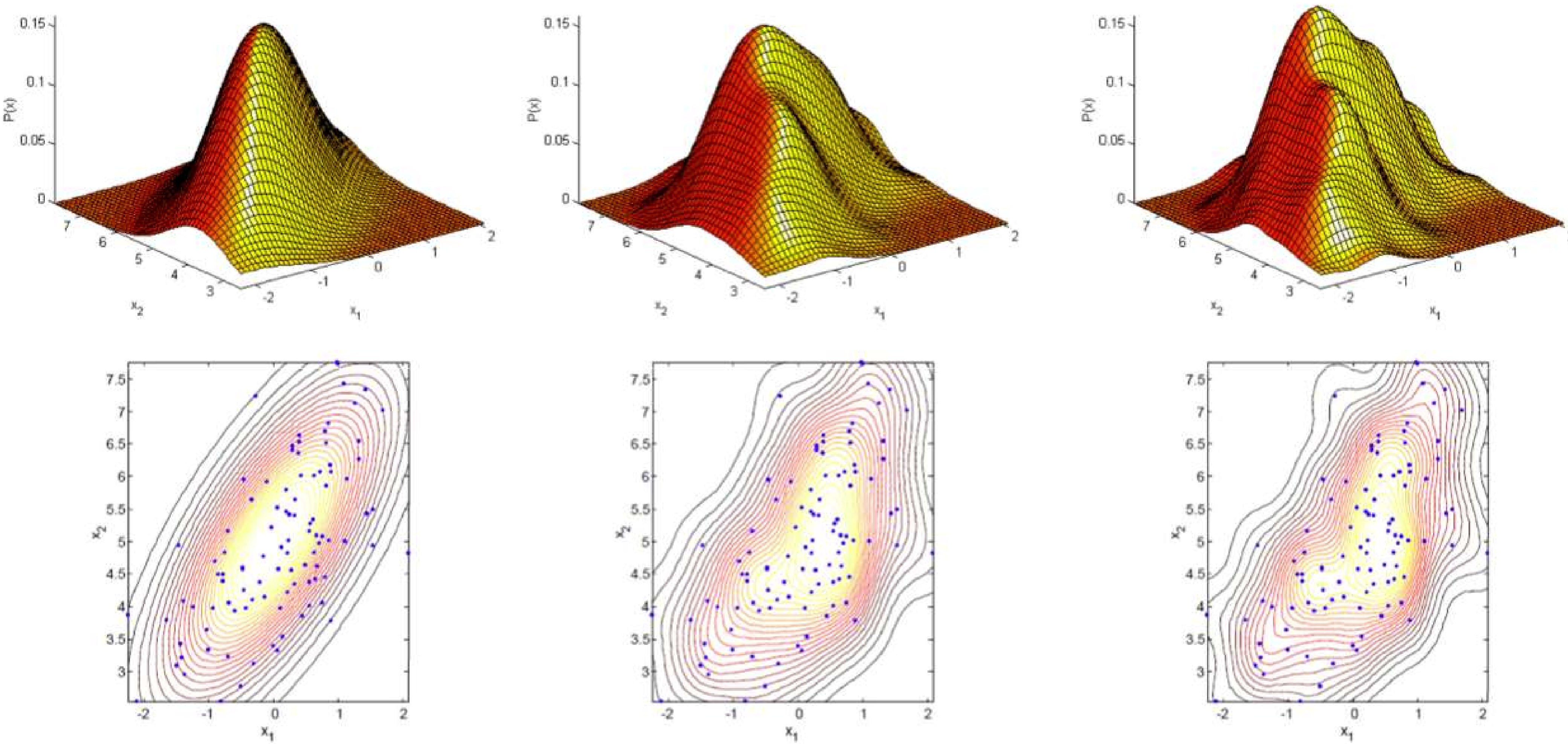
```
for i=1:size(x,1)
    u = (x(i)-data)./h; % length of u is 200
    Ku = exp(-.5*u.^2)/sqrt(2*pi);
    Ku = Ku./h;
    px(i) = mean(Ku);
end
```

```
plot(x,truepdf)
plot(x,px)
```



Two-Dimensional Examples

- This example shows the product KDE of a bivariate <u>unimodal</u> Gaussian
  - 100 data points were drawn from the distribution
  - The figures show the true density (left) and the estimates using $h = 1.06\sigma N^{-1/5}$ (middle) and $h = 0.9AN^{-1/5}$ (right)



Choosing the Kernel Bandwidth

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

$$h \approx 1.06\, \hat{\sigma}\, m^{-1/5}$$

where is the standard deviation of the samples

- A better but more computational 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

Summary

- Parametric density estimation
  - Maximum likelihood estimation
  - Different parametric forms
- Nonparametric density estimation
  - Histogram
  - Kernel density estimation