CSE/ISyE 6740
Computational Data Analysis

Density Estimation

08/27/2025

Kai Wang, Assistant Professor in Computational Science and Engineering
kwang692@gatech.edu

Georgia Tech

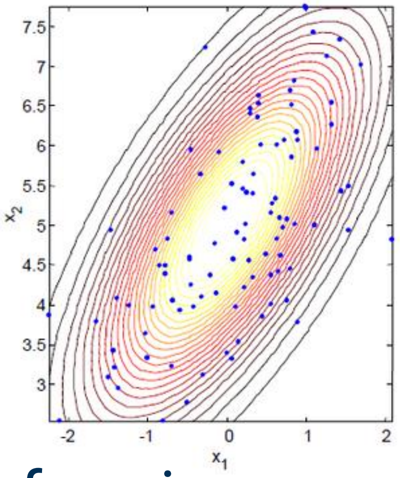
Outline

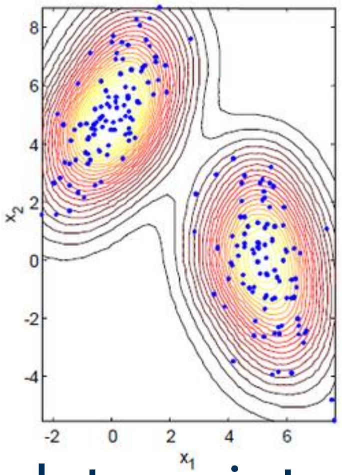
- **Unsupervised learning**
  - Density estimation
    - Parametric models
    - Non-parametric models
    - Kernel density estimation
    - Gaussian mixture models (lecture 5 - 6)
      - Expectation-Maximization algorithm

Density Estimation

Why Do We Need Density Estimation?

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



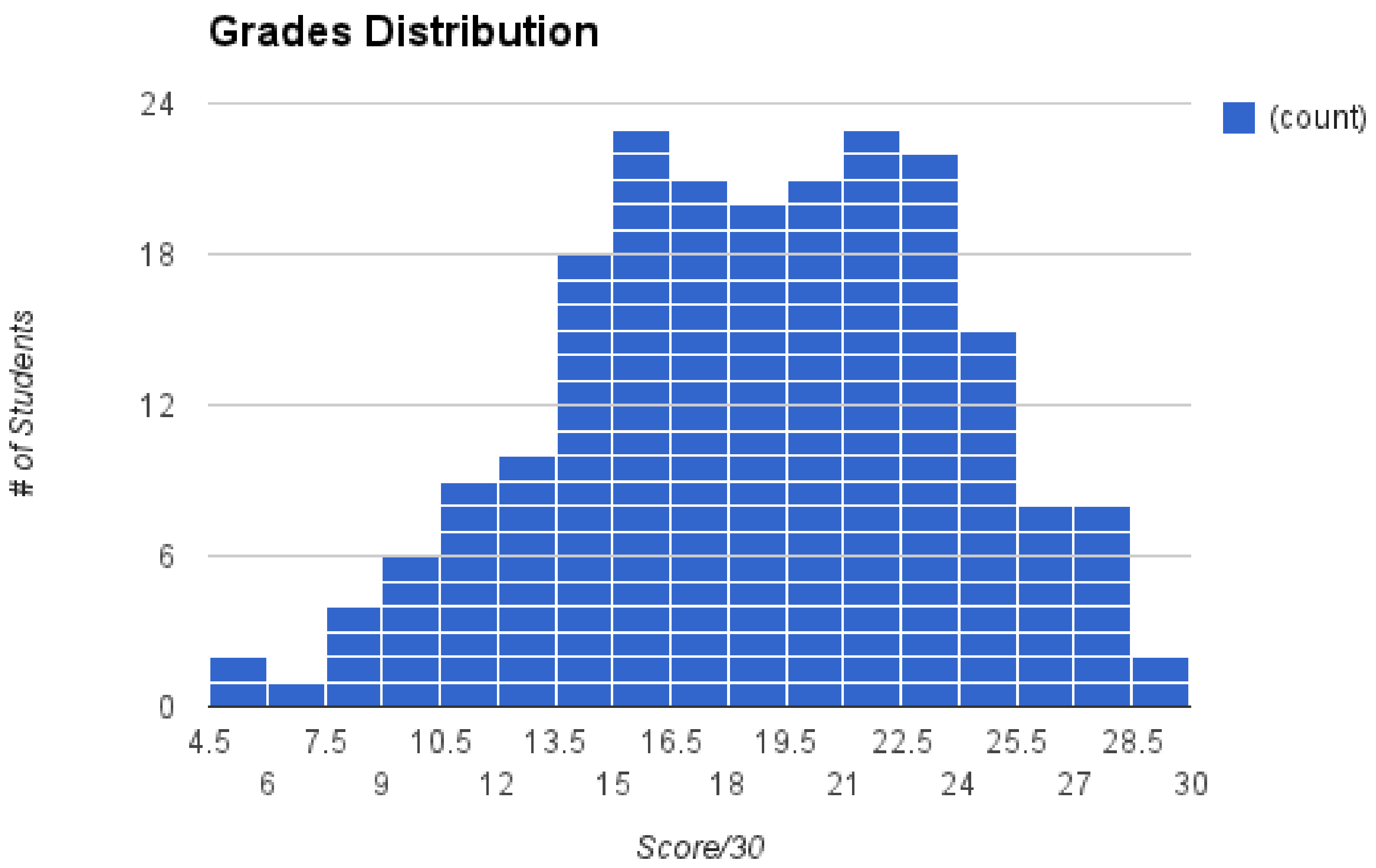


- Assess the likelihood of seeing a particular data point
  - Is this a typical data point? (high density value)
  - Is this an abnormal data point / outlier? (low density value)

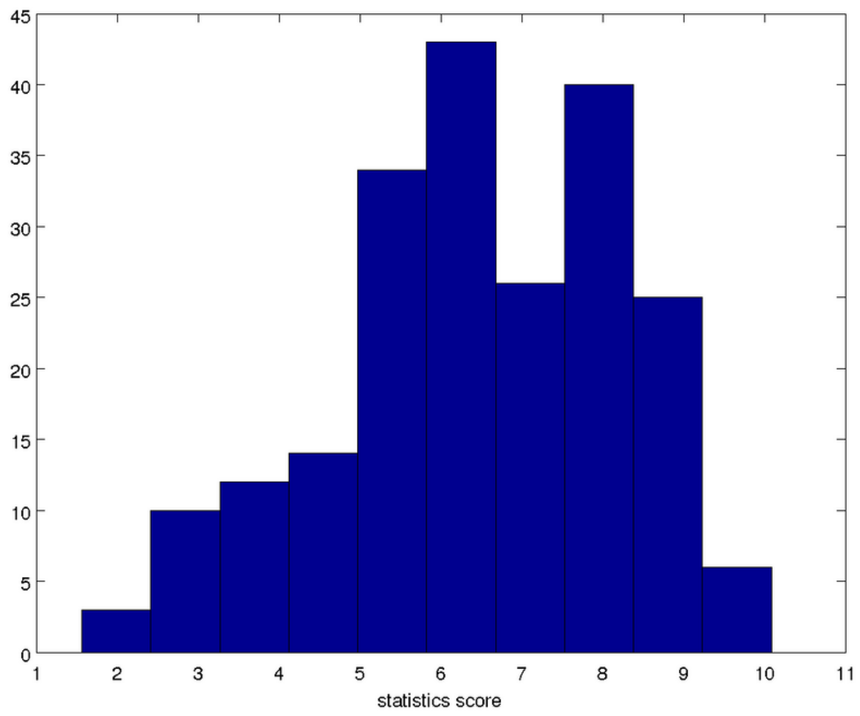
- Building block for more sophisticated learning algorithms
  - Classification, regression, graphical models…
  - A simple recommendation system

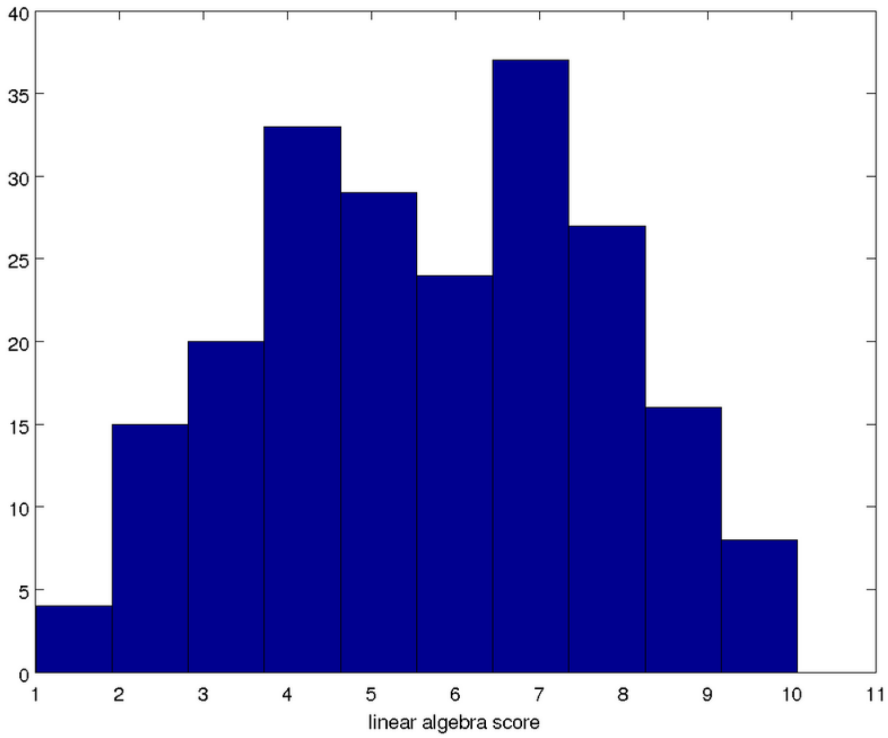
Georgia Tech

Example: Test Scores

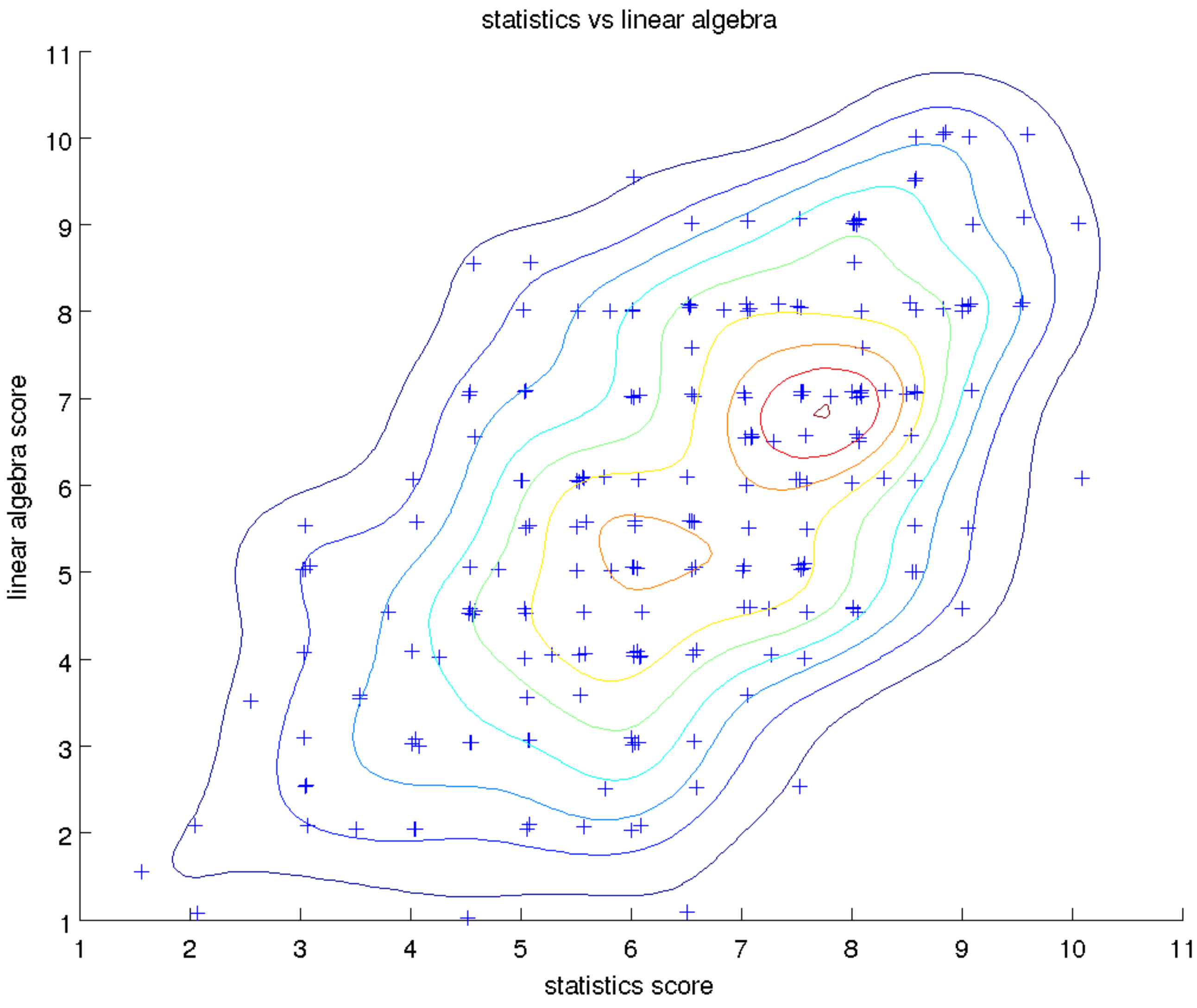


Example: Test Scores (conti.)





Example: Test Scores (conti.)



Parametric Models

Parametric Models

- Models 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 $\theta \in [0,1]$, which generates a family of models: $\mathcal{F} = \{P(x|\theta) \mid \theta \in [0,1]\}$

- **Continuous case**: e.g., multivariate Gaussian distribution

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

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

Estimation of 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 $D = \{x^1, x^2, \ldots, x^n\}$ drawn independently and identically (i.i.d.) from some distribution $P^*(x)$

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

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

Example Problem

- Estimate the probability θ of landing in heads using a biased coin

- Given a sequence of n independently and identically distributed (i.i.d.) flips
  - E.g., $D = \{x^1, x^2, \dots, x^n\}$ $(e.g., \{1,0,1,\dots,0\})$. where $x^i \in \{0,1\}$

- **Model**: $P(\mathrm{x}|\theta) = \theta^x(1-\theta)^x$
  - $P(x|\theta) = \begin{cases} 1-\theta. & for\ x = 0 \\ \theta. & for\ x = 1 \end{cases}$

- Likelihood of a single observation x^i?
  - $P(\mathrm{x}^i|\theta) = \theta^{x^i}(1-\theta)^{1-x^i}$

MLE of Biased Coin

- Objective function, log likelihood

$$l(\theta; D) = \log P(D|\theta) = \log \theta^{n_{head}}(1-\theta)^{n_{tail}}$$
$$= n_{head} \log \theta + (n - n_{head}) \log(1-\theta)$$

  - n_{head}: number of heads, n_{tail}: number of tails

- Maximize $l(\theta; D)$ w.r.t. θ

- Take derivatives w.r.t. θ

$$\frac{\partial l}{\partial \theta} = \frac{n_{head}}{\theta} - \frac{(n - n_{head})}{1-\theta} = 0$$

$\Rightarrow \hat{\theta}_{MLE} = \frac{n_{head}}{n}$ or $\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x^i$

Estimating Gaussian Distribution

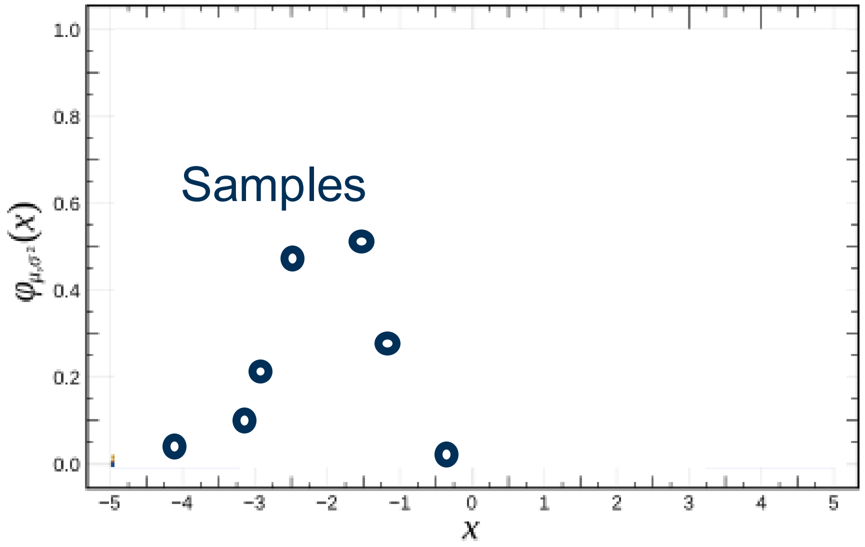
- Univariate Gaussian distribution in $\mathbb{R}$

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

- Need to estimate two sets of parameters μ, σ

- Given n i.i.d. samples

$$D = \{x^1, x^2, \ldots, x^n\}, \qquad x^i \in \mathbb{R}$$



- Likelihood of one data point:

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

The Estimators for μ, σ are Well-Known

- Univariate Gaussian distribution in $\mathbb{R}$

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

- **Mean**

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

- **Variance**

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

MLE for Gaussian Distribution

- Objective function, log likelihood

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

- Maximize $l(\mu, \sigma; D)$ 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

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

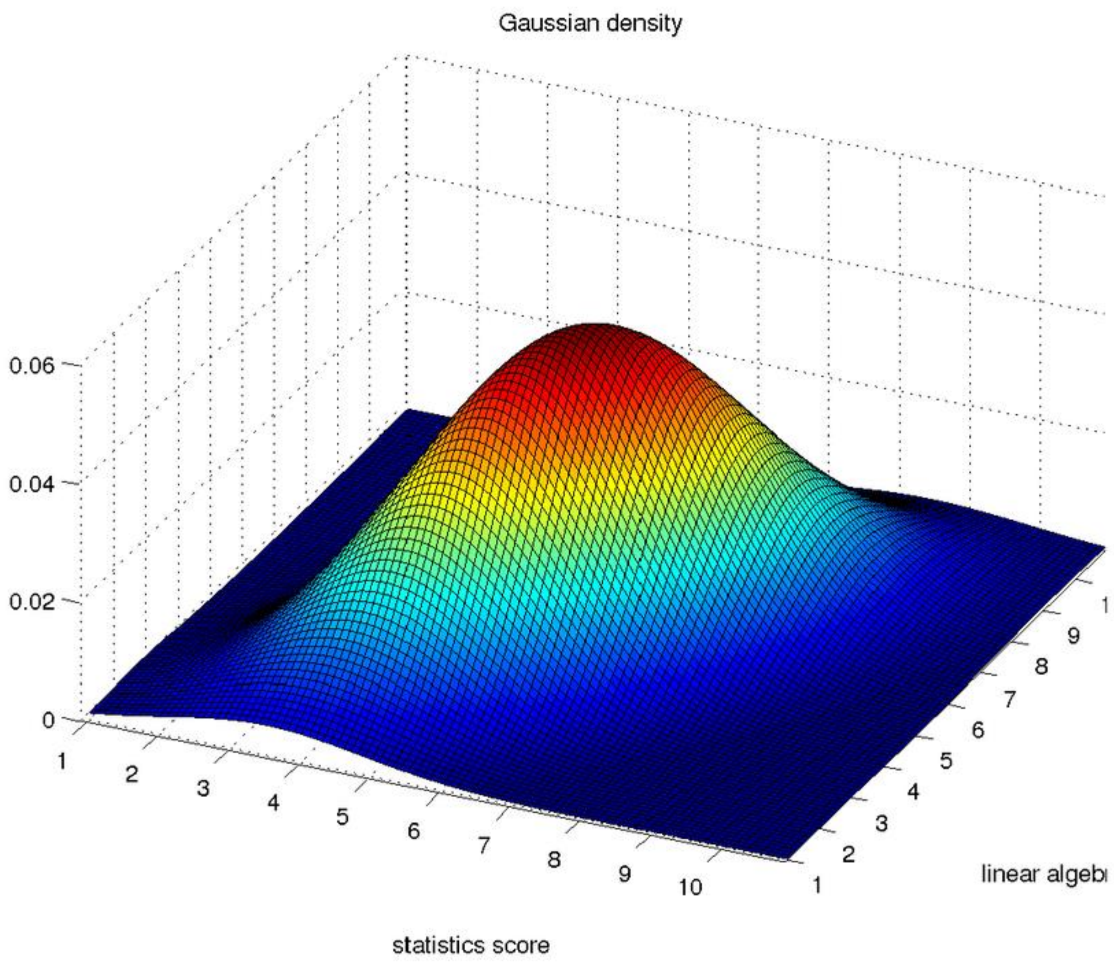
$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{x^i - \mu}{\sigma^2} = 0$$

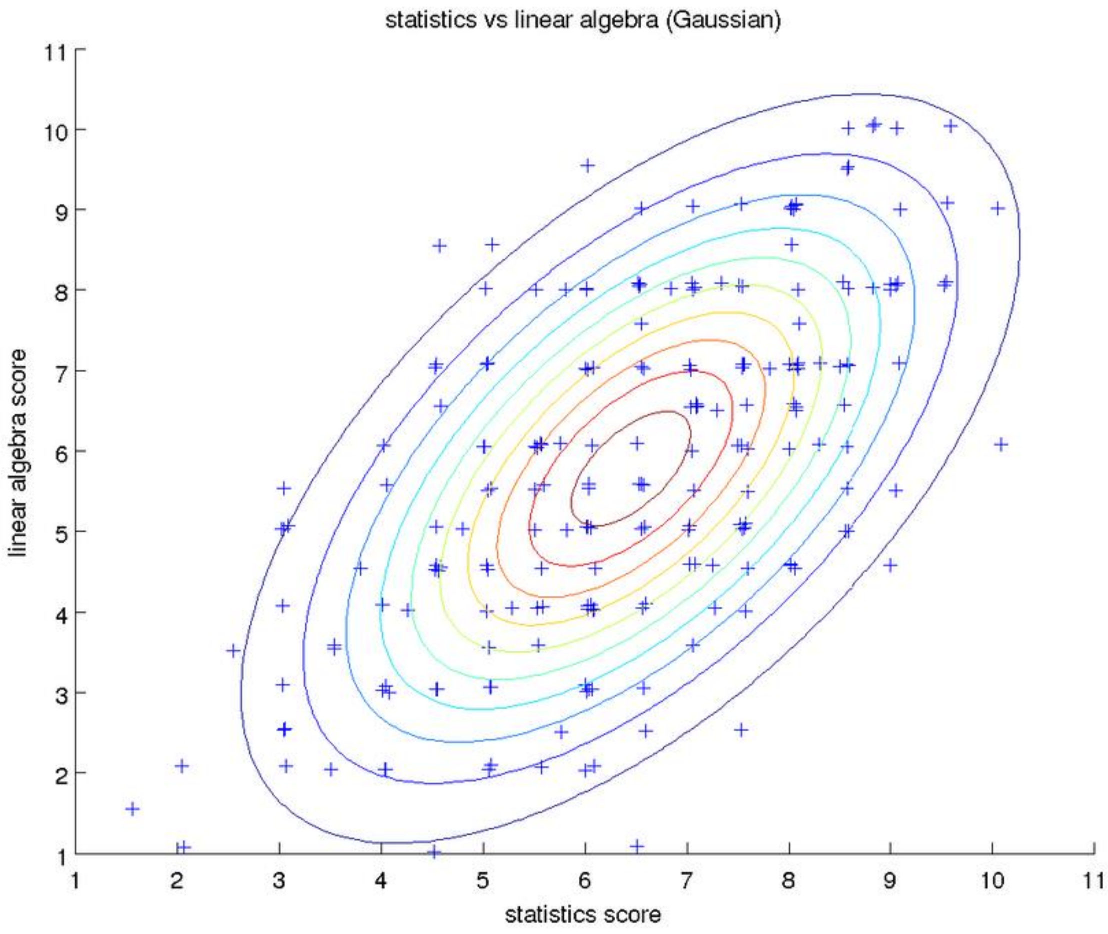
$$\Rightarrow \sum_{i=1}^{n} x^i = n\mu \qquad \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x^i$$

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

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

Density Example



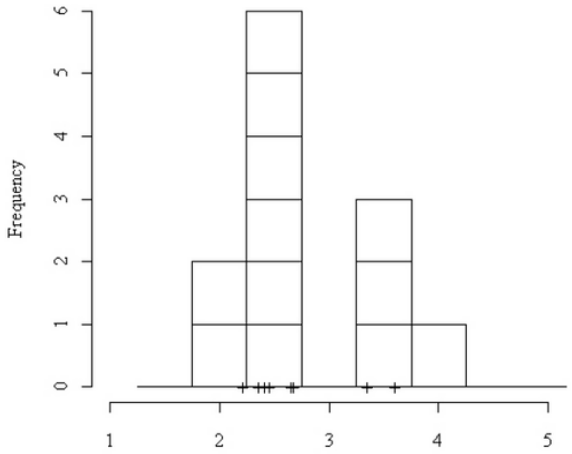


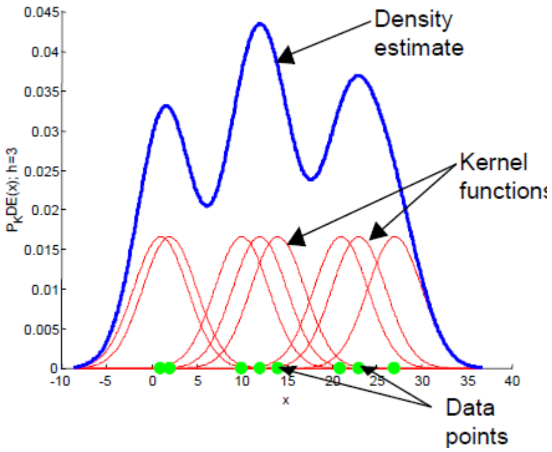
Nonparametric Models

Nonparameteric Models

- E.g., histogram

- E.g., kernel density estimator





- What are nonparametric models?
  - “Nonparametric” does **NOT** mean there are no parameters
  - Nonparametric models can **NOT** be described by **a fixed number of parameters**
  - One can think of there are many many (infinite) parameters

1-D Histogram

- One of the simplest nonparameteric density estimator
- Given n i.i.d. samples $D = \{x^1, x^2, \ldots, x^n\}, x^i \in [0,1)$
- Split $[0,1)$ into m bins

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

- Count the number of points: c_1 points in B_1, c_2 points in B_2,…
- For a new test point x

$$p(x) = \sum_{j=1}^{m} \frac{mc_j}{n} I(x \in B_j) \quad \text{(probability density function)}$$

Georgia Tech

Why is Histogram Valid?

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

- For histogram,

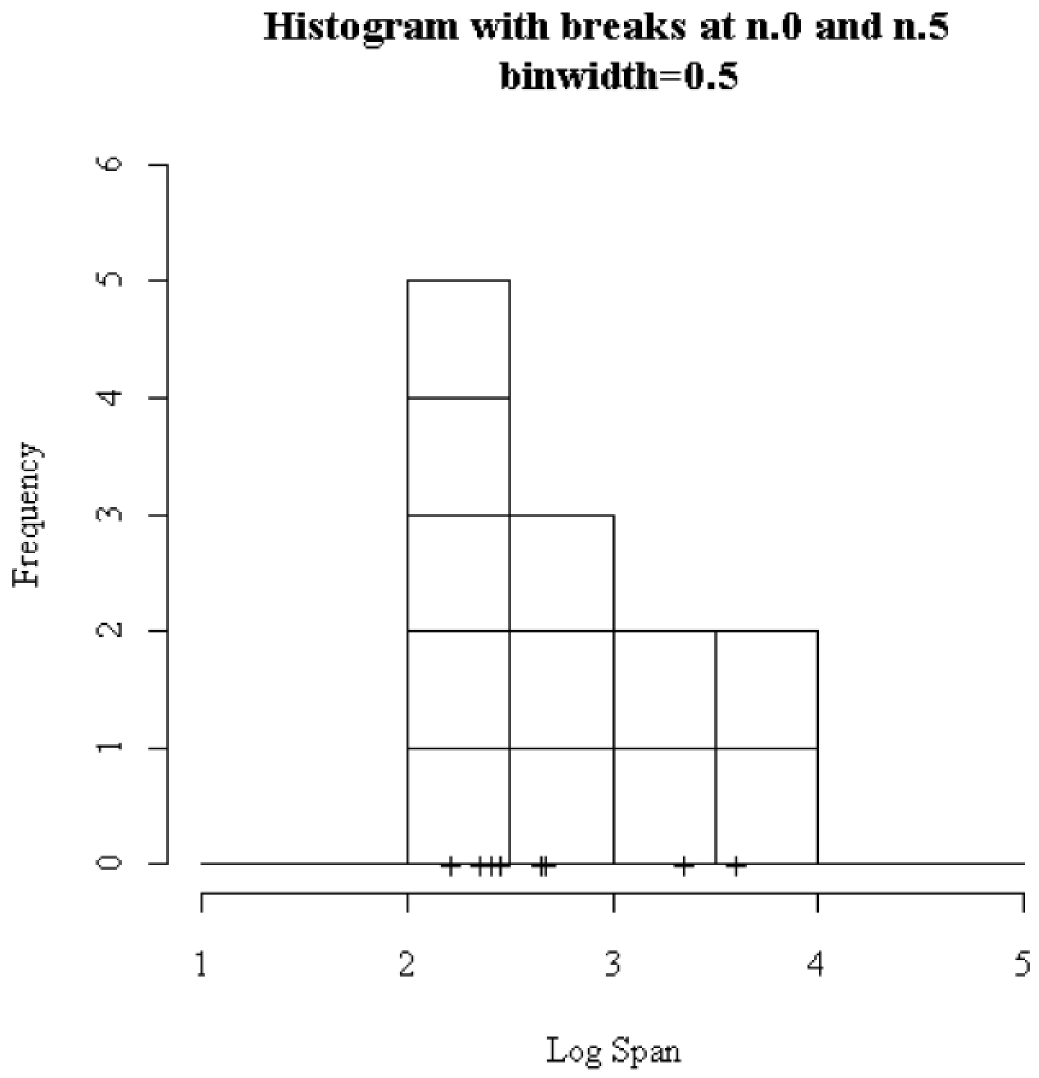
$$\int_\Omega p(x)dx = \int_{[0,1)} \sum_{j=1}^{m} \frac{mc_j}{n} I(x \in B_j)\, dx$$

$$= \sum_{j=1}^{m} \int_{\left[\frac{j-1}{m}, \frac{j}{m}\right)} \frac{mc_j}{n}\, dx$$

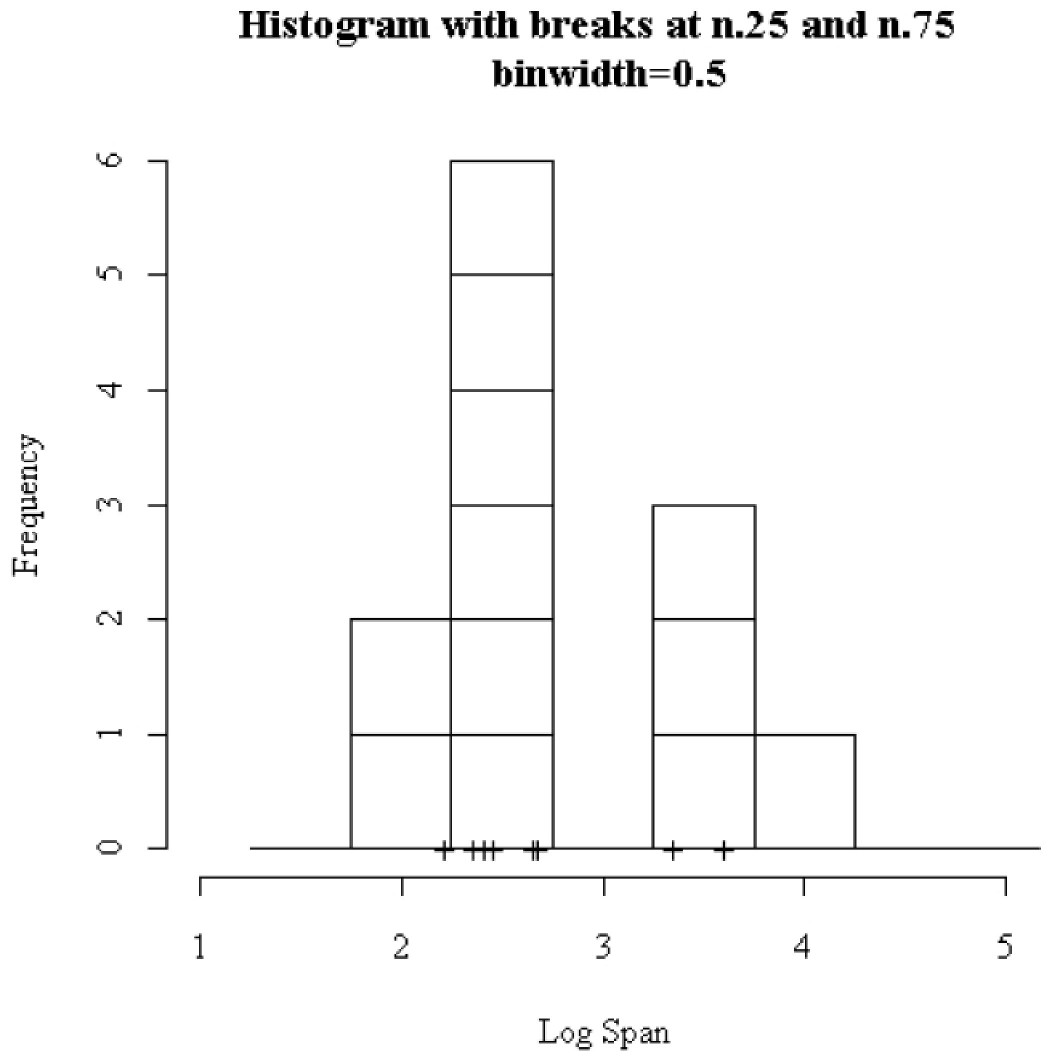
$$= \sum_{j=1}^{m} \frac{c_j}{n} = 1$$

Georgia Tech

Output Depends on Where You Put the Bins



Output Depends on Where You Put the Bins



Higher Dimensional Histogram

- Given n i.i.d. samples $D = \{x^1, x^2, \dots, x^n\}, x^i \in [0,1)^d$

- Split $[0,1)^d$ evenly into m^d bins

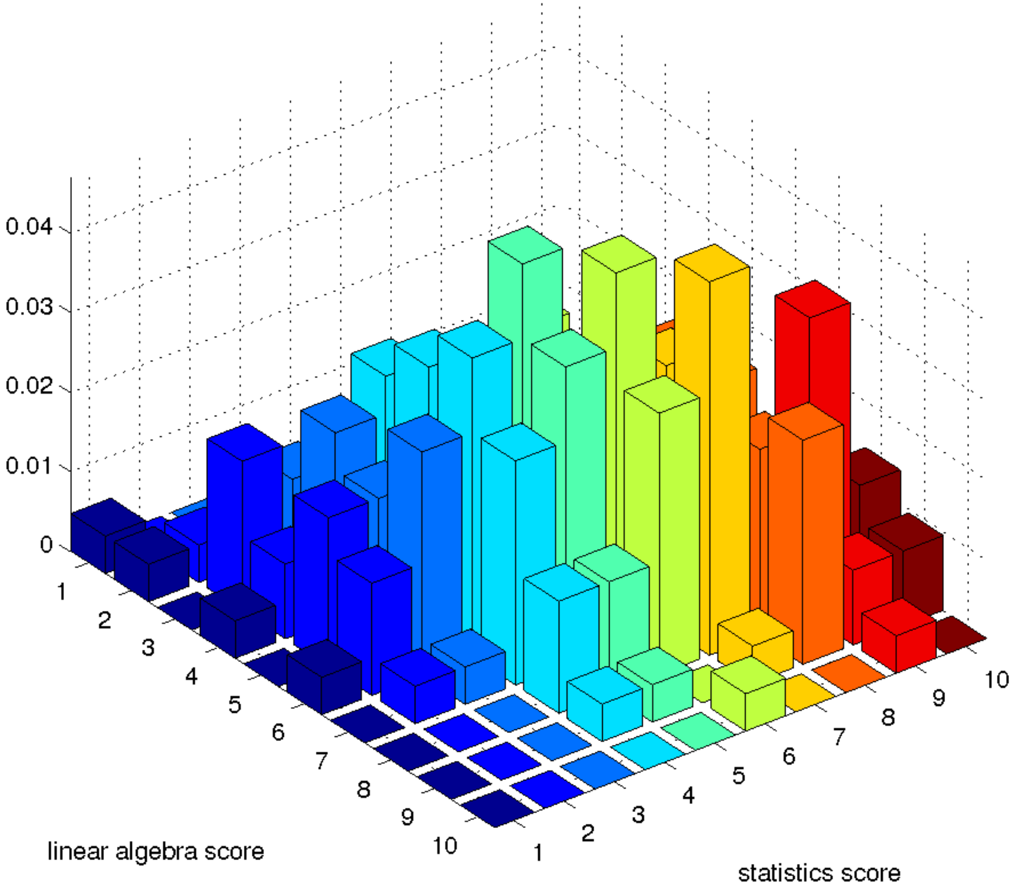
$$B_1 = \left[0, \frac{1}{m}\right) \times \left[0, \frac{1}{m}\right) \times \cdots \times \left[0, \frac{1}{m}\right),$$
$$B_2 = \left[\frac{1}{m}, \frac{2}{m}\right) \times \left[0, \frac{1}{m}\right) \times \cdots \times \left[0, \frac{1}{m}\right),$$

…

$$B_{m^d} = \left[\frac{m-1}{m}, 1\right) \times \left[\frac{m-1}{m}, 1\right) \times \cdots \times \left[\frac{m-1}{m}, 1\right)$$

- Bin size is $h = \frac{1}{m}$

Class Scores



Computation and Statistical Considerations

- **Problem I**: too many bins! Not good for high dimensional data
  - If m^d is larger than the number of samples n, most bins are empty
  - E.g., $m = 10, d = 6$, then we need ~1 million data samples

- **Problem II**: statistically histogram is not the best
  - Integrated risk:

$$r(\hat{p}, p) := \int_{\mathbb{R}} \mathbb{E}_X\left[\left(\hat{p}(x) - p(x)\right)^2\right] dx$$

  - Histogram (with bin size $h \sim n^{-1/3}$)

$$r(\hat{p}, p) \sim \frac{C}{n^{2/3}}$$

  - Difference even bigger for high dimensional data

Kernel Density Estimation

Kernel Density Estimation

- Kernel density estimator

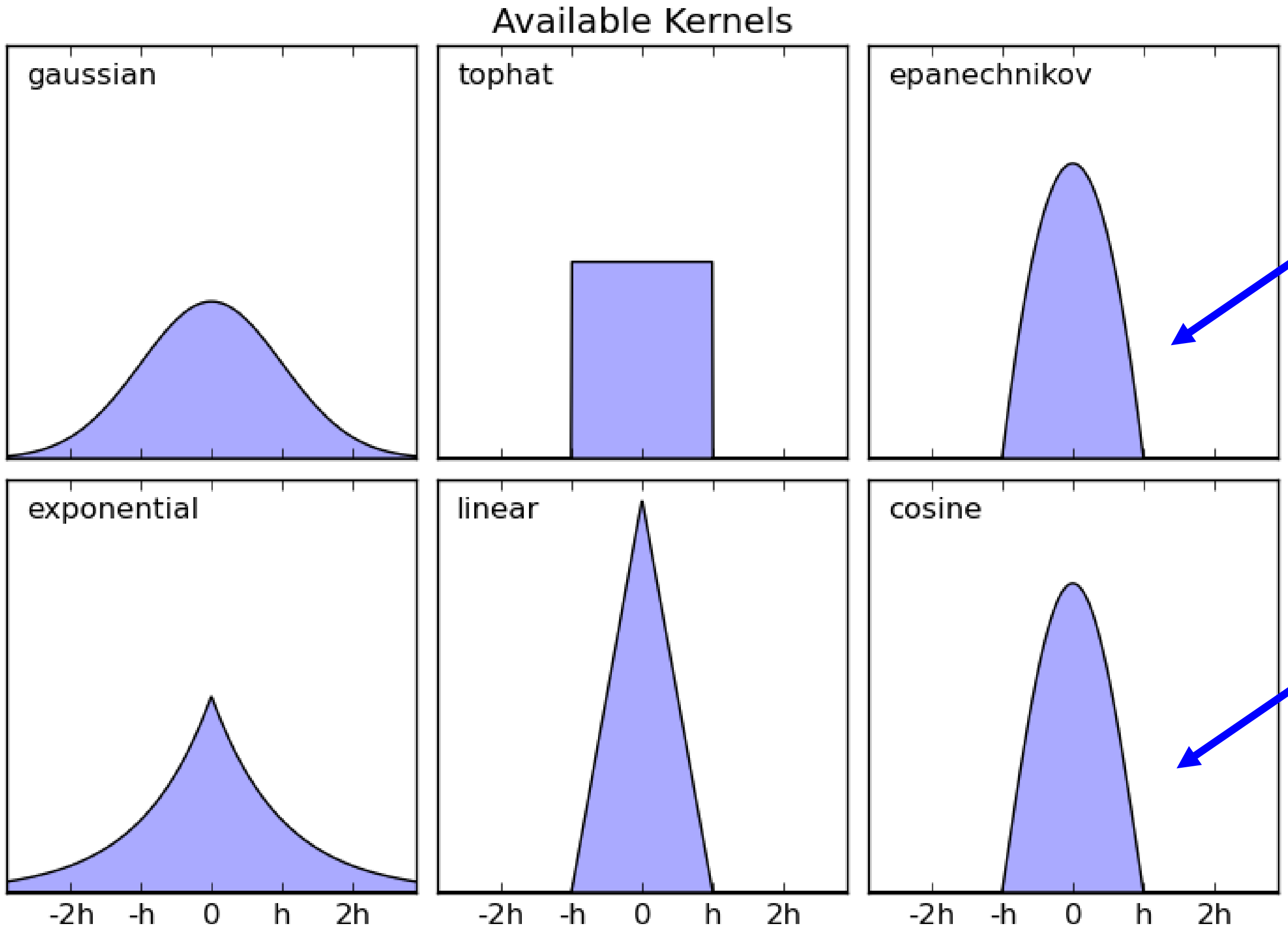
$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x^i - x}{h}\right)$$

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

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

Smoothing Kernel Functions

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

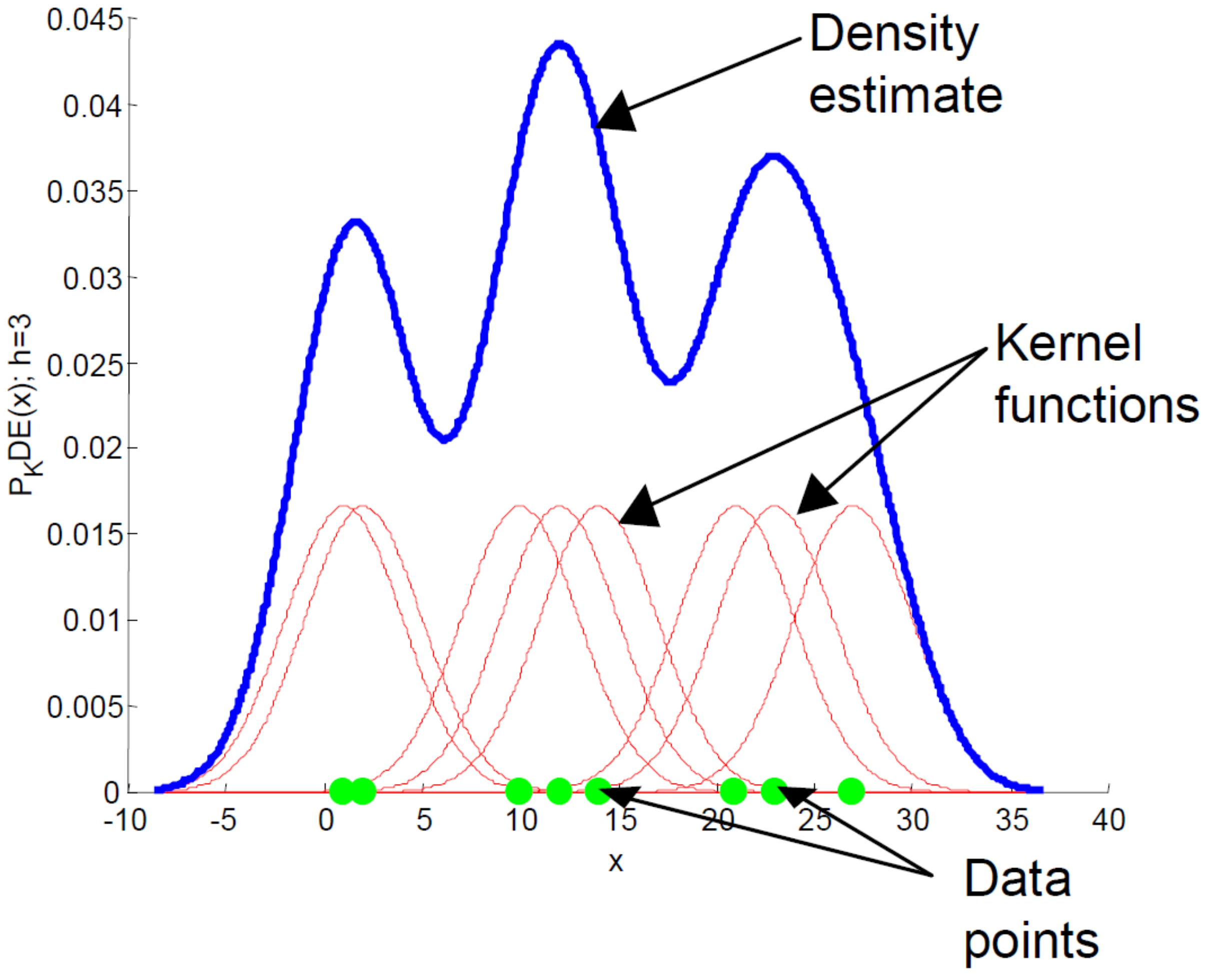


$$K(u) = \frac{3}{4}(1 - u^2) I(|u| \leq 1)$$

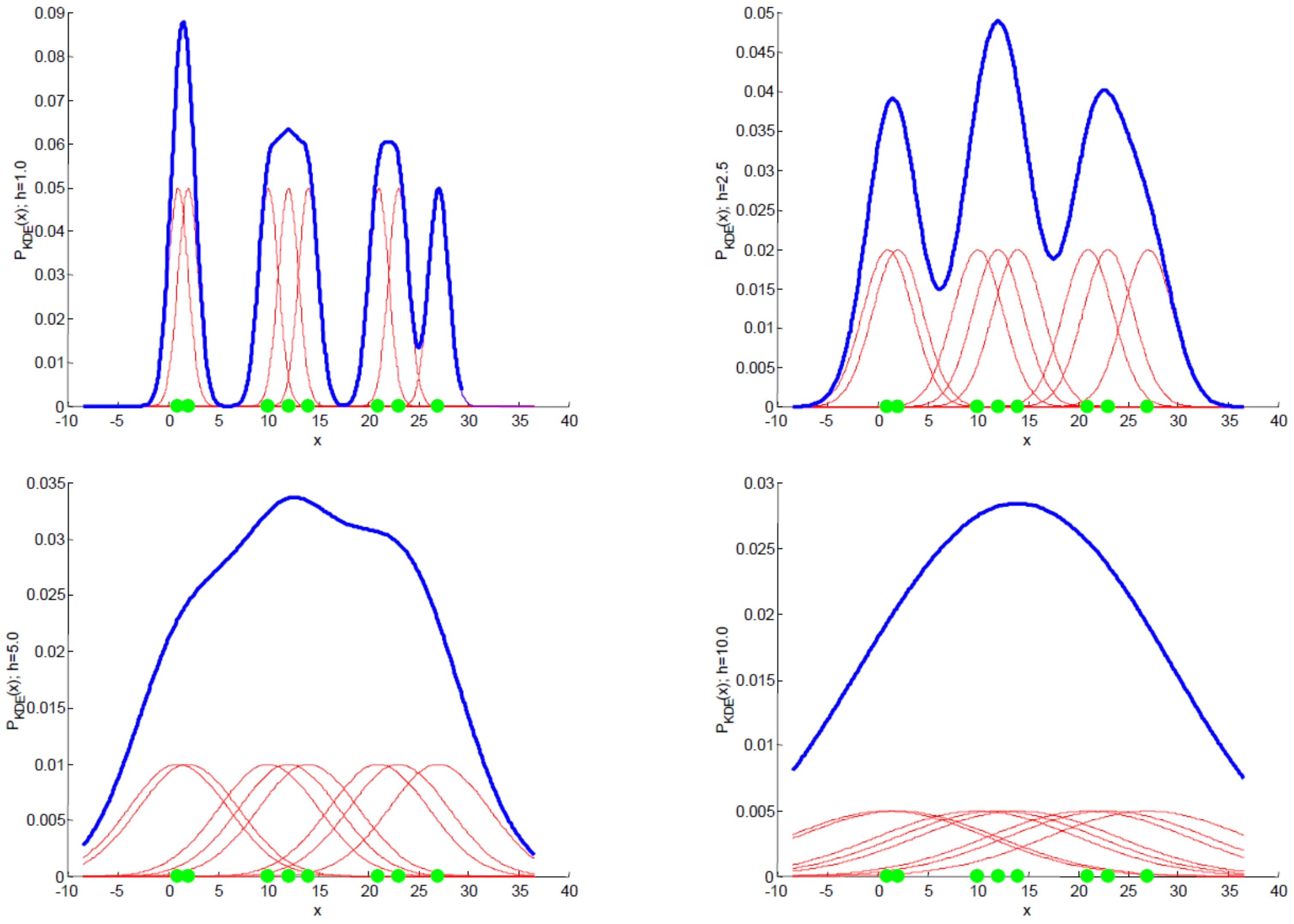
$$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right) I(|u| \leq 1)$$

Example

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

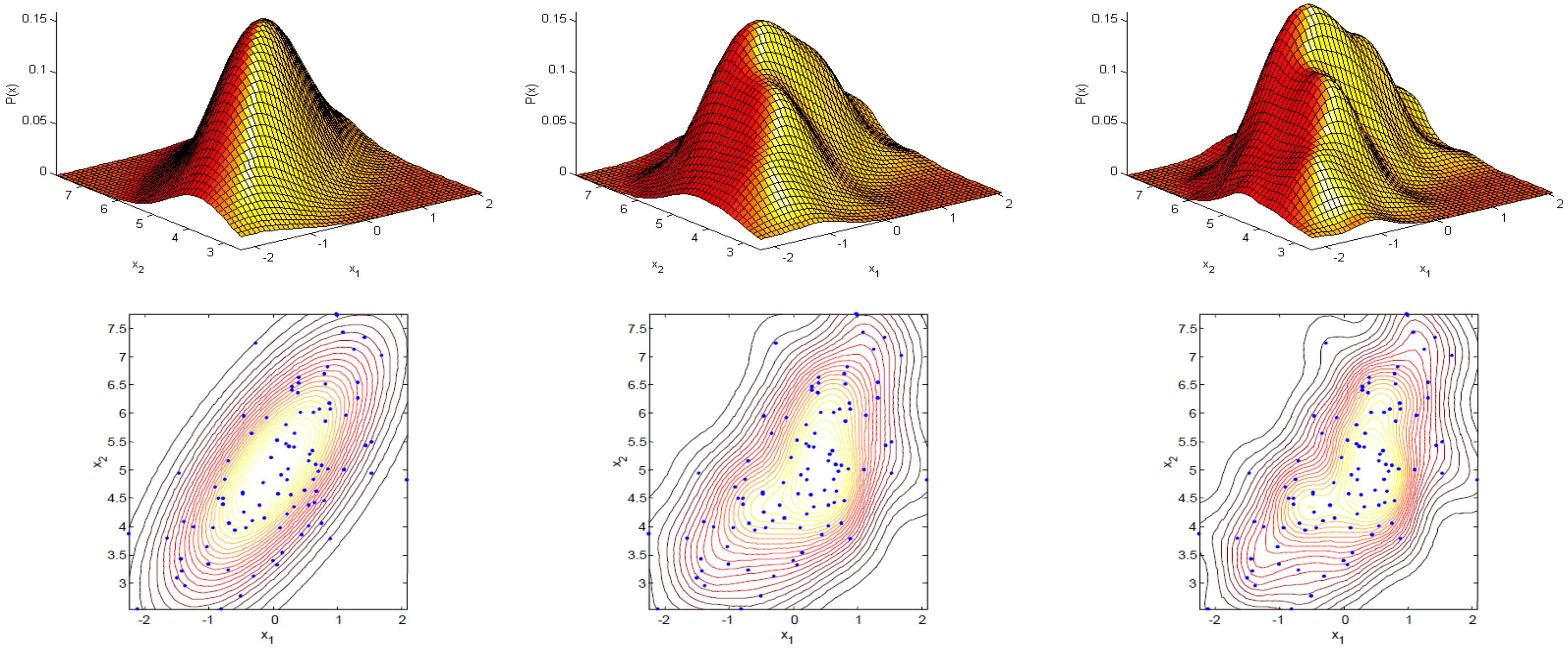


Effect of the Kernel Bandwidth h



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

Two-dimensional Example



What is the Best Kernel Bandwidth?

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

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

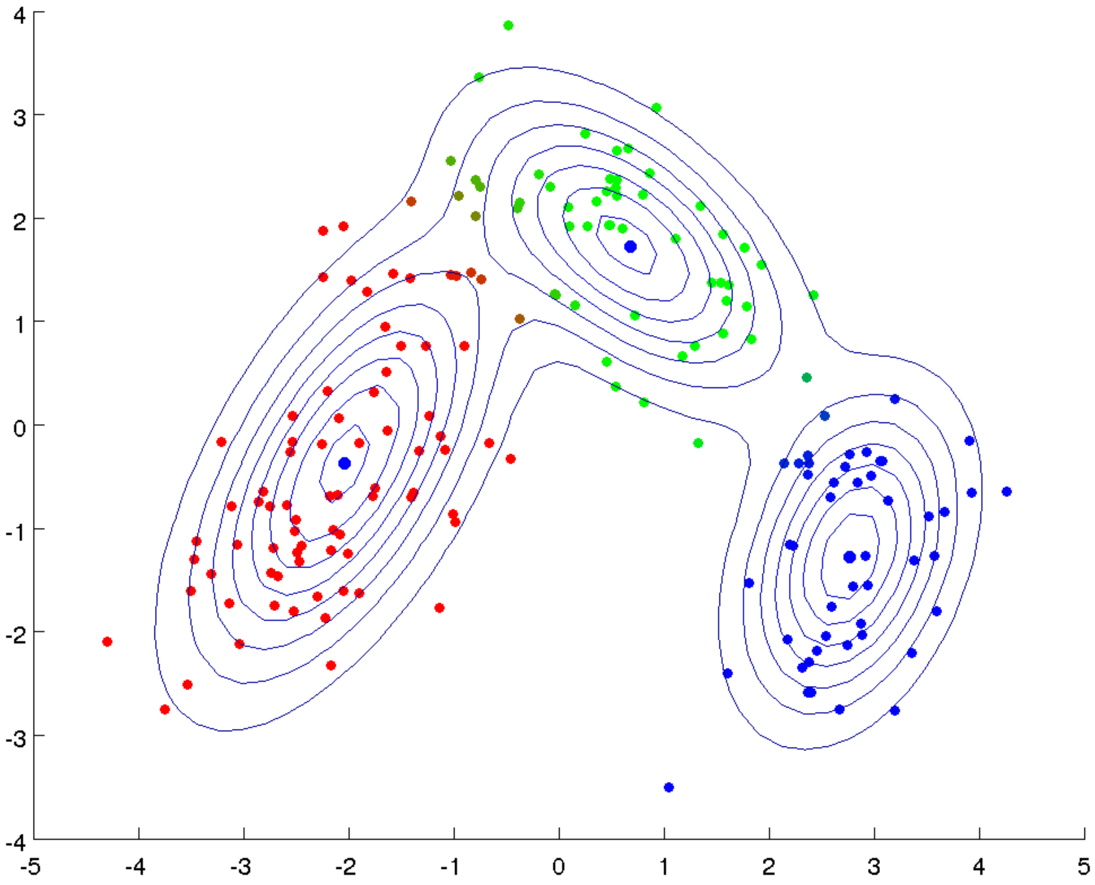
where $\hat{\sigma}$ is the standard deviation of the samples

- A better but more computational intensive approach:
  - Randomly split the data into two sets
  - Obtain a kernel density estimate for the first set
  - Measure the likelihood of the second set
  - Repeat over many random splits and take the average

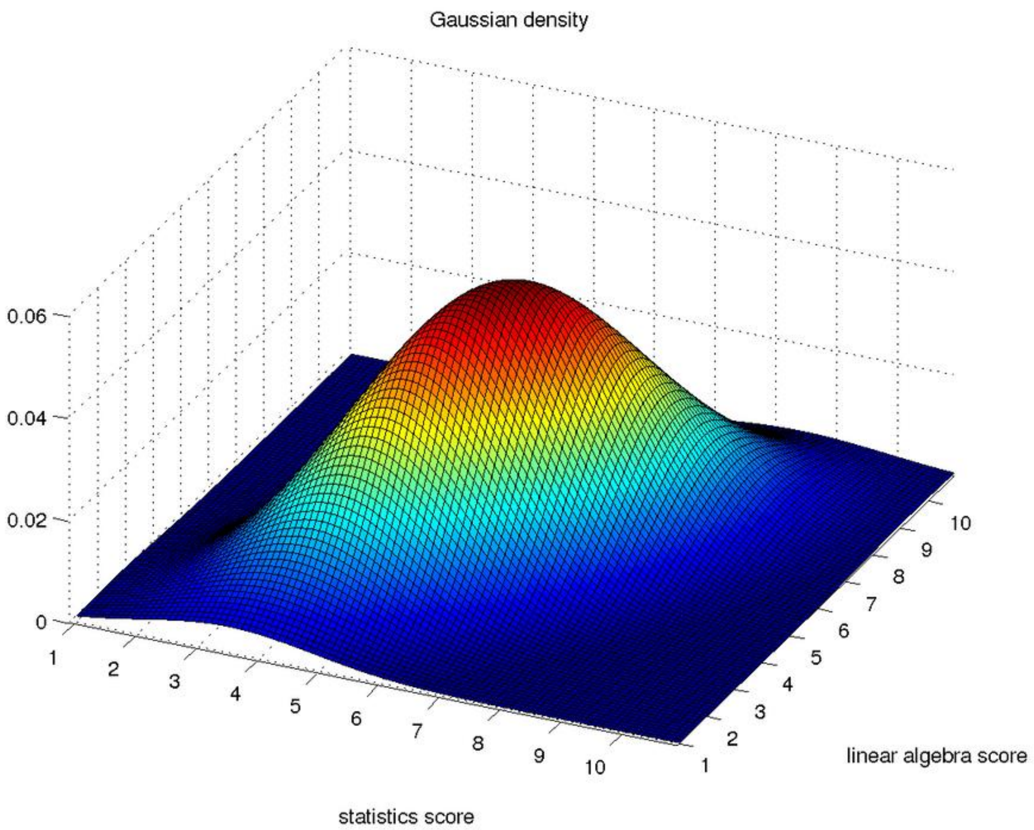
Wine Dataset (http://archive.ics.uci.edu/ml/datasets/Wine+Quality)

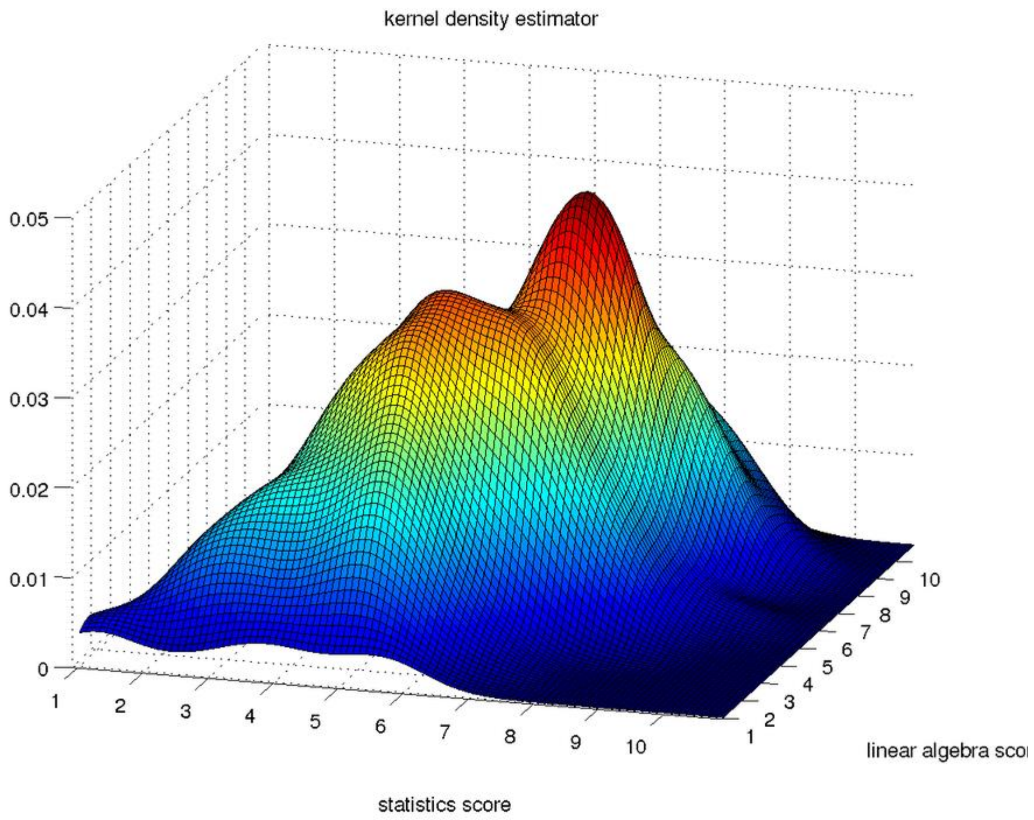
- These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. Feature include
  - 1) Alcohol
  - 2) Malic acid
  - 3) Ash
  - 4) Alcalinity of ash
  - 5) Magnesium
  - 6) Toal pheonols
  - 7) Flavanoids
  - 8) Nonflavanoid phenols
  - 9) Proanthocyanins
  - 10) Color intensity
  - 11) Hue
  - 12) OD280/OD315 of diluted wines
  - 13) Proline

Demo: test_wine.py

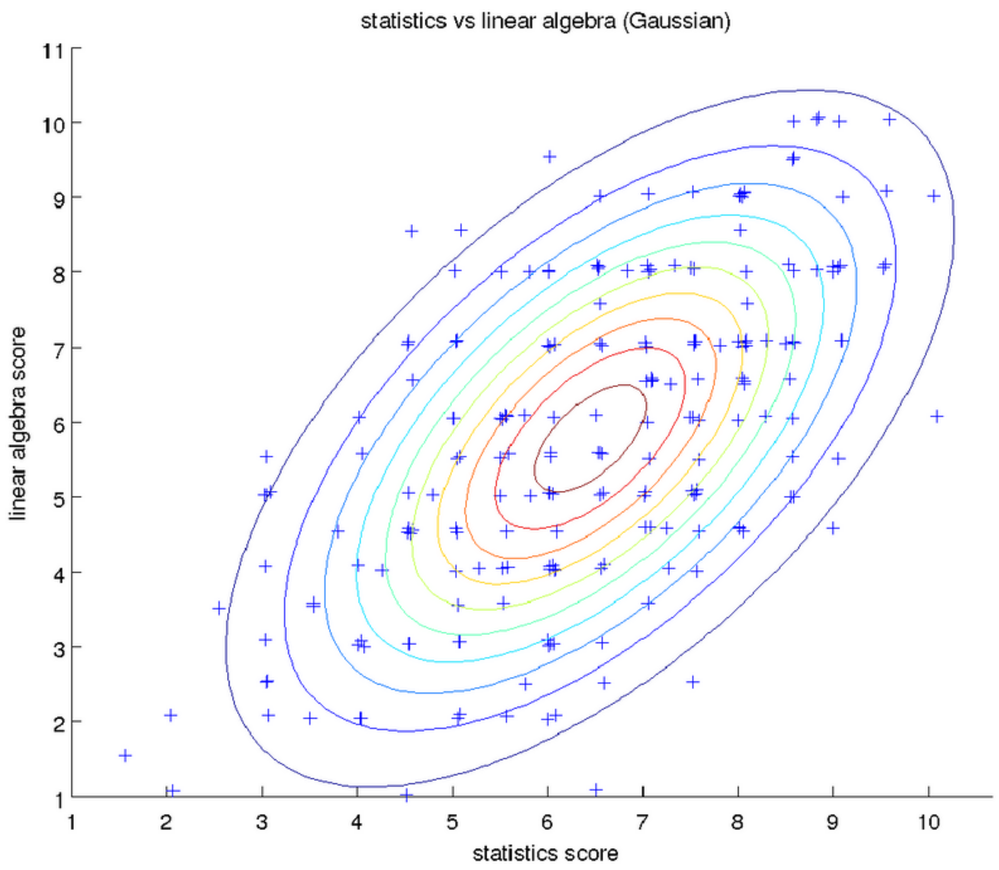


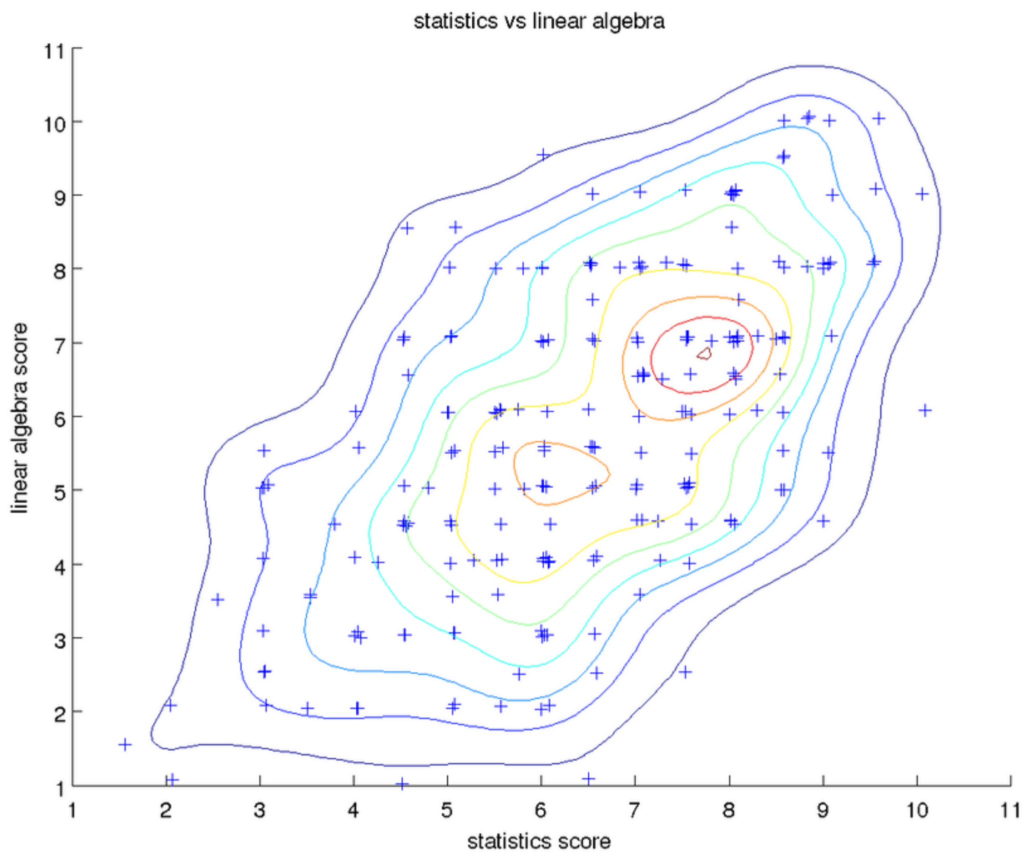
Parametric v.s. Nonparametric





Parametric v.s. Nonparametric





Parametric v.s. Nonparametric

- **Nonparametric** models place very mild assumptions on the data distribution and provide good models for complex data
- **Parametric** models rely on very strong (simplistic) distributional assumptions

- **Nonparametric** models (not histograms) requires storing and computing with the entire data set.
- **Parametric** models, once fitted, are much more efficient in terms of storage and computation.

Parametric v.s. Nonparametric

- Data $x \in \mathbb{R}^d$ with fixed dimension d
- Given n training data points $\{x^1, x^2, \ldots, x^n\}$
- Partition m bins in each dimension

Aspects	Gaussian	Histogram	KDE
Flexible	No	Yes	Yes
Assumption	Strong	Mild	Mild
Parameter number	Fixed	Increased with m	Increased with n
Memory requirement	$d + d^2$	m^d	nd
Training computation	Closed form	Binning and counting	Nothing
Test computation	Plug in formula	Find the bin	Evaluate n functions
Statistical guarantee	Only Gaussian case	Arbitrary (worse)	Arbitrary (better)

Announcement

- No class on **Sep 1st**. Happy long weekend!

- Hw1 will be released on **Aug 31st (Sunday)** and due on **Sep 14th (Sunday)**. Please start working on it earlier!