

Density Estimation

Mahdi Roozbahani
Georgia Tech

Why we love exponential terms?



When I try to take derivative of Exp

Outline

- Overview 
- Parametric Density Estimation
- Nonparametric Density Estimation

Continuous variable

Continuous probability distribution

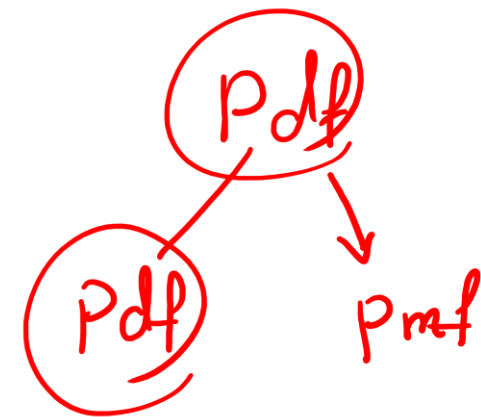
Probability density function $\rightarrow$ pdf

Density value

Temperature (real number)

Gaussian Distribution

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



Discrete variable

Discrete probability distribution

Probability mass function $\rightarrow$ pmf

Probability value

Coin flip (integer)

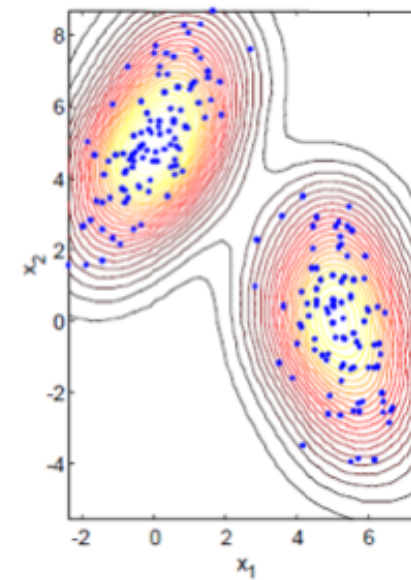
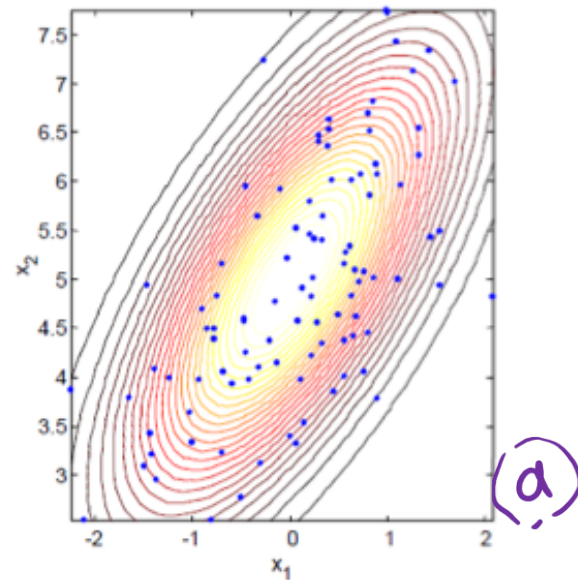
Bernoulli distribution

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

Why Density Estimation?

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

pdf
↓
Multivariate
Gaussian

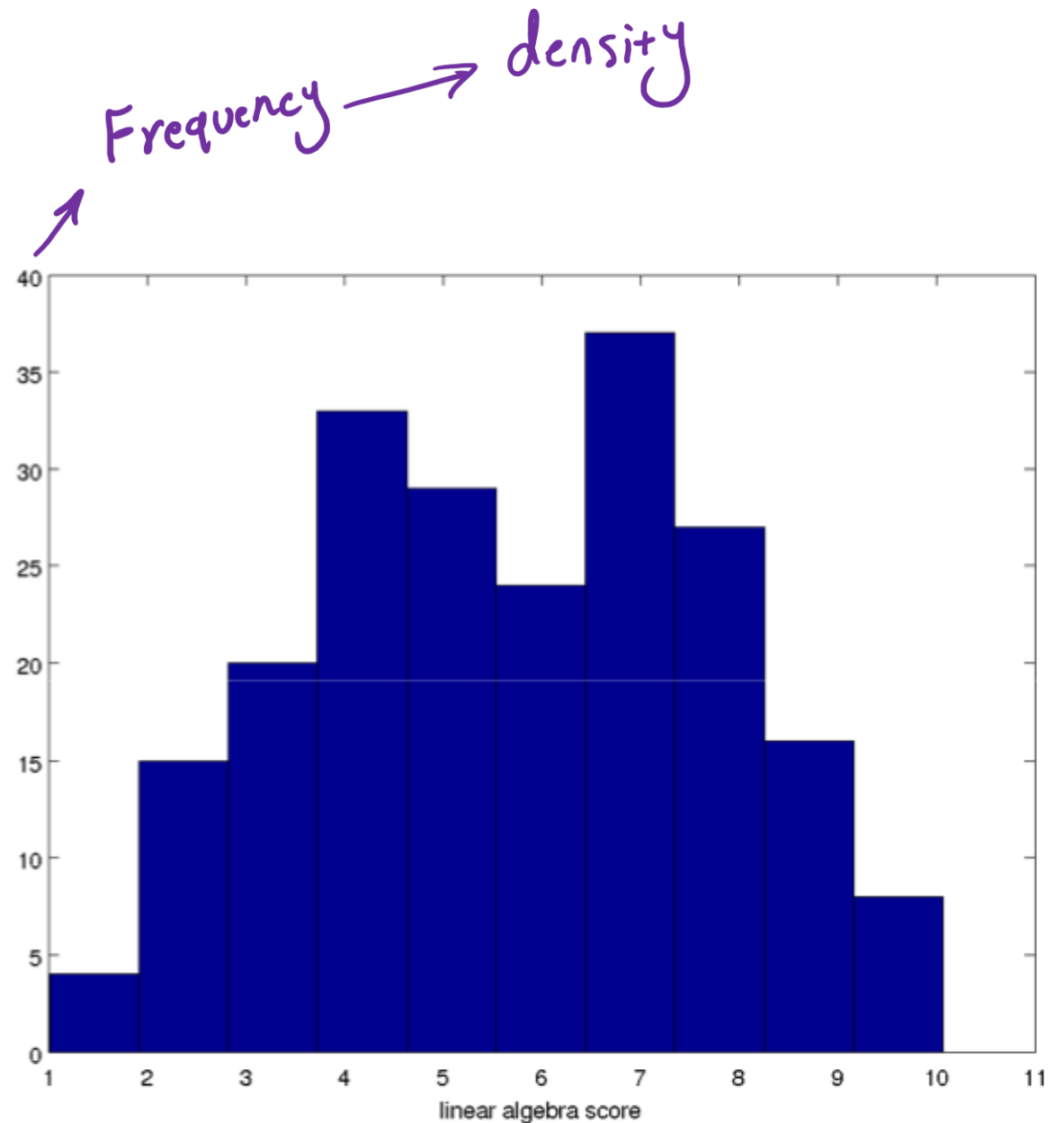
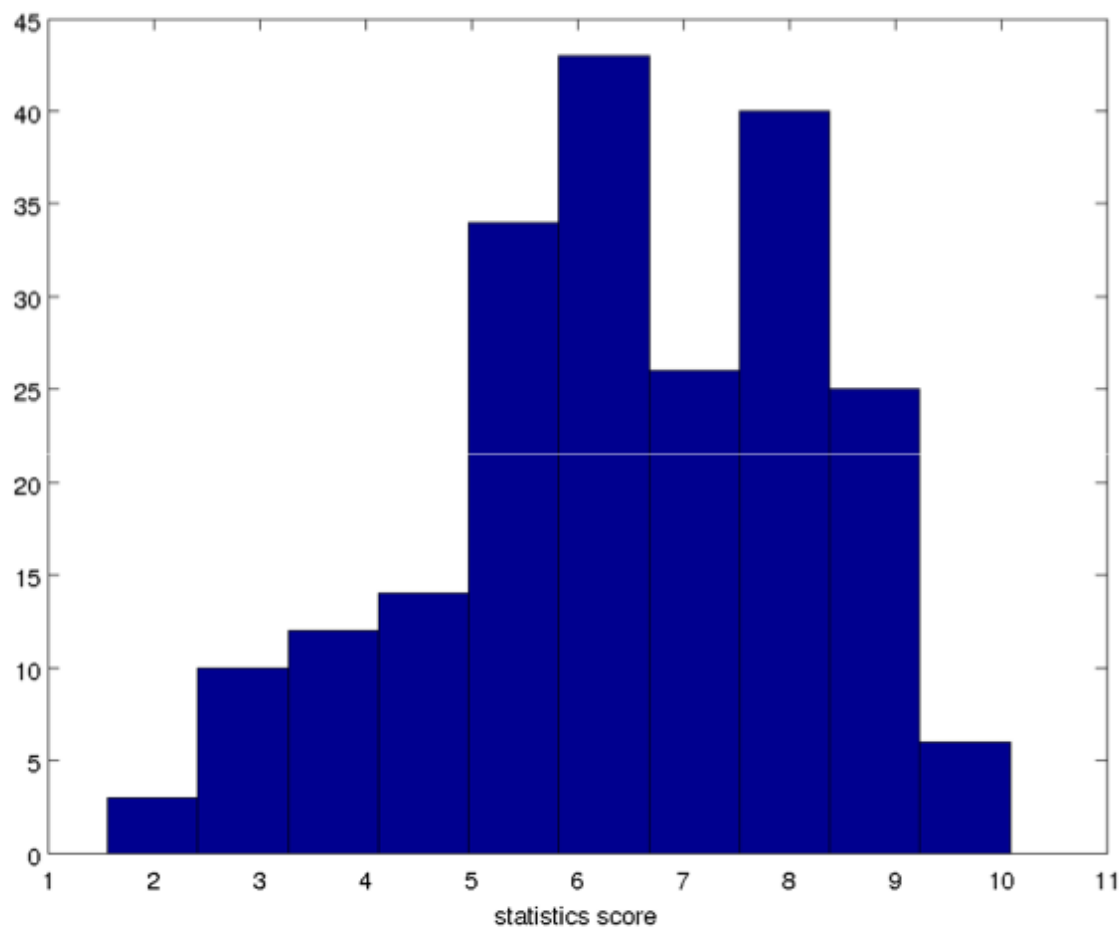


pdf

$$P(x|\theta) = P(x|\mu, \Sigma, \Pi) = \Pi_0 N_0 + \Pi_1 N_1$$

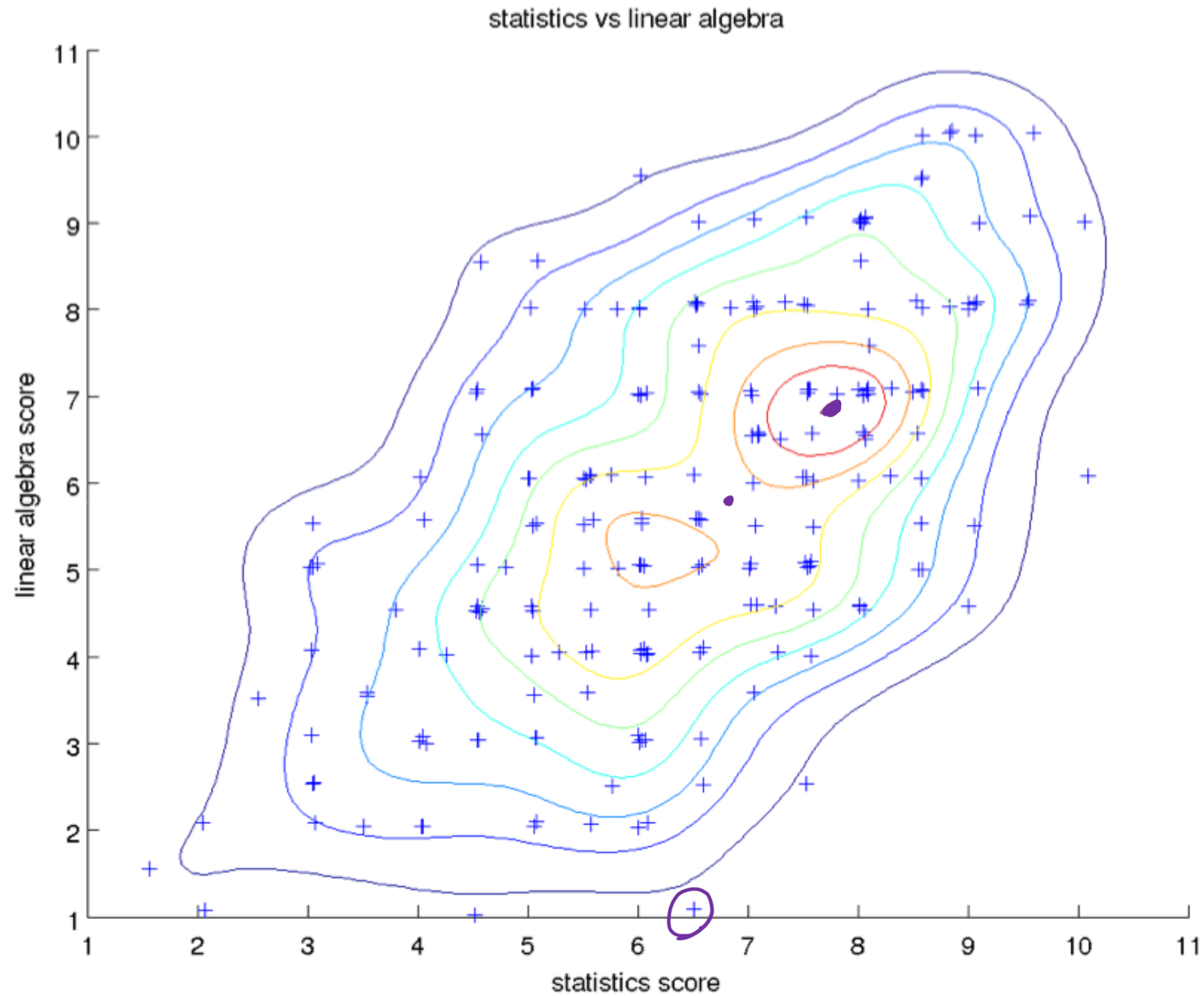
- Access the density of seeing a particular data point
 - Is this a typical data point? (high density value)
 - Is this an abnormal data point / outlier? (low density value)
- Building block for more sophisticated learning algorithms
 - Classification, regression, graphical models ...
 - A simple recommendation 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

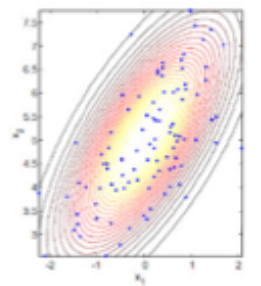


- Continuous case: eg. Gaussian distribution in R^d

Multivariate Gaussian $\rightarrow$

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

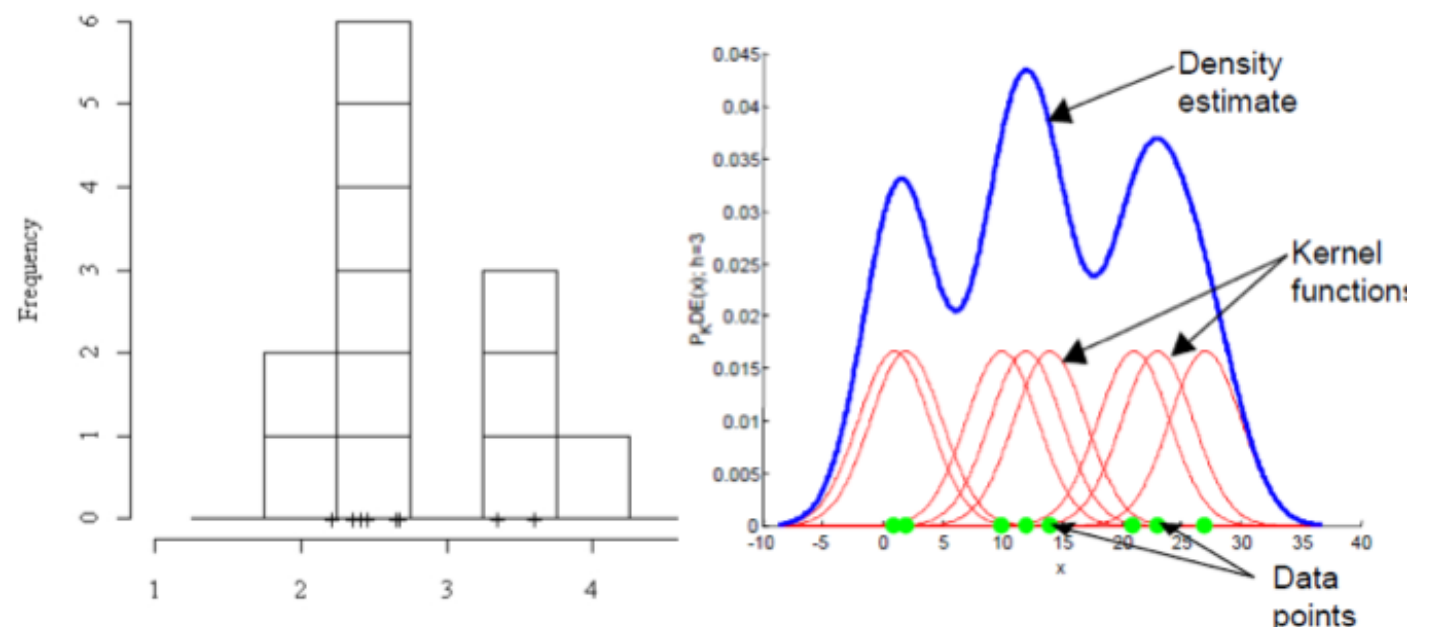
Annotations: Σ is circled with $1 \times d$ above and $d \times d$ below. Σ^{-1} is circled.



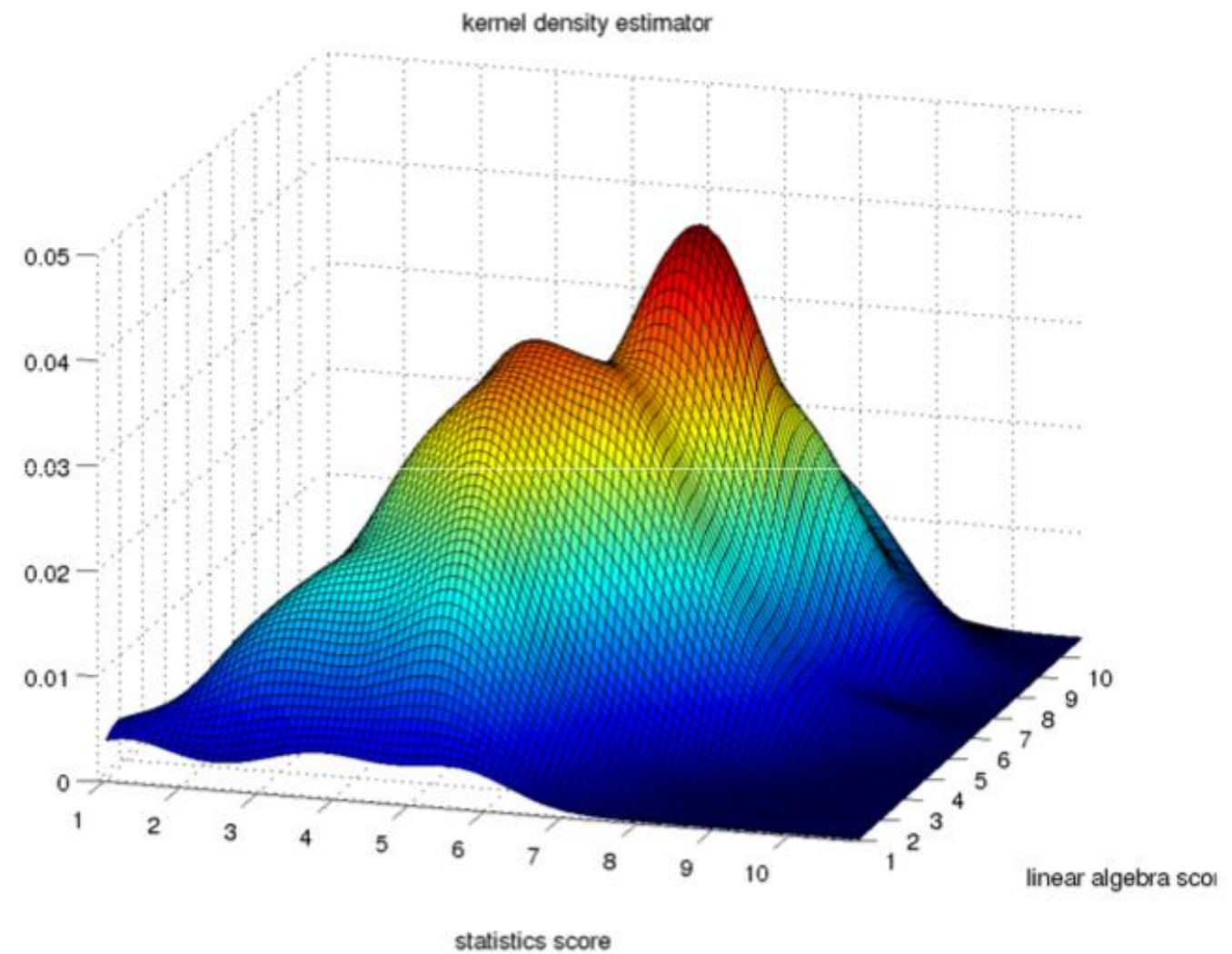
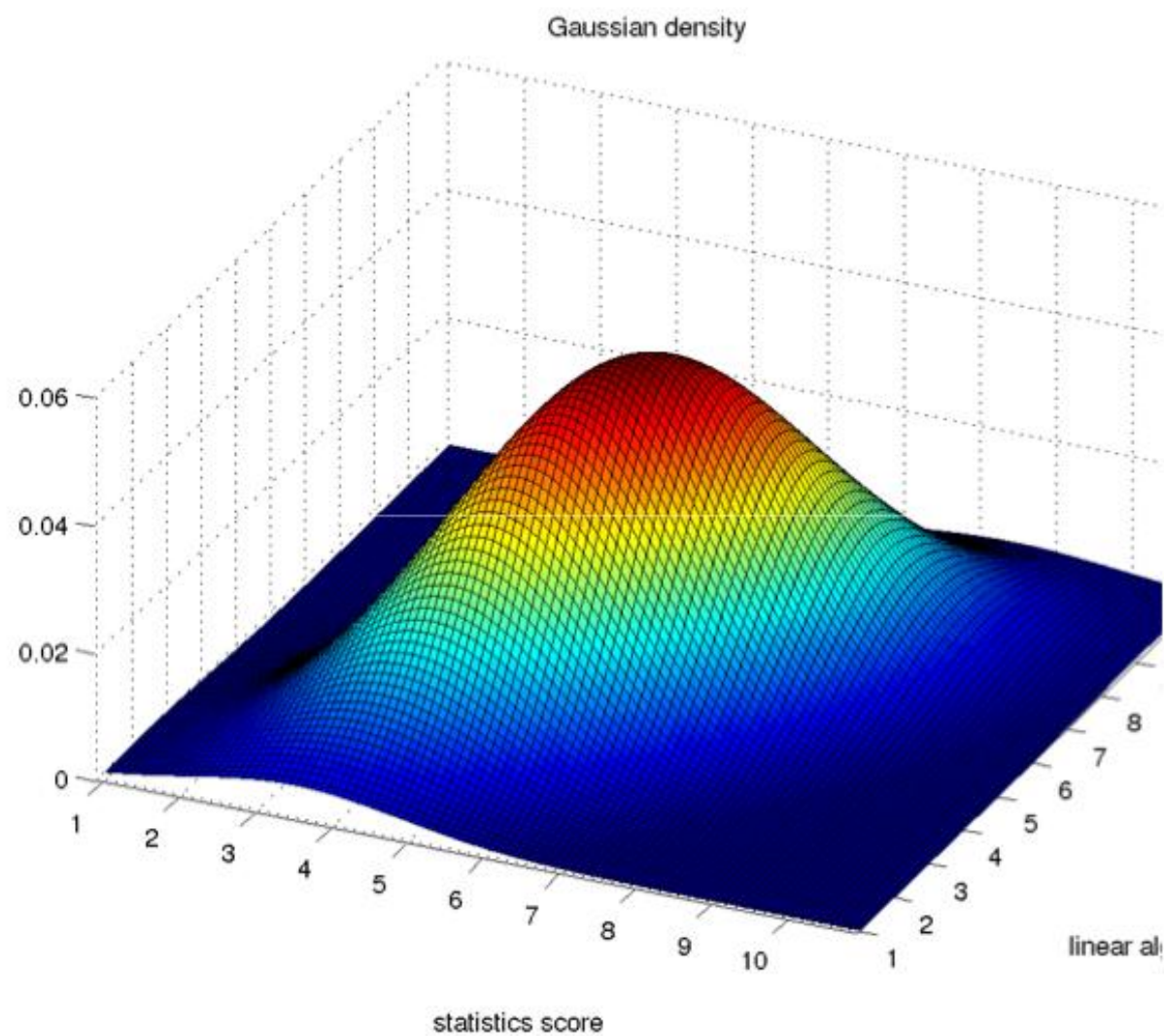
Two sets of parameters $\{\mu, \Sigma\}$, which again generate a family of models, $\mathcal{F} = \{p(x|\mu, \Sigma) \mid \mu \in R^d, \Sigma \in R^{d \times d} \text{ and PSD}\}$,

Nonparametric Density Estimation

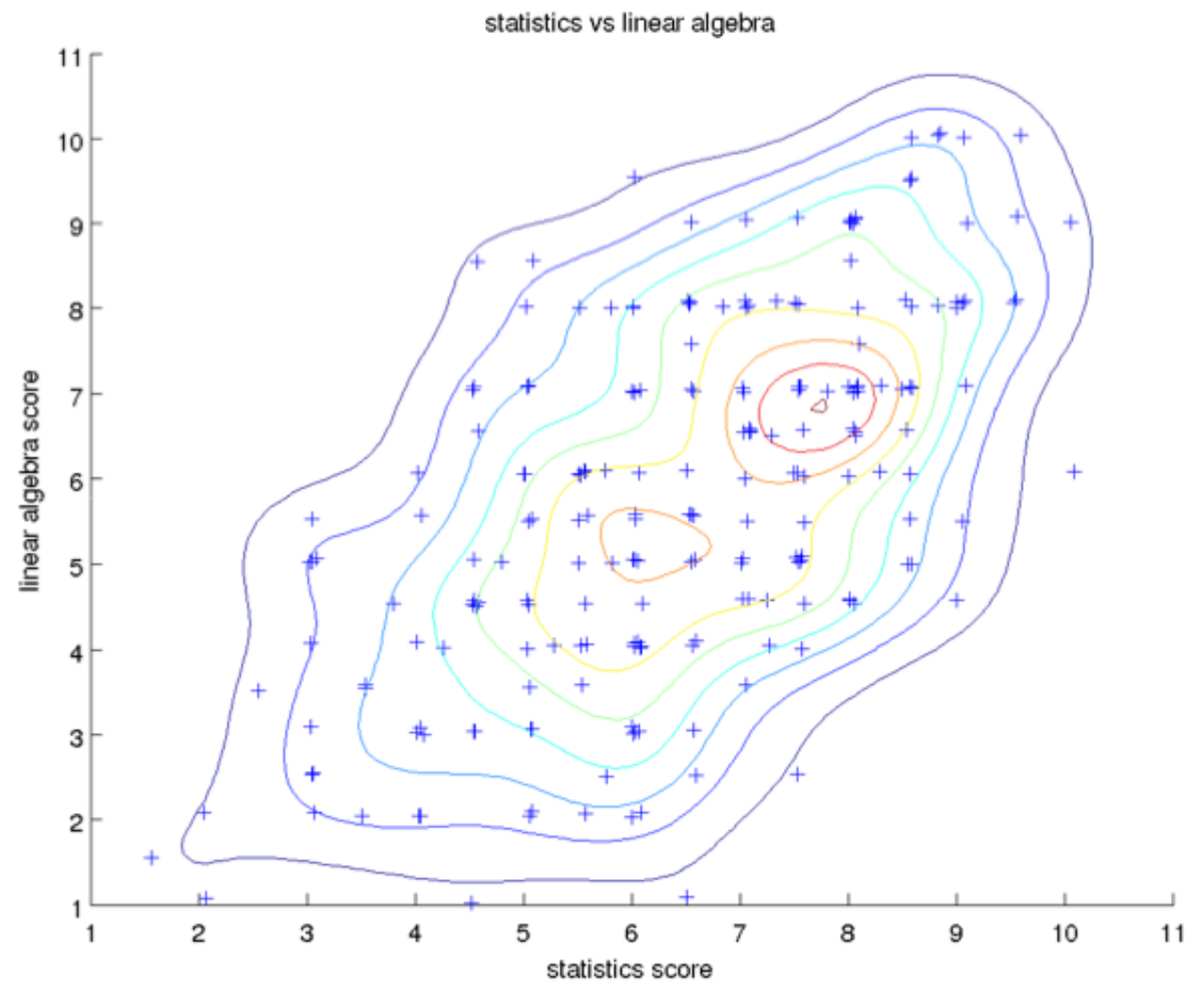
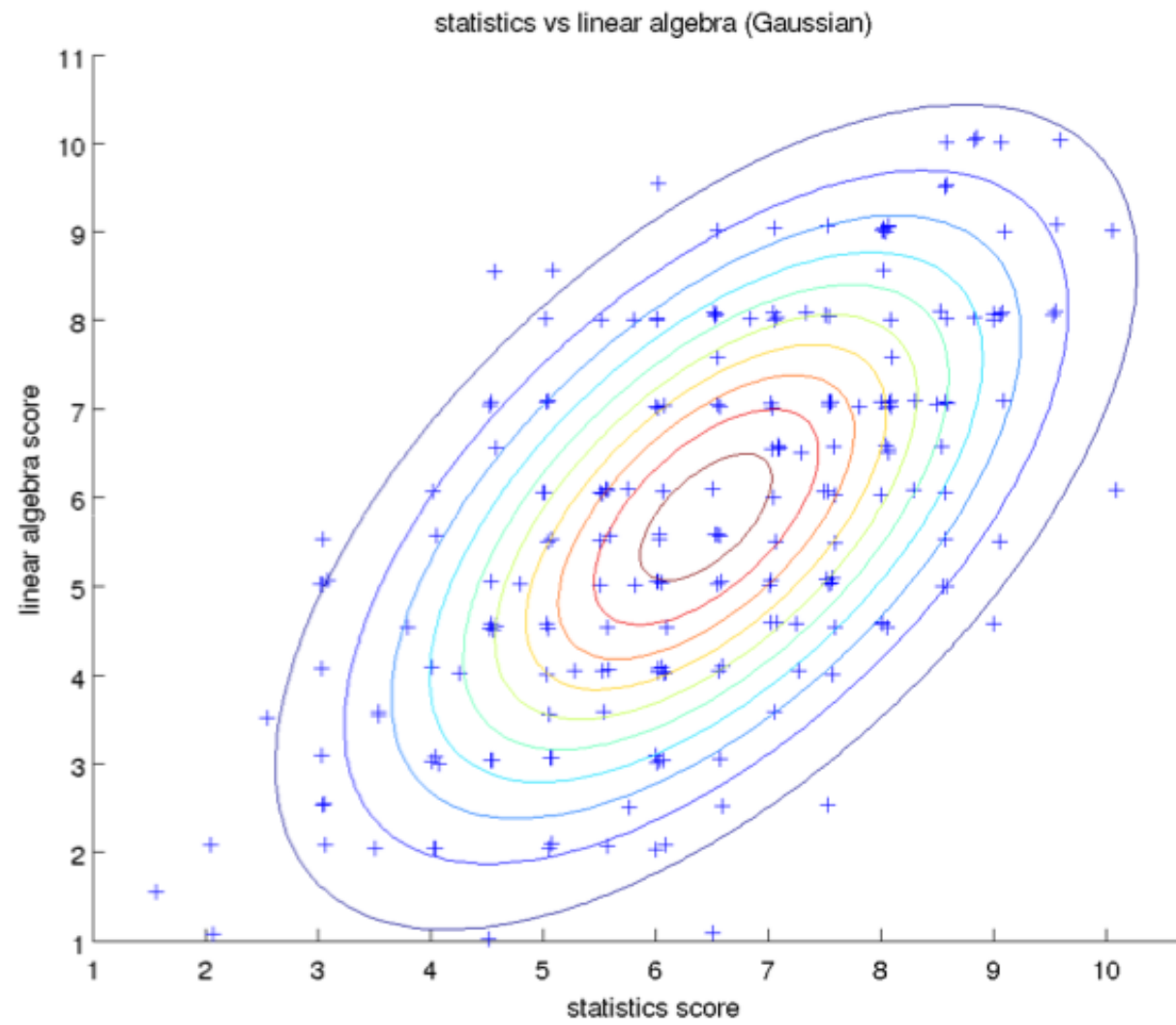
- What are nonparametric models?
 - “nonparametric” does **not** mean there are no parameters
 - can not be described by a fixed number of parameters
 - one can think of there are many parameters
- Eg. Histogram
- Eg. Kernel density estimator



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 n data points $X = \{x_1, x_2, \dots, x_n\}$ 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

$$\theta = \underset{\theta}{\operatorname{argmax}} \log P(X|\theta) = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^N P(x_i|\theta)$$

Handwritten notes: $l(\theta|x)$ is written above the first term, and the entire equation is enclosed in a purple box.

Example Problem

- Estimate the probability θ of landing in heads using a biased coin
- Given a sequence of n independently and identically distributed (iid) flips
 - Eg. $X = \{x_1, x_2, \dots, x_n\} = \{1, 0, 1, \dots, 0\}, x_i \in \{0, 1\}$
- Model: $P(x|\theta) = \theta^x(1 - \theta)^{1-x}$
 - $P(x|\theta) = \begin{cases} 1 - \theta, & \text{for } x = 0 \\ \theta, & \text{for } x = 1 \end{cases}$
- Likelihood of a single observation x_i ?
 - $L(\theta|x_i) = p(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$



MLE for Biased Coin

- Objective function, log-likelihood

$$l(\theta|\mathbf{X}) = \log L(\theta|\mathbf{X}) = \log \prod_{i=1}^N \theta^{x_i} (1 - \theta)^{1-x_i} = \log(\theta^{N_H} (1 - \theta)^{N_T})$$
$$= N_H \times \log \theta + N_T \times \log(1 - \theta)$$

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: $N_H = 78, N_T = 22 \rightarrow \theta = 0.78$

Estimating Gaussian Distributions

- Gaussian distribution in R

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

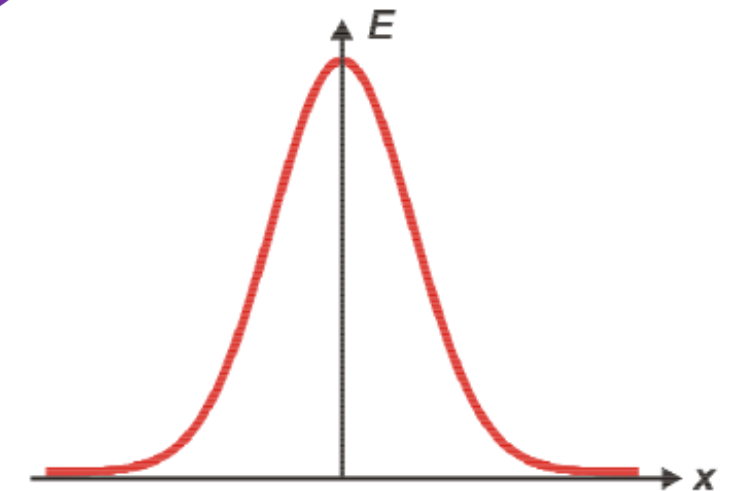
$$L(\theta|x) = \prod_{i=1}^N P(x_i|\mu, \sigma) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)$$

- Need to estimate two sets of parameters μ, σ

$$\log L(\theta|x) = \ell(\theta|x) = \log \sum_{i=1}^N (\cdot)$$

- Given n iid samples

$$X = \{x_1, x_2, \dots, x_n\}, x_i \in R$$



- 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

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

MLE for Gaussian Distribution

- Objective function, log likelihood

$$l(\theta|X) = l(\mu, \sigma|X) = \log \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{1}{2}} \sigma} \exp \left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

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

Handwritten notes: $\log(\exp(u)) = u$, $b = (2\pi)^{\frac{1}{2}} \sigma$

- Maximize $l(\mu, \sigma; X)$ with respect to μ, σ

- Take derivatives w.r.t. μ, σ^2

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

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

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

$$-\log b = -\log (2\pi\sigma^2)^{\frac{1}{2}} = -\frac{1}{2} \log (2\pi\sigma^2)$$

$$= n \left(-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 \right)$$

MLE for Gaussian Distribution

$$l(\mu, \sigma | X) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

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

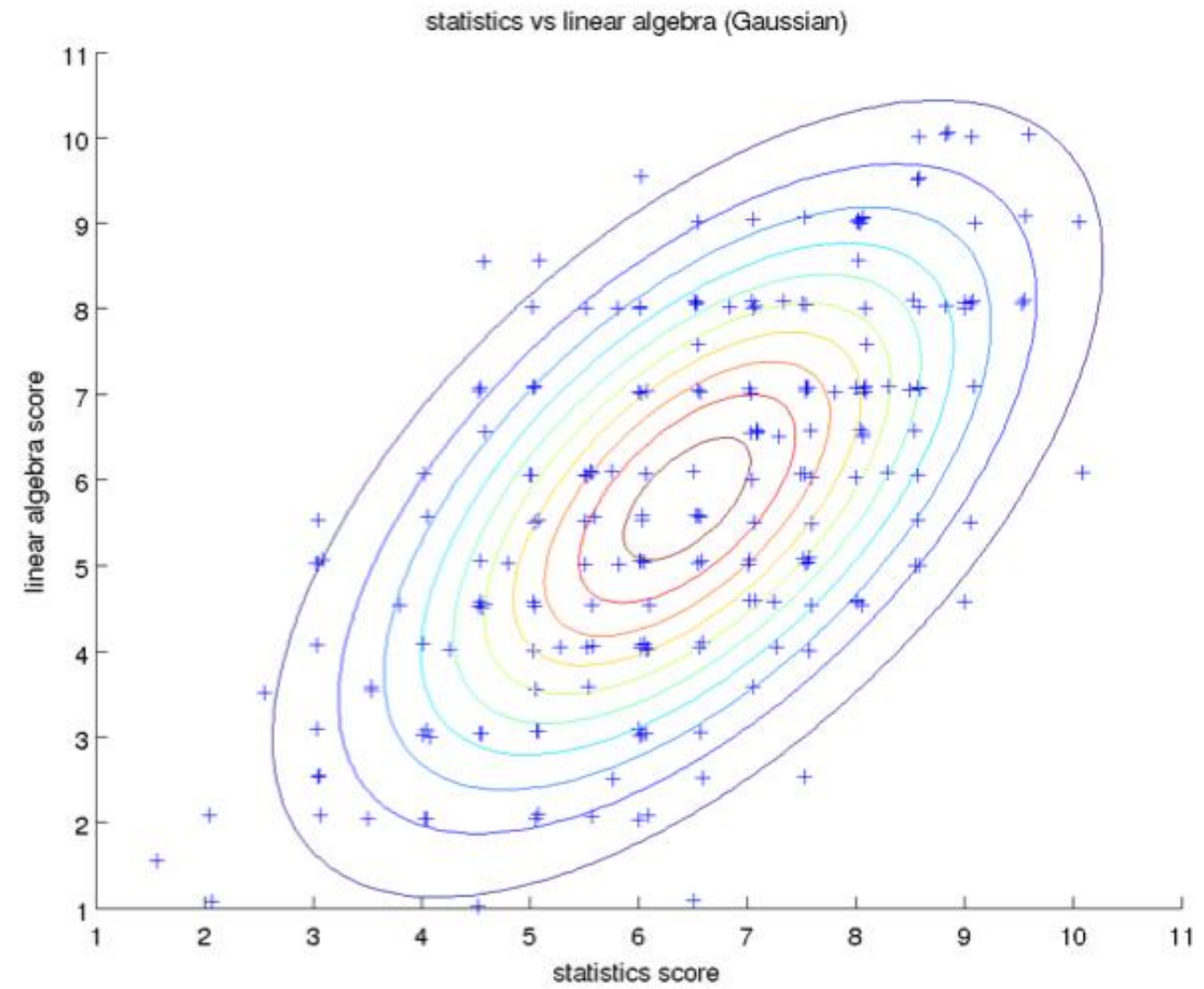
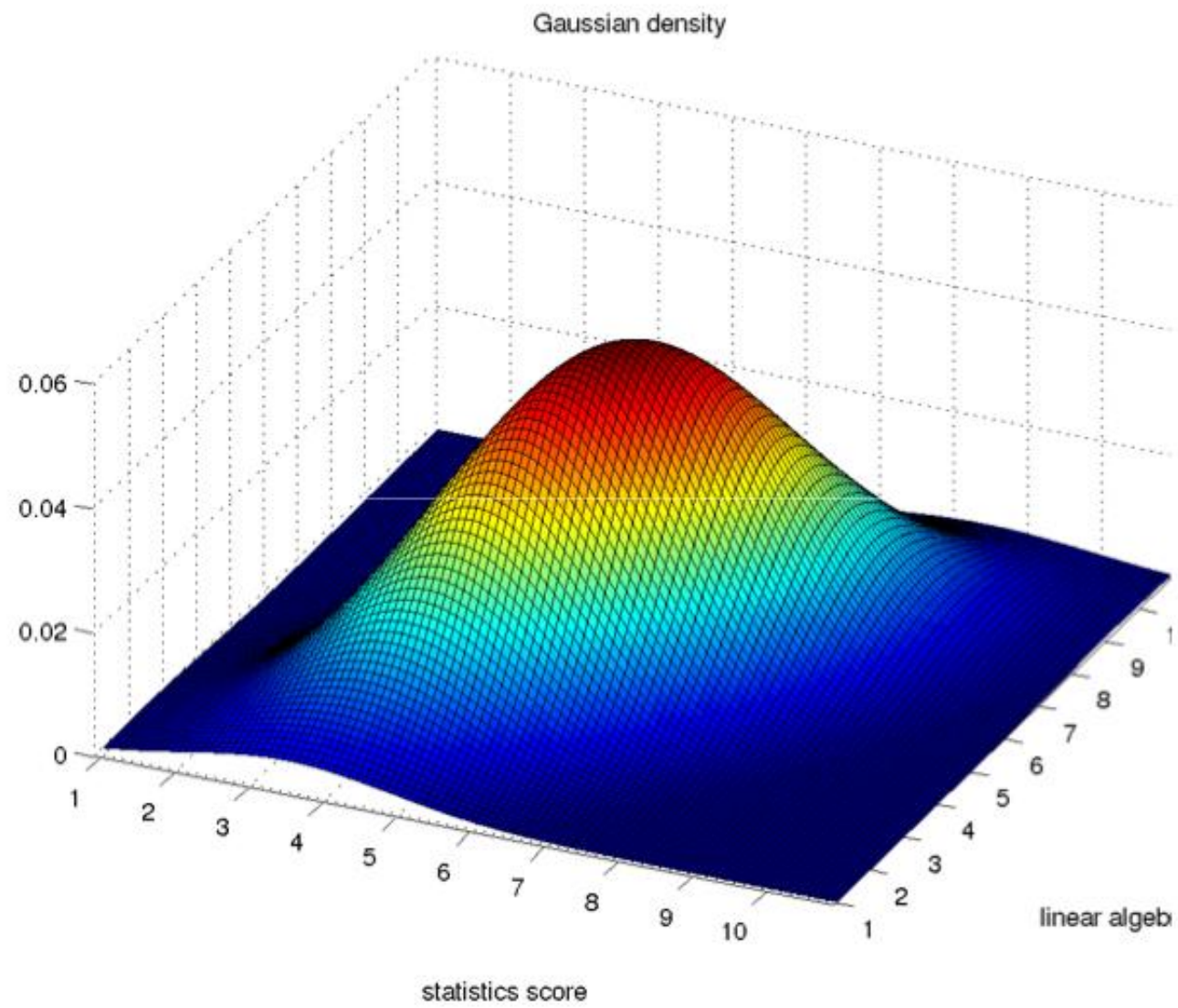
$$\Rightarrow \sum_i x_i = n \mu \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^N x_i$$

$$\frac{\partial (\log(x))}{\partial x} = \frac{1}{x}$$

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

$$\Rightarrow \sum_i (x_i - \mu)^2 = n \sigma^2 \Rightarrow \frac{1}{n} \sum_{i=1}^N (x_i - \mu)^2 = \sigma^2$$

Example



Outline

- Overview
- Parametric Density Estimation $\rightsquigarrow$ Gaussian
- Nonparametric Density Estimation $\leftarrow$

$$p(x | \overset{1 \times d}{\mu}, \overset{d \times d}{\Sigma})$$

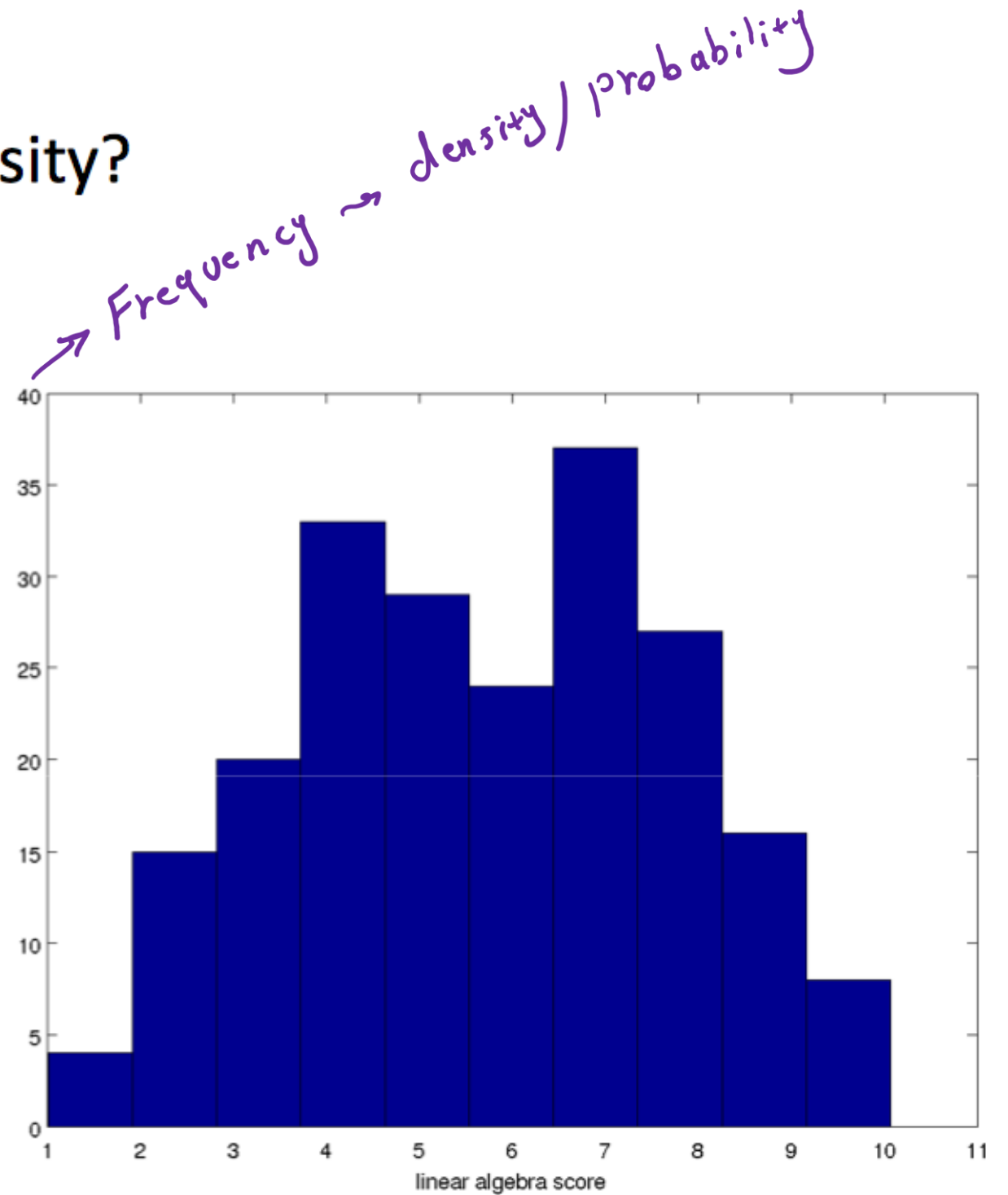
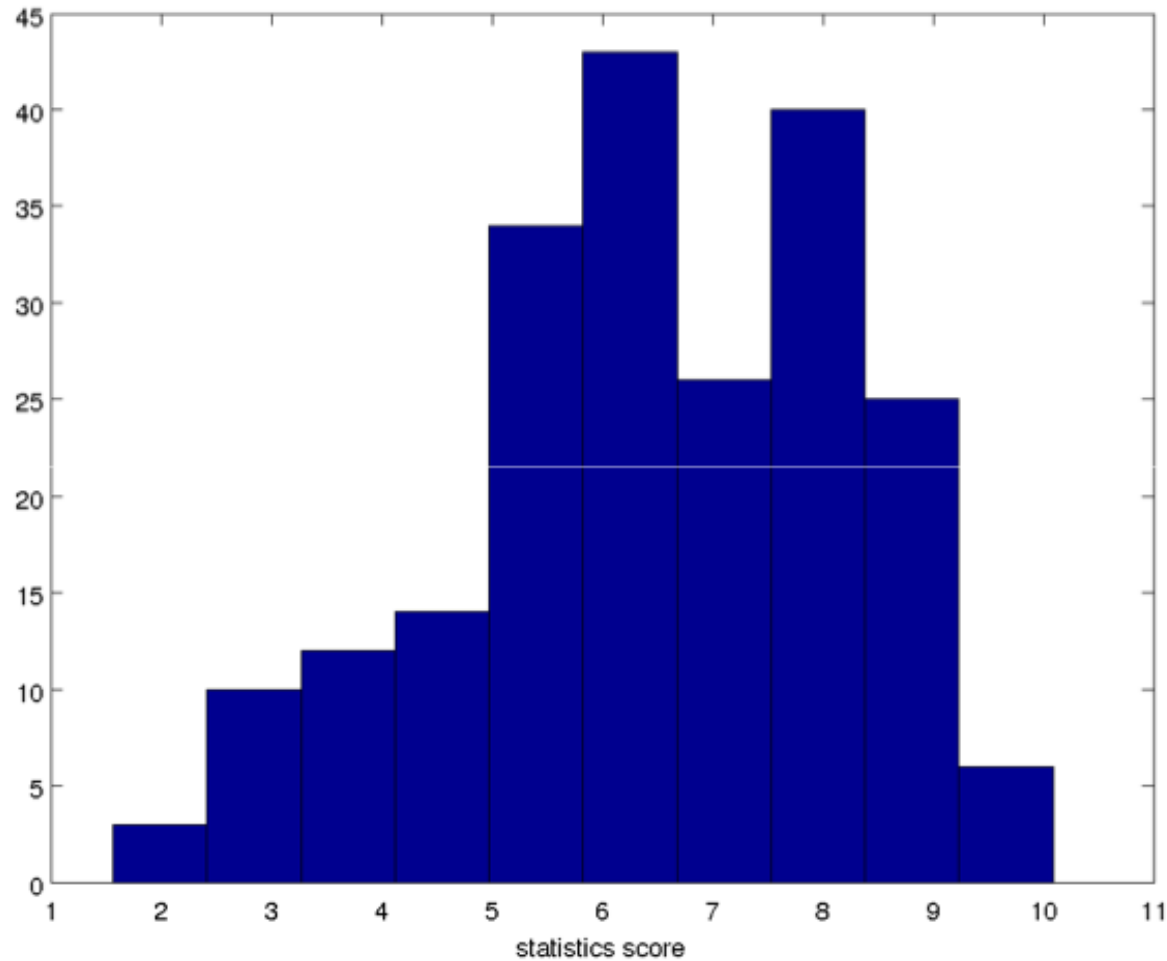
$\uparrow$
MLE

Can be used for:

- Visualization
- Classification $\rightsquigarrow$ NB
- Regression

Example: Test Scores

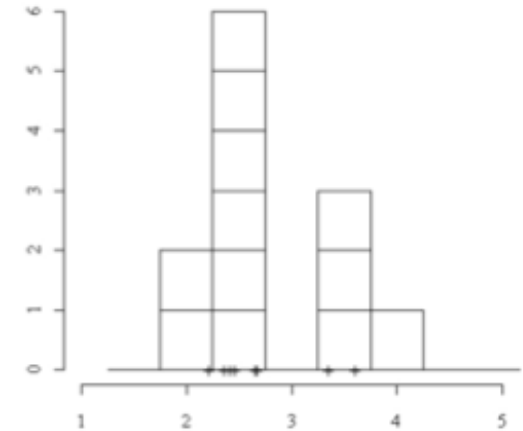
- What is missing if we want density?



1-D Histogram

- One the simplest nonparametric density estimator

- Given N iid samples $X = \{x_{\{1\}}, x_{\{2\}}, \dots, x_{\{N\}}\} = x_{\{i\}} \in [0, 1)$



- Split $[0, 1)$ into M bins

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \dots, B_l = \left[\frac{l-1}{M}, \frac{l}{M}\right), \dots, B_M = \left[\frac{M-1}{M}, 1\right)$$

$\underbrace{\quad\quad\quad}_{\frac{1}{M}} \quad \underbrace{\quad\quad\quad}_{\frac{1}{M}} \quad \underbrace{\quad\quad\quad}_{\frac{1}{M}}$

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

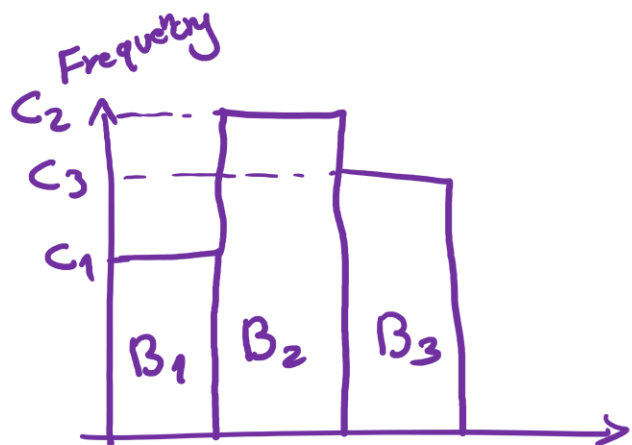
- For a new test data point x which belongs to B_l

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

$$p(x) = \frac{M}{N} \sum_{i=1}^N 1(x_i \in B_l) = \frac{\text{number of points in bin } B_l (c_l)}{\text{total number of data points} \times \text{bin width}}$$

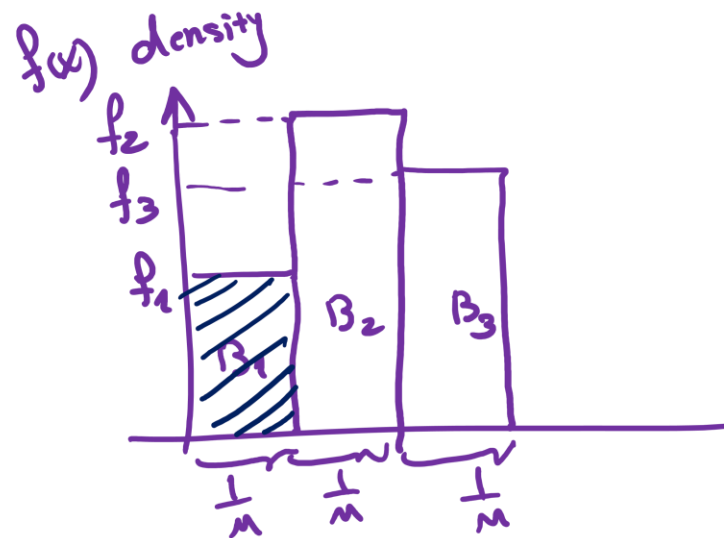
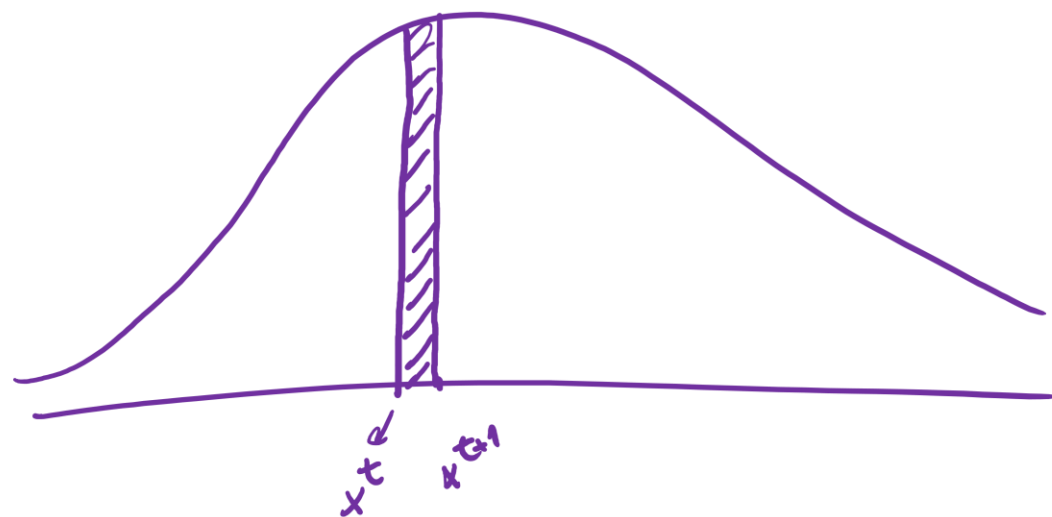
$\frac{1}{M}$

$$\int p(x) dx \quad ?$$



$$P(B_1) = \frac{C_1}{C_1 + C_2 + C_3} = \frac{C_1}{N}$$

$$P(B_2) = \frac{C_2}{N}$$



$$P(B_1) = f_1 * \frac{1}{M}$$

$$\frac{C_1}{N} = f_1 * \frac{1}{M} \Rightarrow f_1 = \frac{C_1}{N * \frac{1}{M}}$$

Why is Histogram Valid?

- Requirement for density $p(x)$

- $p(x) \geq 0, \int_{\Omega} p(x) dx = 1$

- For histogram, $\int_{[0,1)} p(x) dx = \int_0^1 \frac{M}{N} \sum_{i=1}^N 1(x_i \in B_l) dx$

$$= \int_0^{\frac{1}{M}} \frac{M}{N} \sum_{i=1}^N 1(x_i \in B_l) dx + \int_{\frac{1}{M}}^{\frac{2}{M}} \frac{M}{N} \sum_{i=1}^N 1(x_i \in B_l) dx + \dots + \int_{\frac{l-1}{M}}^{\frac{l}{M}} \frac{M}{N} \sum_{i=1}^N 1(x_i \in B_l) dx =$$

$$= \frac{M}{N} \left[\int_0^{\frac{1}{M}} c_1 dx + \int_{\frac{1}{M}}^{\frac{2}{M}} c_2 dx + \dots + \int_{\frac{l-1}{M}}^{\frac{l}{M}} c_l dx + \dots + \int_{\frac{M-1}{M}}^1 c_M dx \right] =$$

$$= \frac{M}{N} \sum_{l=1}^M \int_{\frac{l-1}{M}}^{\frac{l}{M}} c_l dx = \frac{M}{N} \sum_{j=1}^M c_l \left[\frac{l}{M} - \frac{l-1}{M} \right] = \sum_{l=1}^M \frac{c_l}{N} = 1$$

$\uparrow = \frac{c_1 + c_2 + \dots + c_M}{N}$

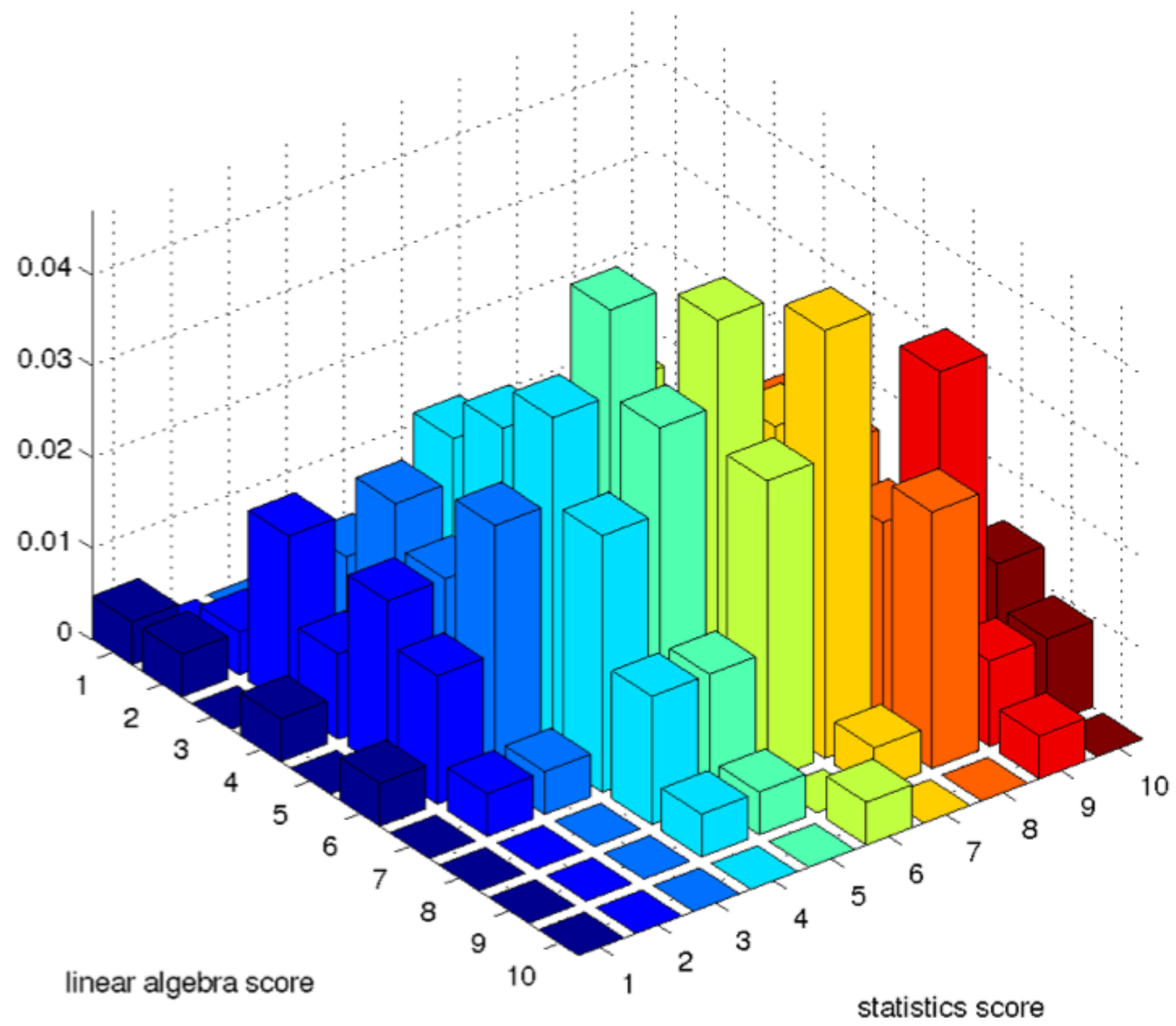
Higher-Dimensional Histogram

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

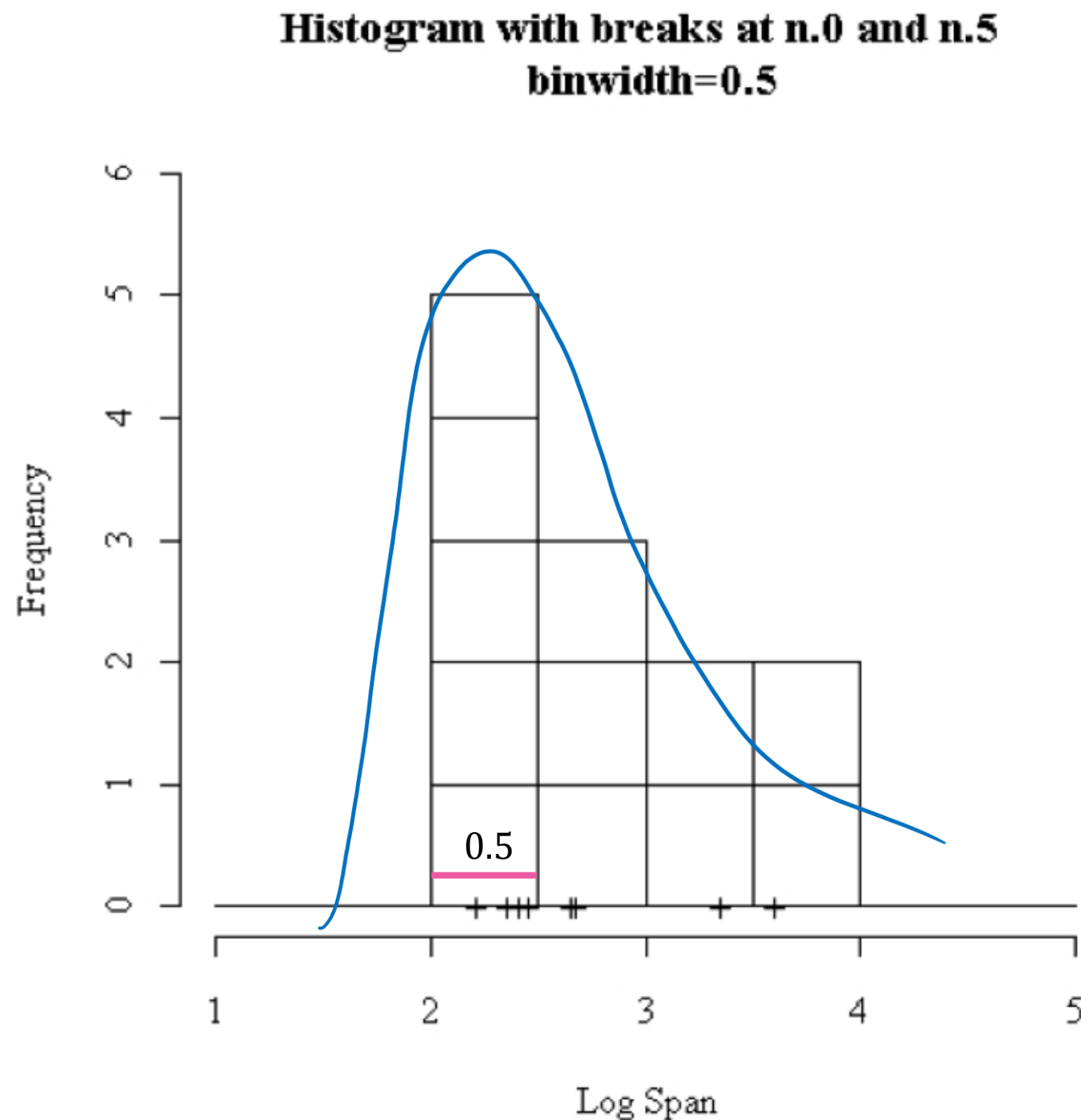
Two Dimensional data:

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

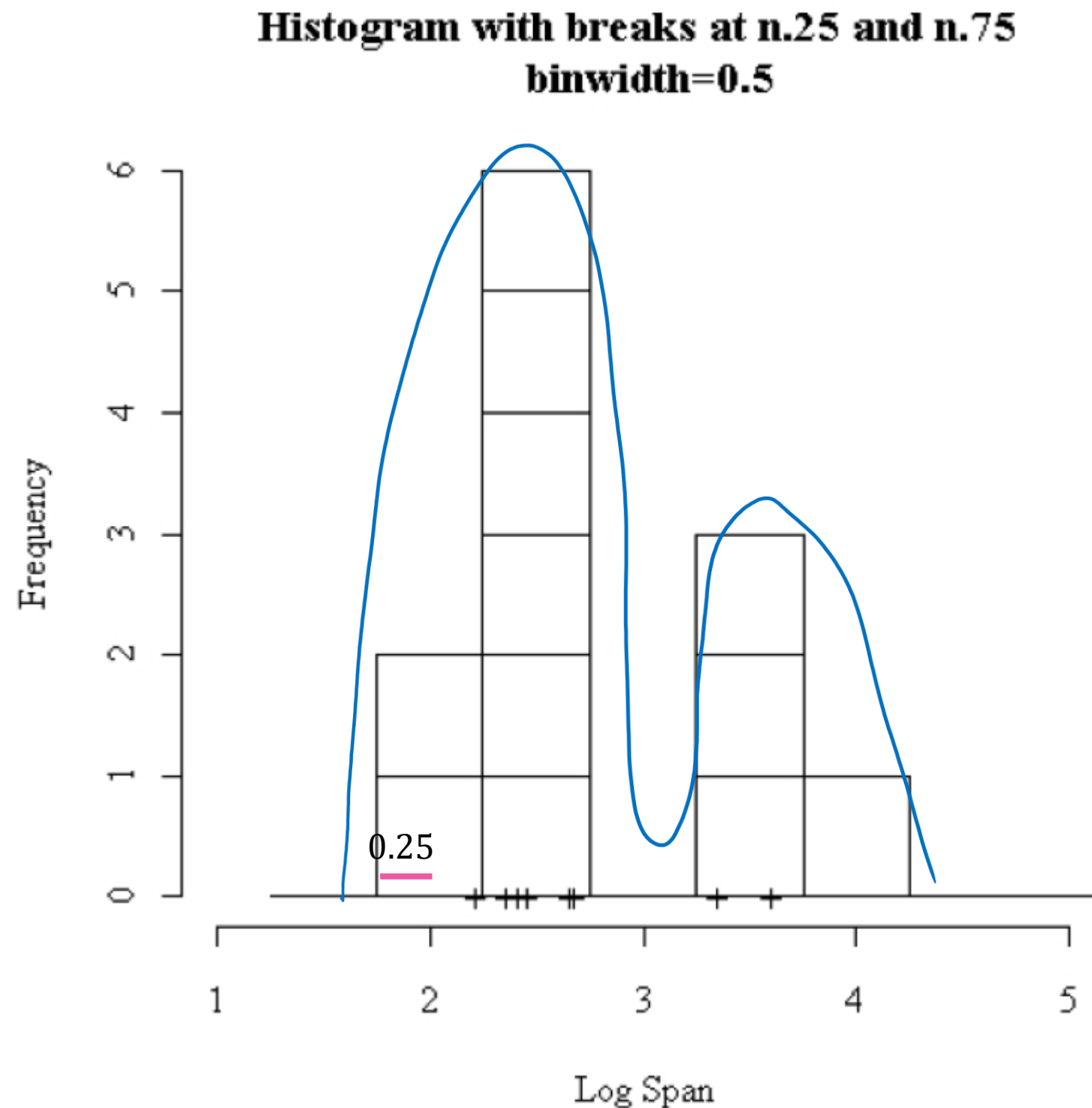
$M^2 = 100$ (total number of bins for two dimensional data)



Output Depends on Where You Put the Bins



Output Depends on Where You Put the Bins



Kernel Density Estimation KDE

- Kernel density estimator

$$p(x) = \frac{1}{N} \sum_i^N \frac{1}{h} K\left(\frac{x_l - x_i}{h}\right)$$

$K(u)$

$\int p(x) dx = 1 \leftarrow \text{normalization}$

$$x_l = x_{\text{gridline}}$$

- Smoothing kernel function

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

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x_l - x_i)^2}{h^2}\right]$$

$u = \frac{x_l - x_i}{h}$

behaves very similar to μ mean

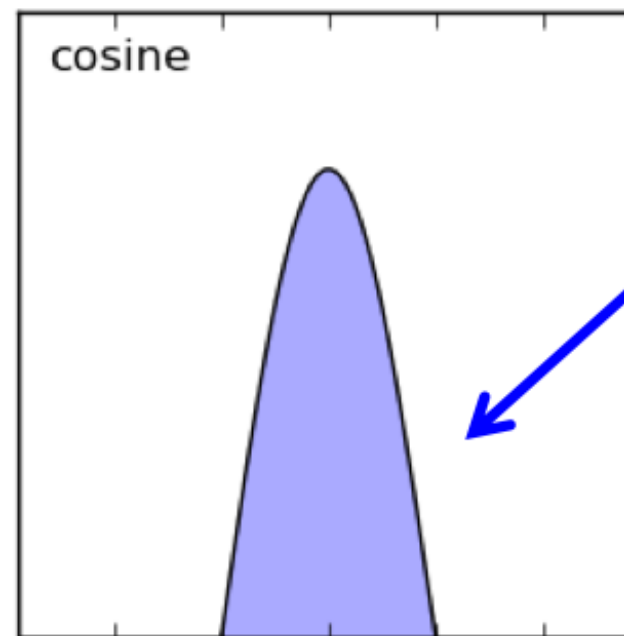
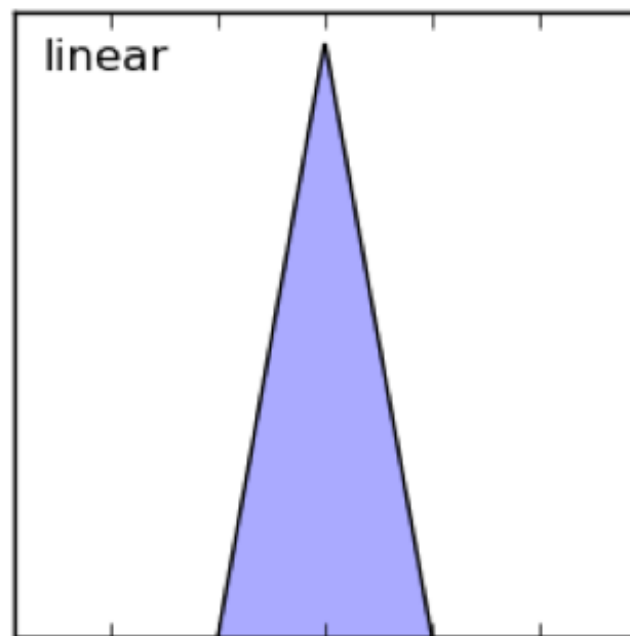
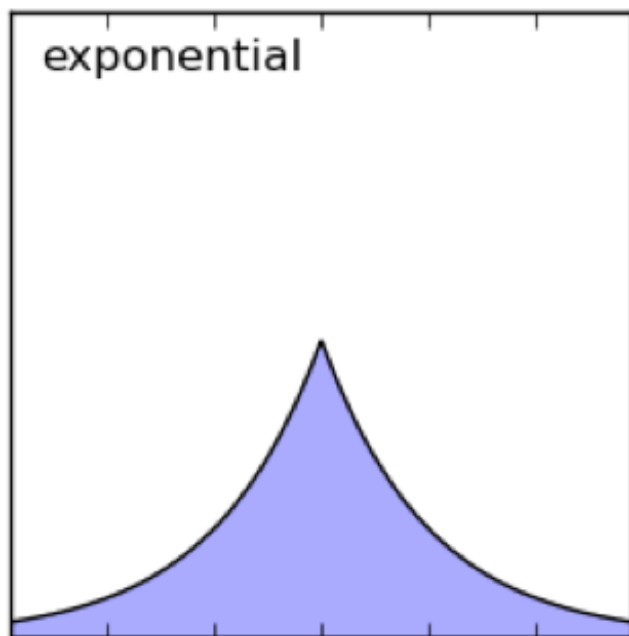
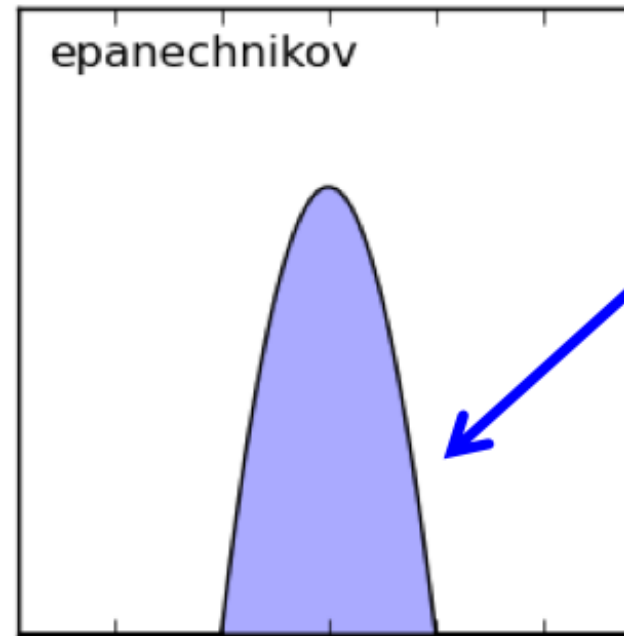
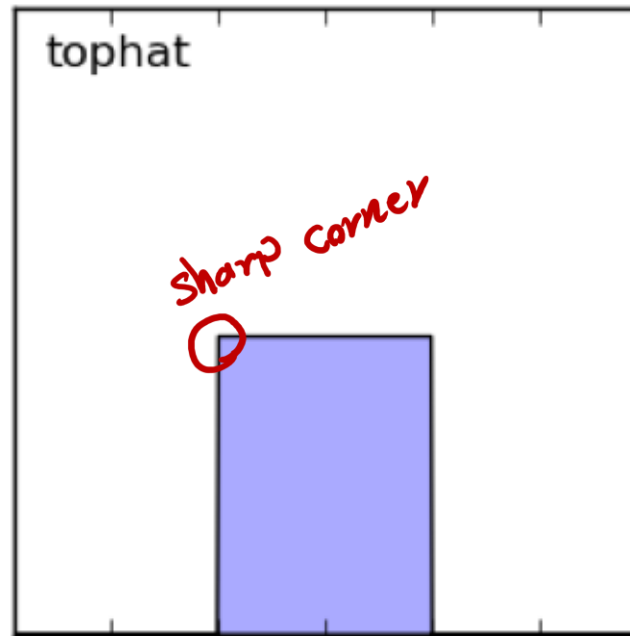
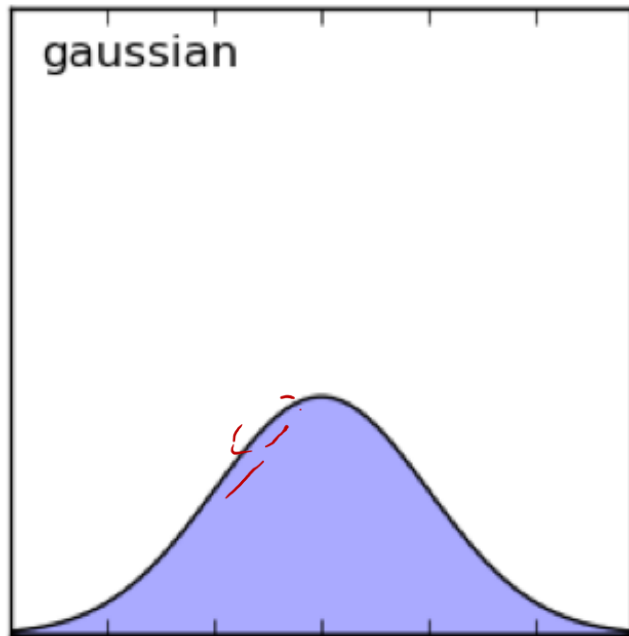
behaves similar to Variance

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

Smoothing Kernel Functions

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

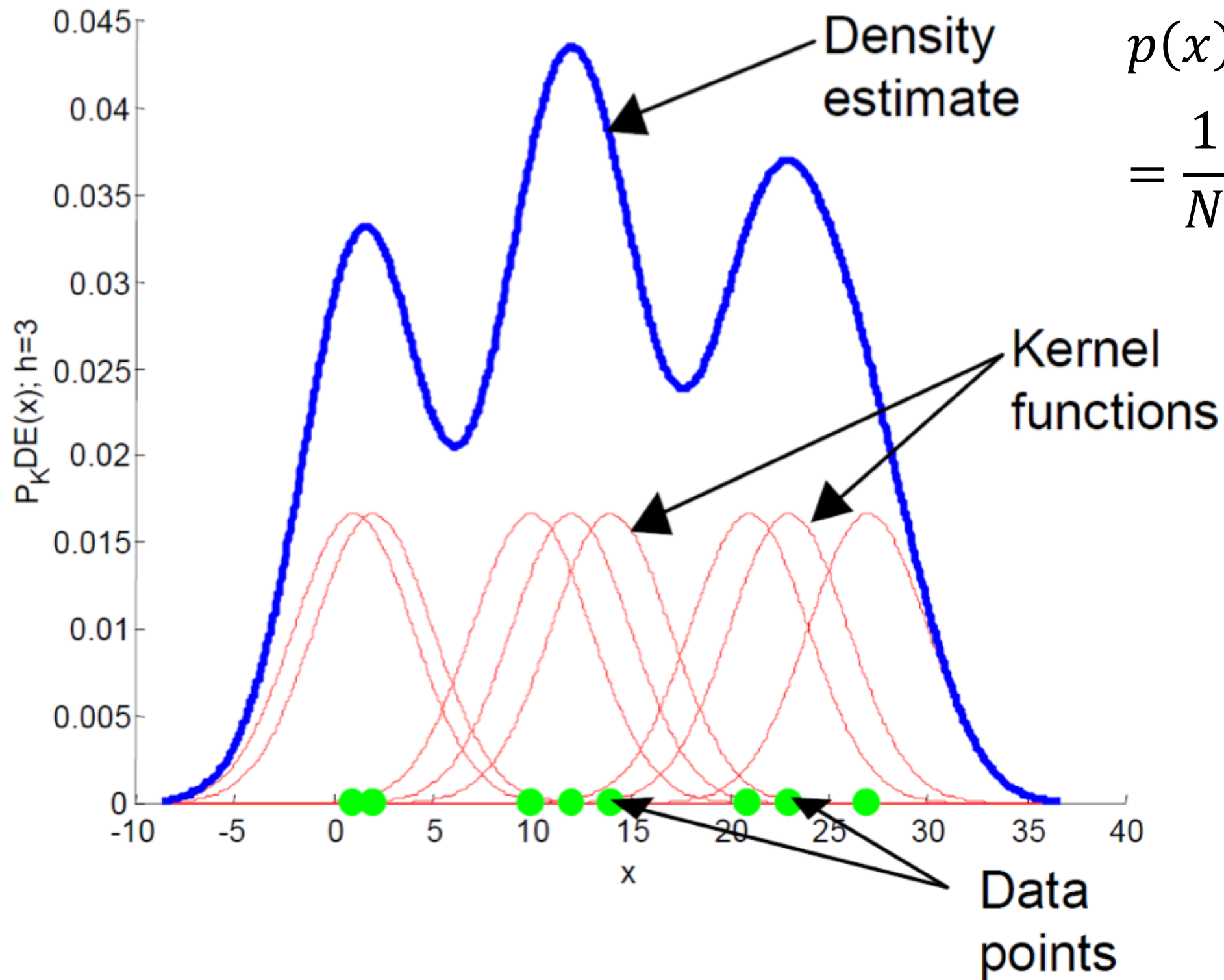
Available Kernels



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

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

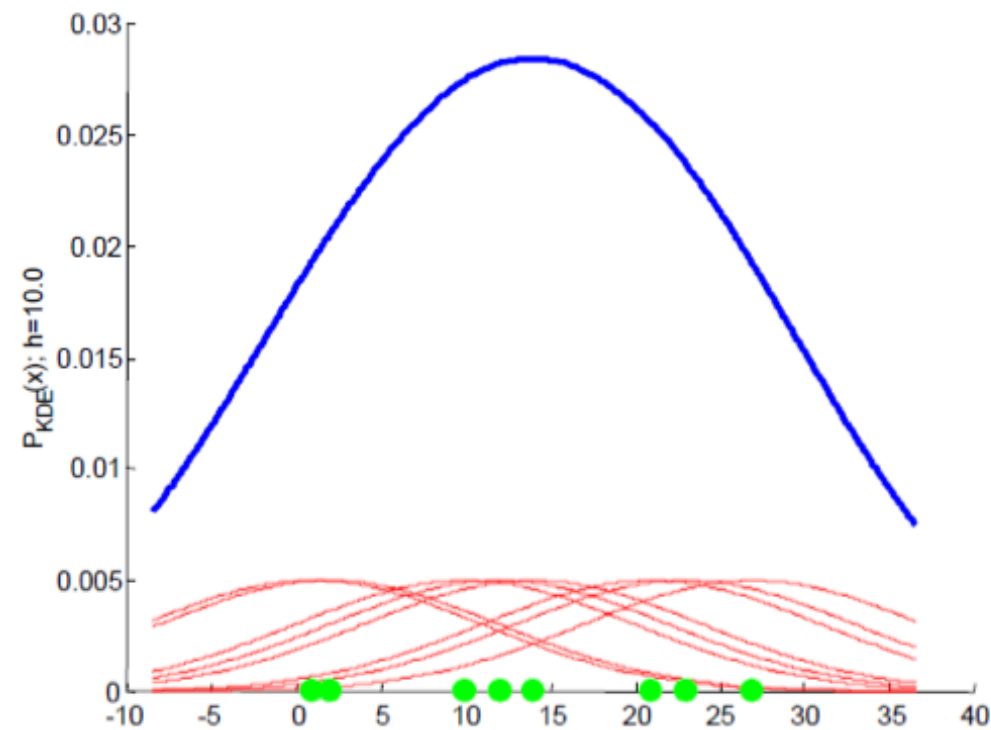
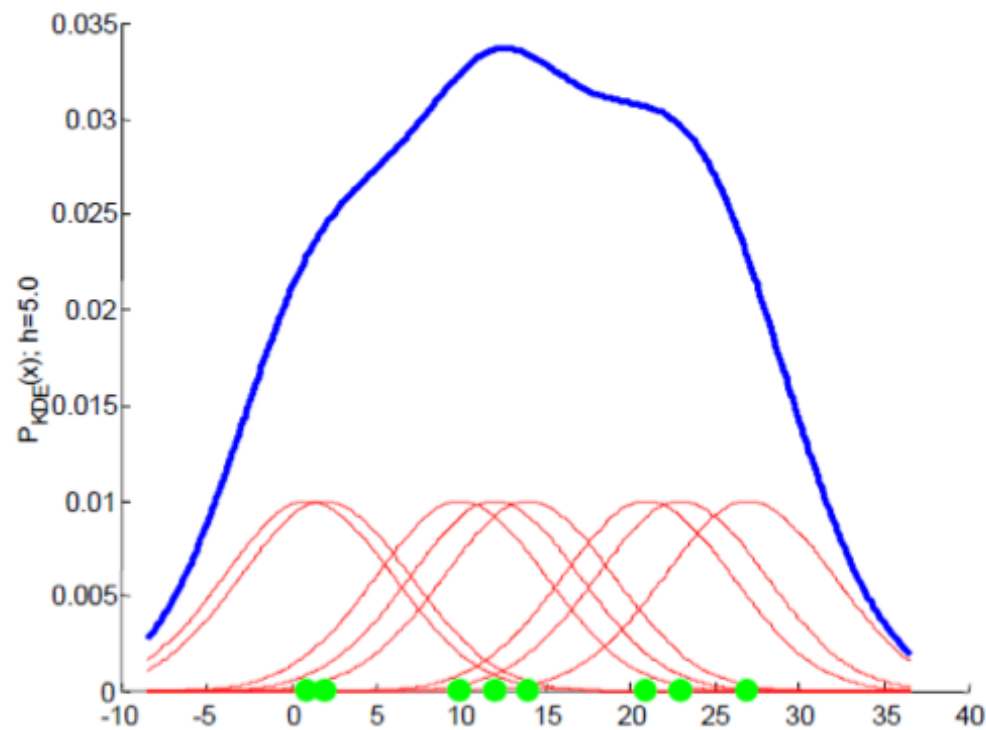
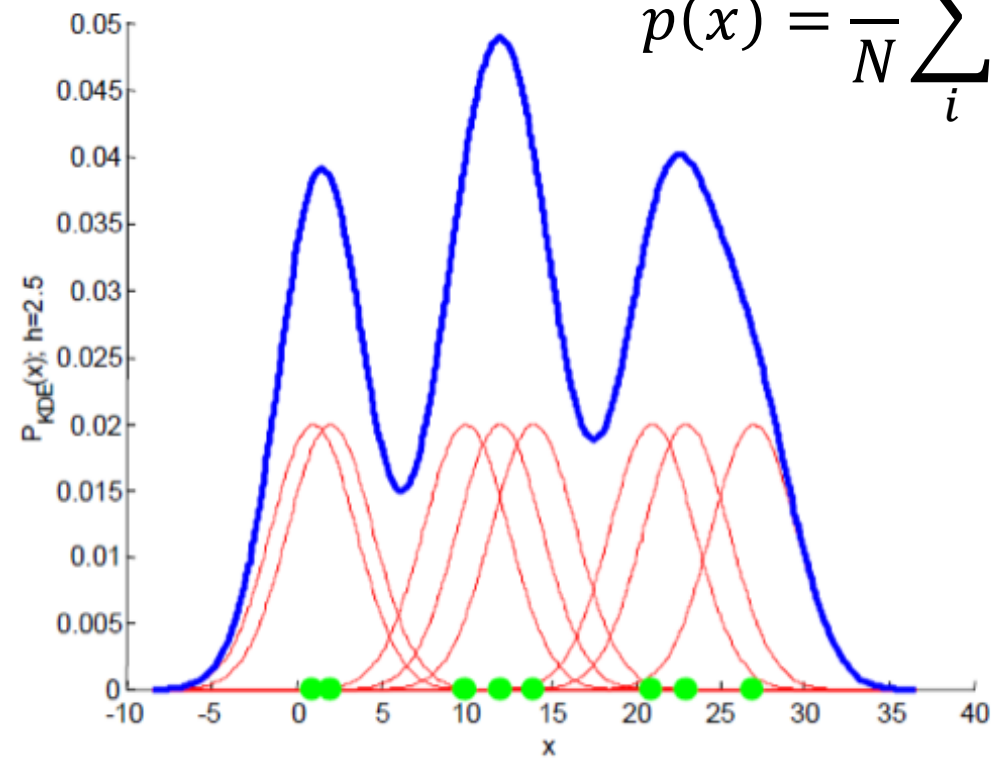
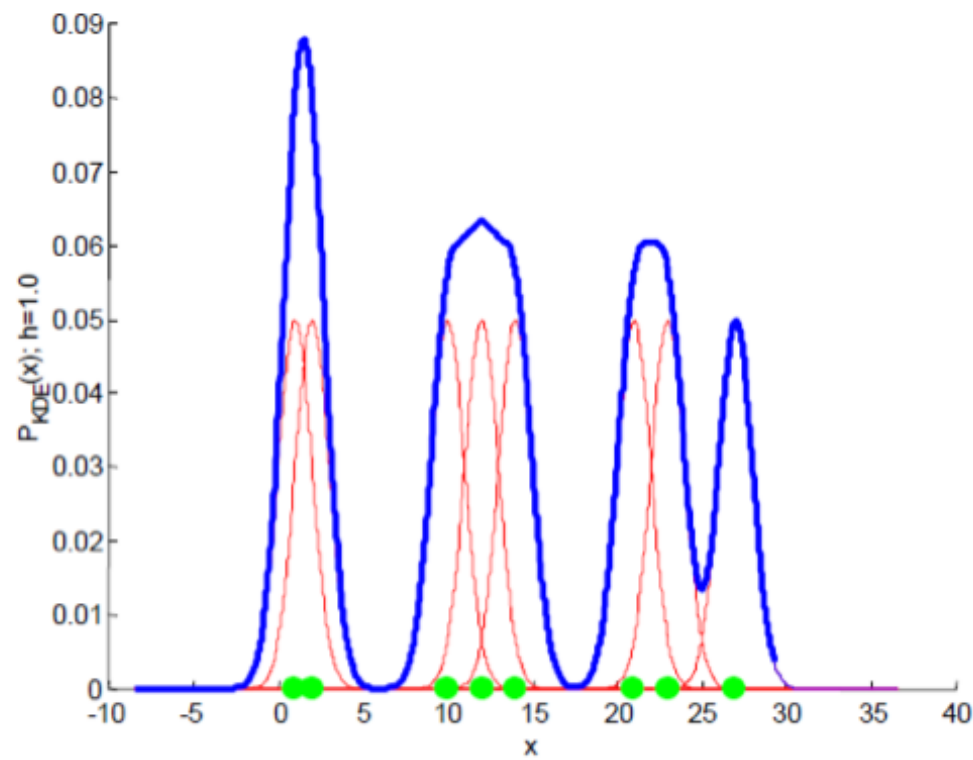
Example



$$p(x) = \frac{1}{N} \sum_i^N \frac{1}{h} K\left(\frac{x_l - x_i}{h}\right)$$

Effect of the Kernel Bandwidth (h)

$$p(x) = \frac{1}{N} \sum_i^N \frac{1}{h} K\left(\frac{x_l - x_i}{h}\right)$$



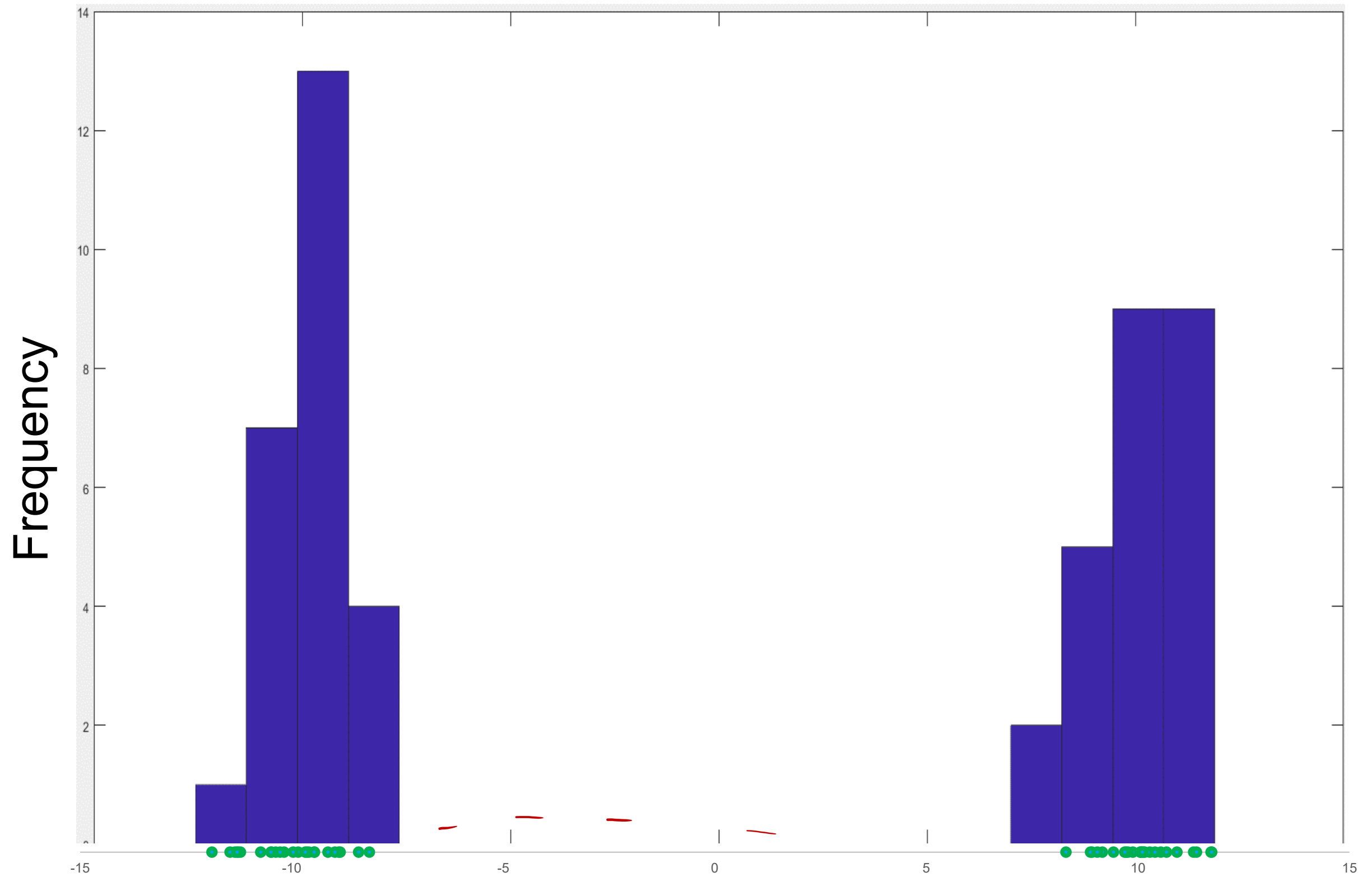
Visual Example

50 datapoints are given to us



Visual Example

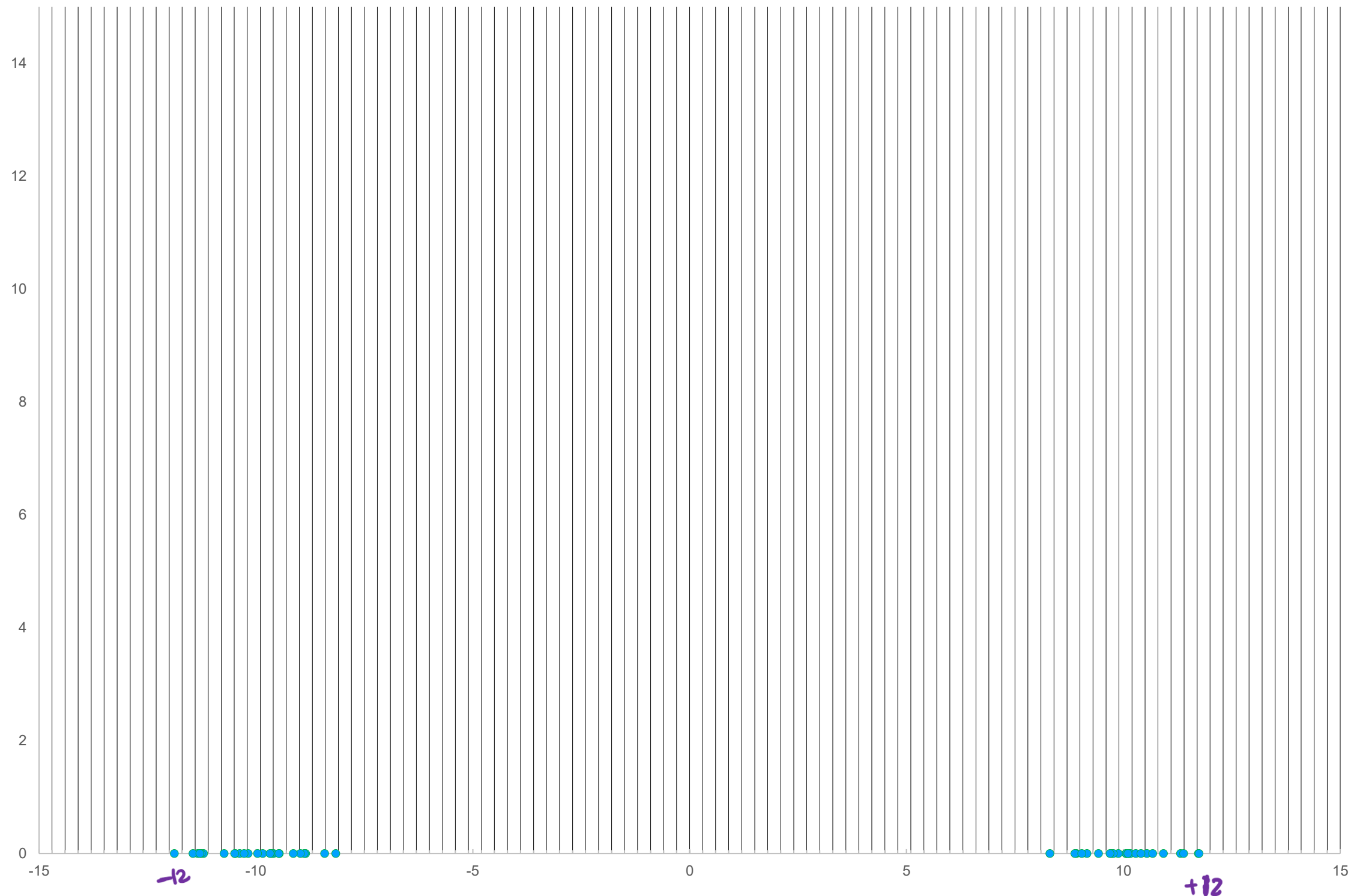
Let's implement 20 bins histogram



Visual Example

np.linspace(min, max, number)
(-15, 15, 200)

Let's create 200 uniform gridlines (x_l) to have a smoother density function
OR simply you can just implement this on each datapoint



For **each** linearly spaced gridline x_l , let's calculate the Gaussian kernel value over the given 50 points

$$p(x) = \underbrace{\frac{1}{N} \sum_i^N}_{\text{avg}} \underbrace{\frac{1}{h} K(u_i)}_{\text{kernel}}$$

$u_1, u_2, \dots, u_{50}$
 $k(u_1), k(u_2) \dots k(u_{50})$

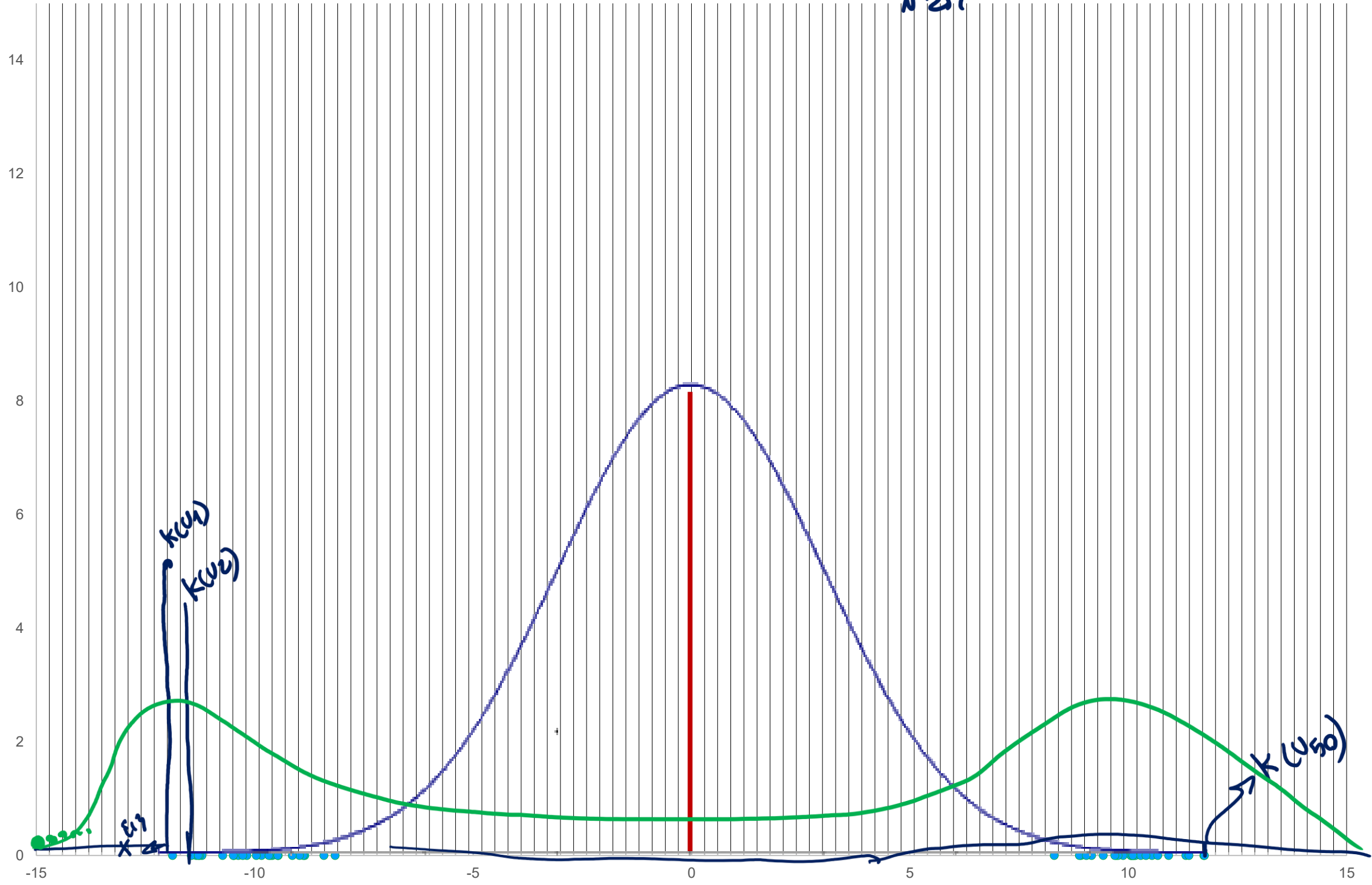
$$u_1 = \frac{x_l - x}{h=1} = -15 + 12 = -3$$

$$u_i = \frac{x_l - x_i}{h}$$

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

$$k(u_1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(-3)^2)$$

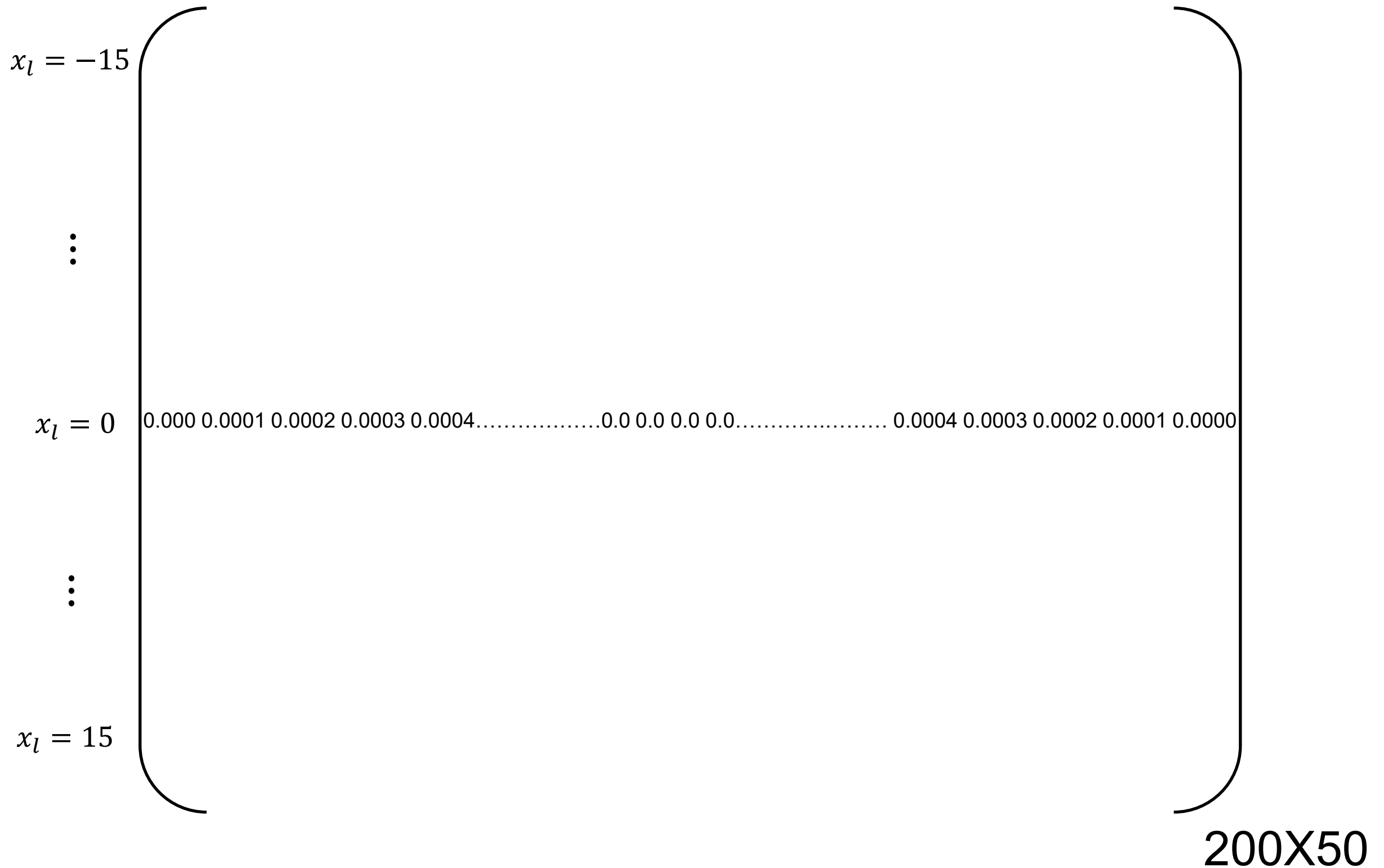
$h=1$



Density value

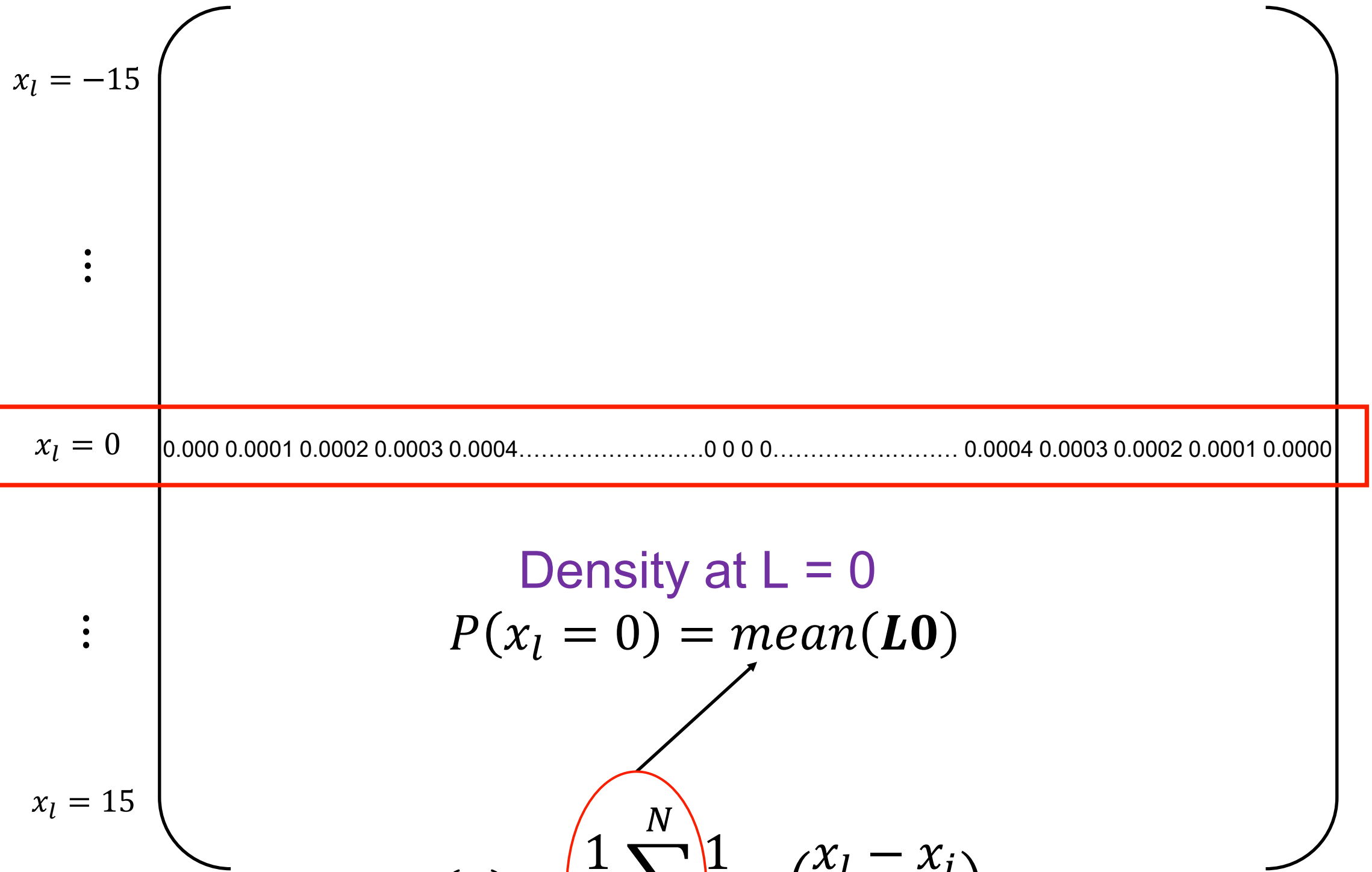
As an example of kernel heights for line at 0

Linearly spaced lines



Density value

Linearly spaced lines



Density at $L = 0$

$$P(x_l = 0) = \text{mean}(\mathbf{L0})$$

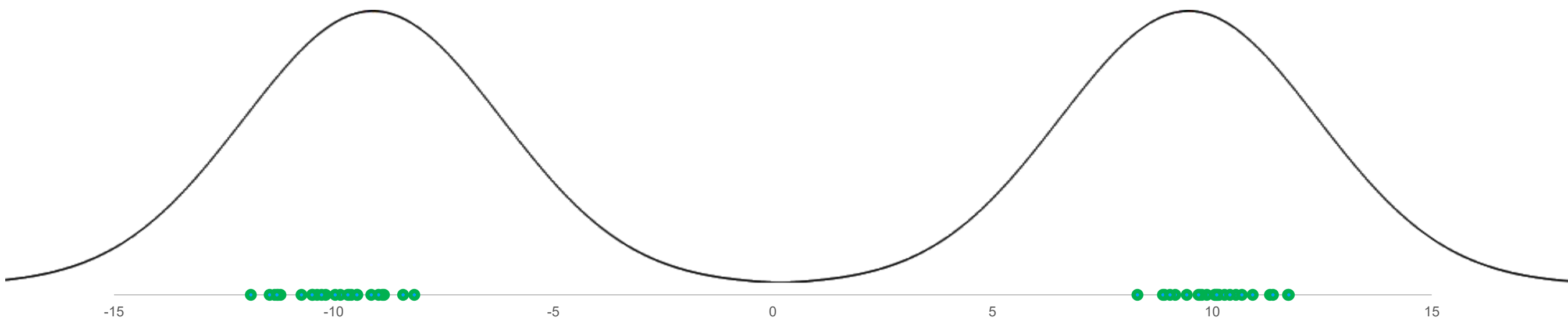
$$p(x) = \frac{1}{N} \sum_i^N \frac{1}{h} K \left(\frac{x_l - x_i}{h} \right)$$

200X50

Visual Example

Based on Gaussian kernel estimator

[Interactive Example](#)



For $\sigma = 1$;

Numerical Example

```
% Data ; There are 200 data points (-13~<data<~13)  
randn('seed',1) % Used for reproducibility  
x = [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(x) * (4/3/numel(x))^(1/5); % Bandwidth estimated by Silverman's Rule of Thumb
```

```
% Let's create apply density estimation over 1000 linearly spaced points ( $x_l$ )
```

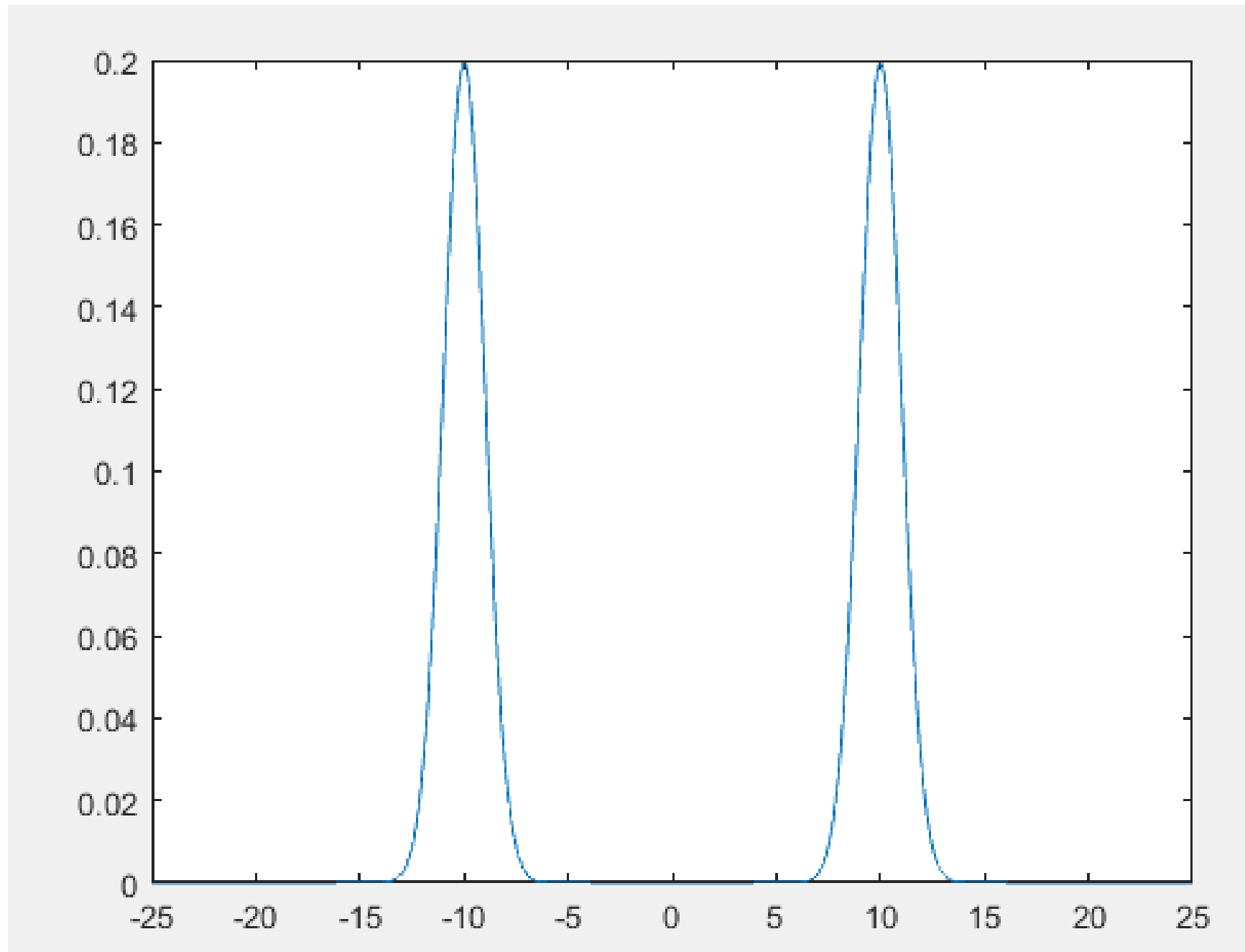
```
x1 = linspace(-25,+25,1000); % gridlines
```

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

```
truepdf_firstnormal = exp(-.5*(x1-10).^2)/sqrt(2*pi);  
truepdf_secondnormal = exp(-.5*(x1+10).^2)/sqrt(2*pi);  
truepdf = truepdf_firstnormal/2 + truepdf_secondnormal/2;  
% divided down by 2, because we are adding density value two times
```

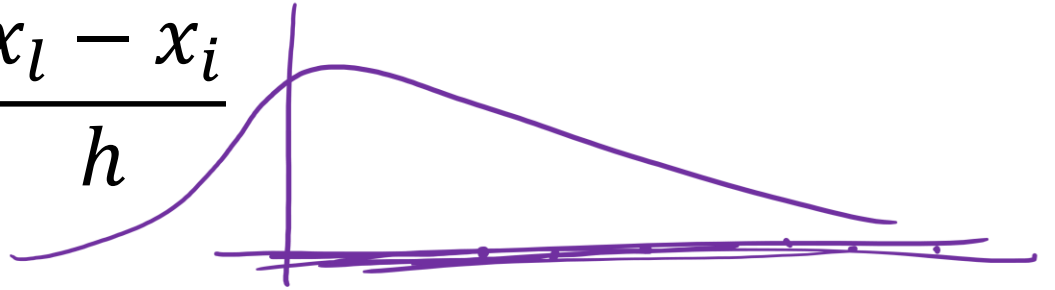
```
plot(x,truepdf)
```

```
% Plot True Density
```



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

$$p(x) = \frac{1}{N} \sum_i^N \frac{1}{h} K(u_i)$$

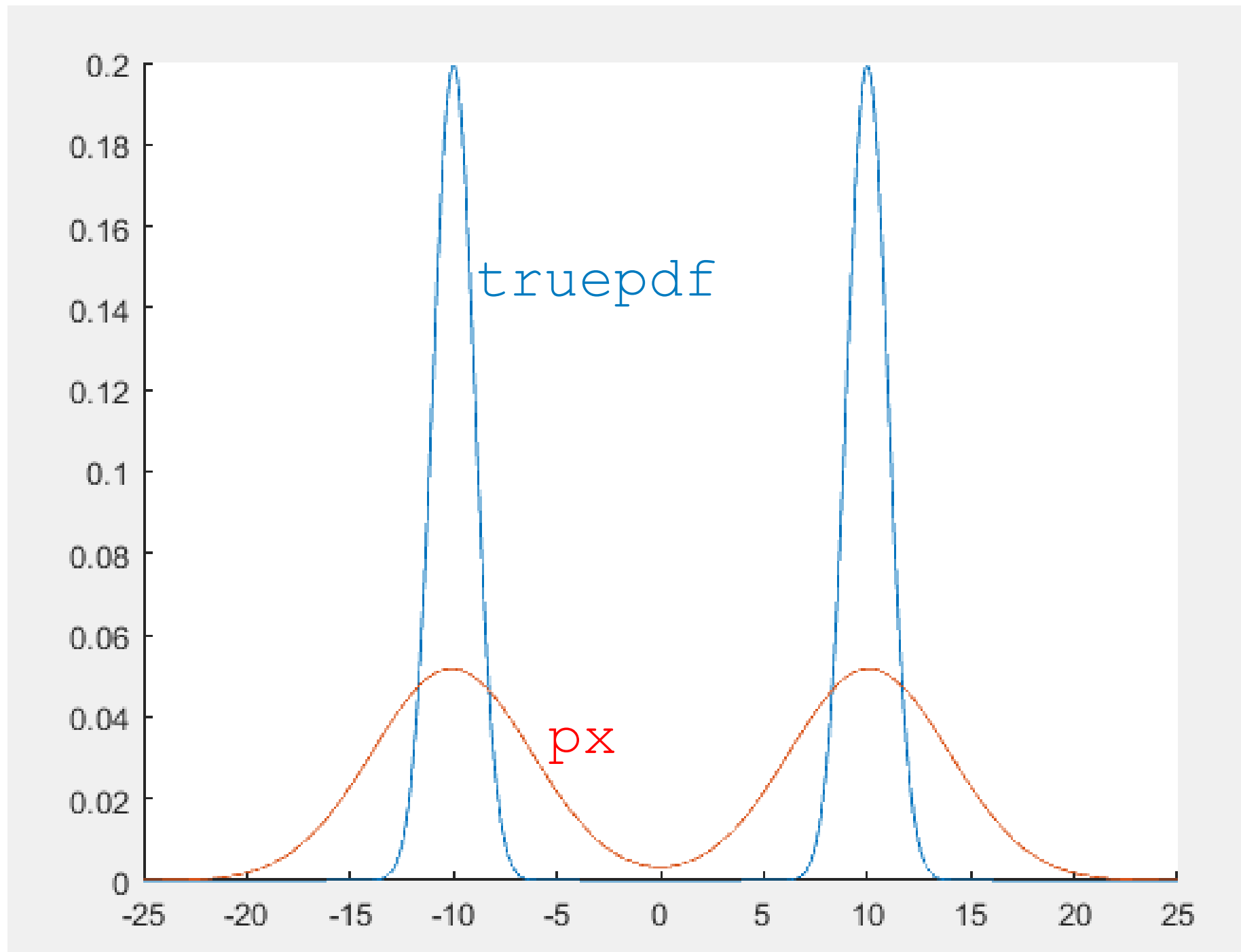
$$u_i = \frac{x_l - x_i}{h}$$


Gaussian kernel

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

```
for l=1:size(xl,1) % let's loop over grid lines (x_l)
    u = (xl(1) - x) ./ h; % length of u is 200
    Ku = exp(-.5*u.^2) / sqrt(2*pi);
    Ku = Ku ./ h;
    px(l) = mean(Ku);
end
```

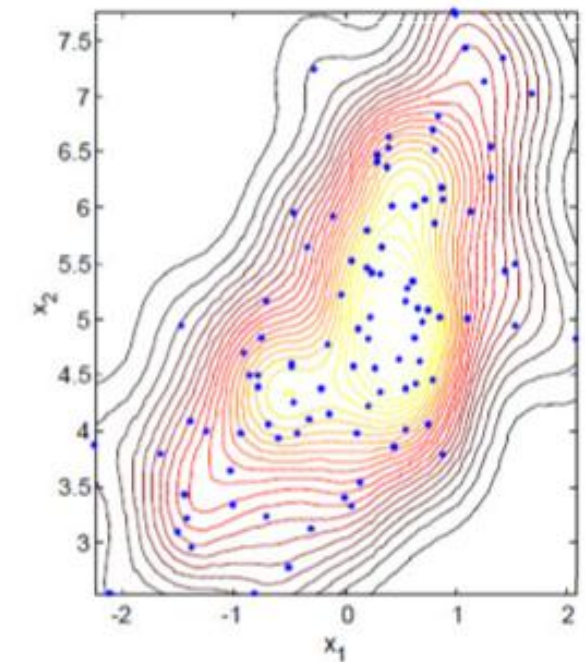
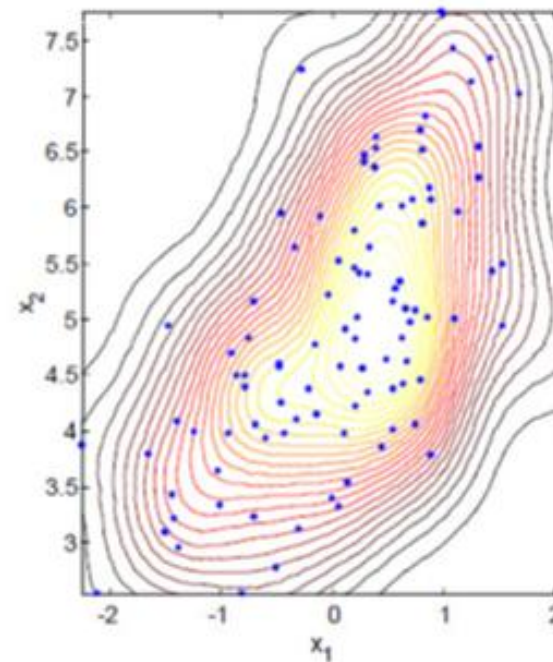
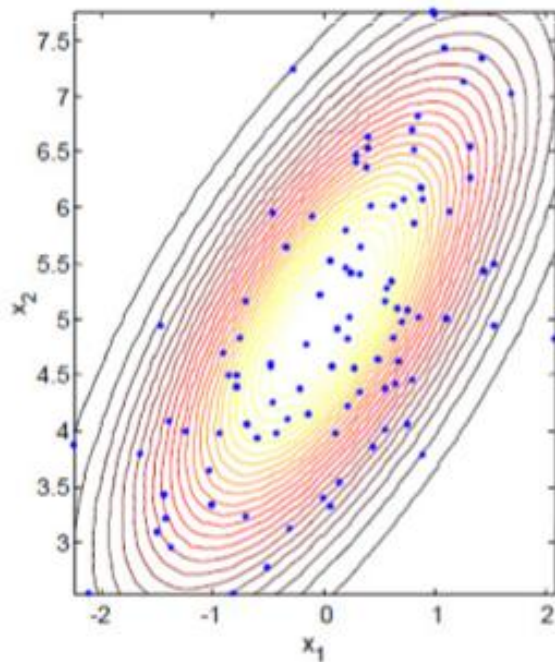
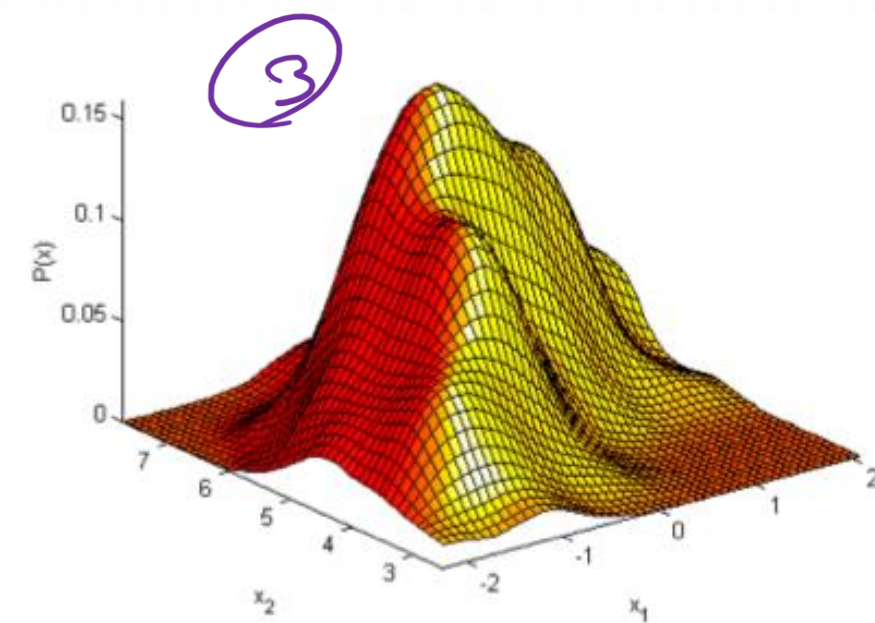
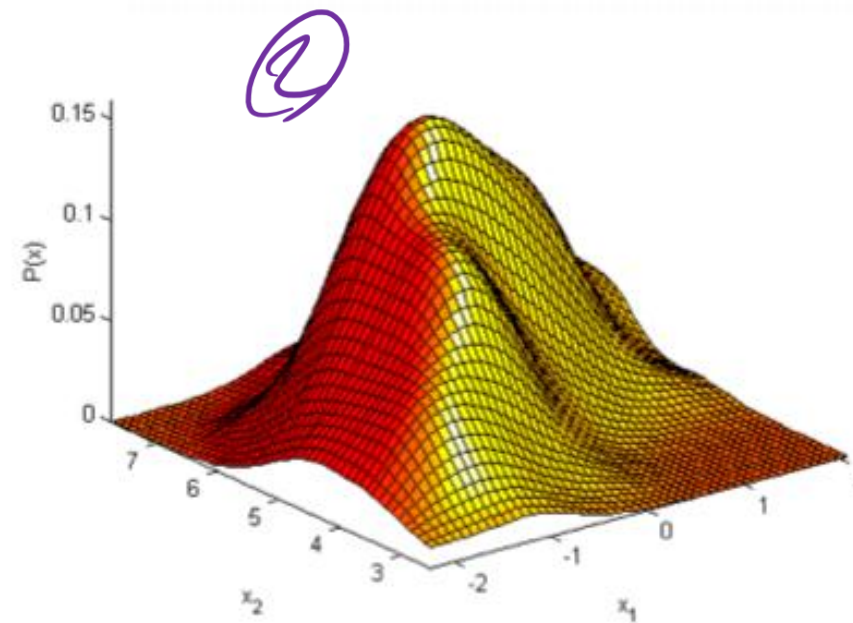
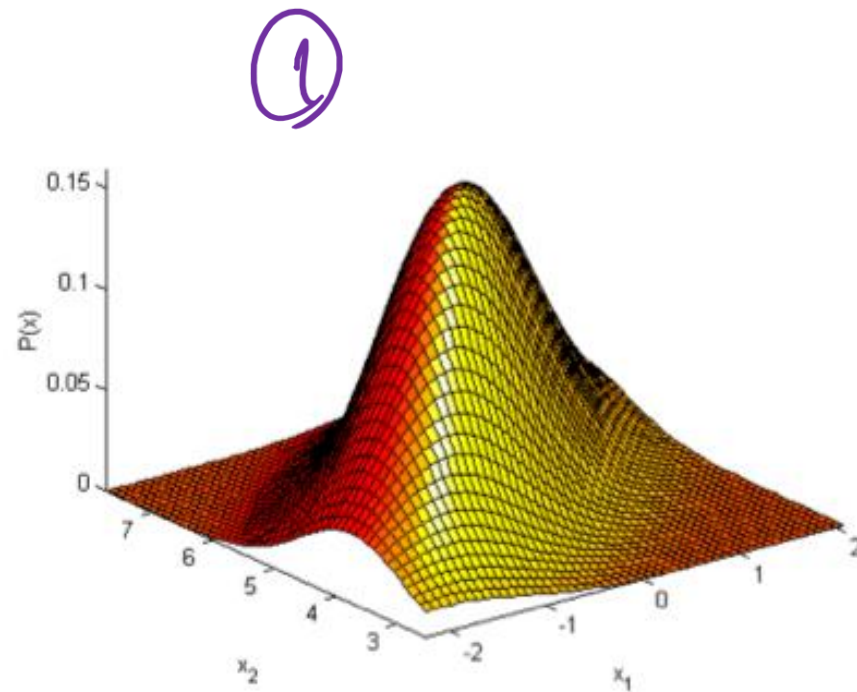
```
plot(x,truepdf)
plot(x,px)
```



Two-Dimensional Examples

— This example shows the product KDE of a bivariate unimodal 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

$$h \approx 1.06\hat{\sigma}N^{-\frac{1}{5}}$$

where $\hat{\sigma}$ 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

Non-parametric vs parametric



kde $\leadsto$ we need to play
around bandwidth

10 peak points



$$P(X|\pi, \mu, \Sigma) = \pi_0 N_0 + \dots + \pi_k N_k$$

30 parameters to optimize

Summary

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