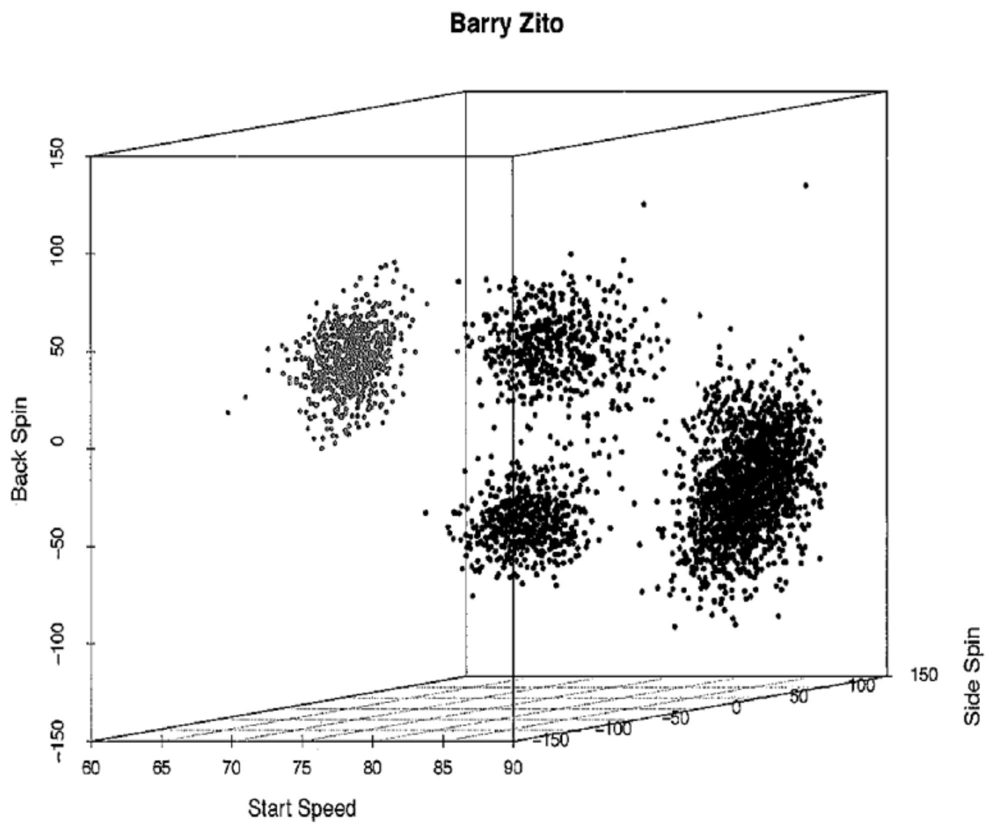
Density Estimation and Mode Finding

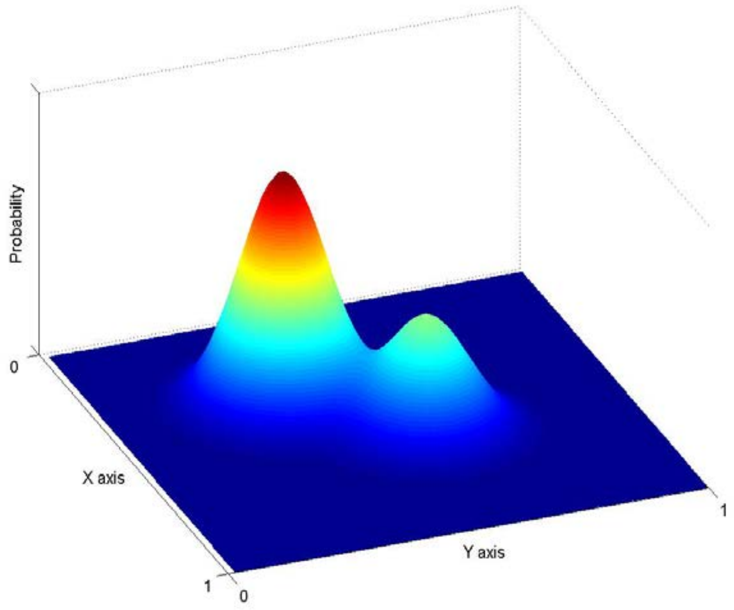
Clustering baseball pitches

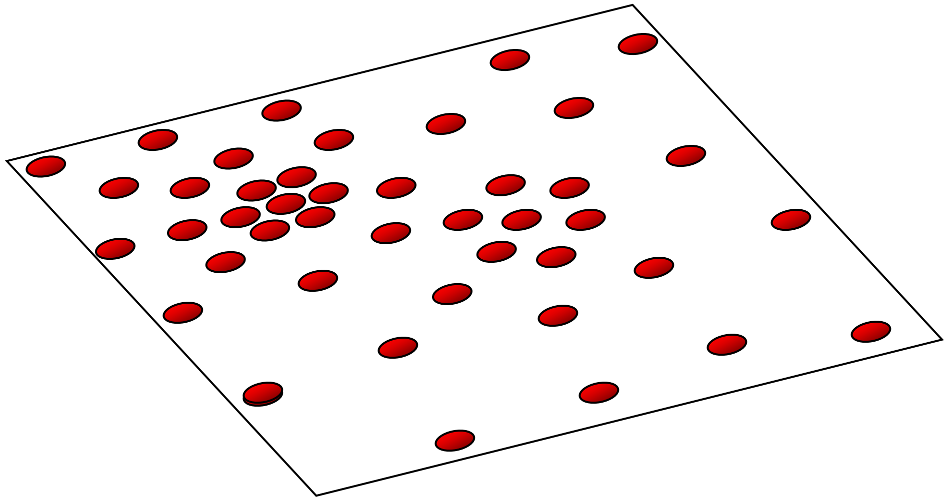


Inferred meaning of clusters: black – fastball, red – sinker, green – changeup, blue – slider, light blue – curveball

Probabilistic Interpretation: Density Estimation

The data points are sampled from an underlying PDF





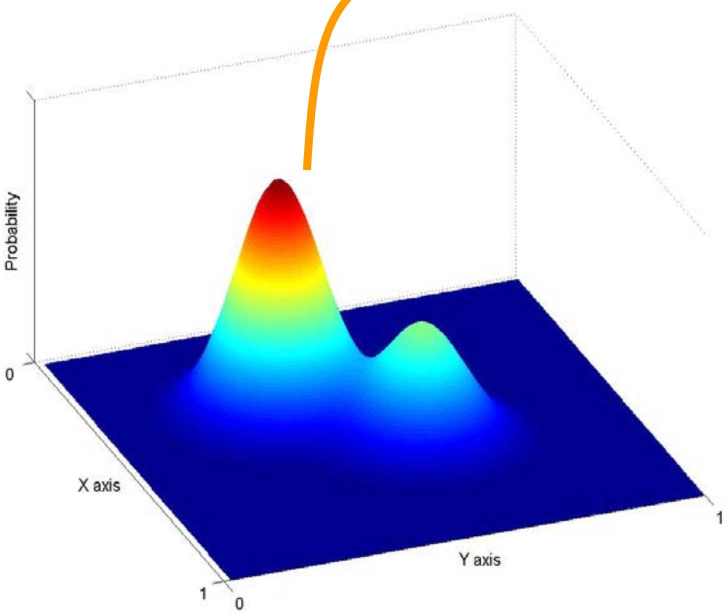
Assumed Underlying PDF

Data Samples

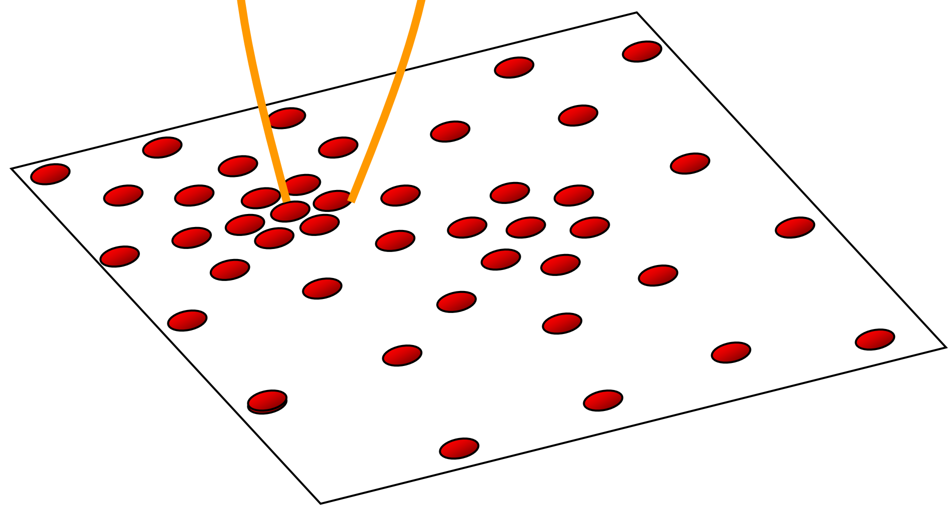
Parametric Density Estimation

Just fit a Gaussian!

$$p(\mathbf{x}) = e^{-\frac{(\mathbf{x}-\mu_i)^2}{2\sigma_i^2}}$$



Estimate from data



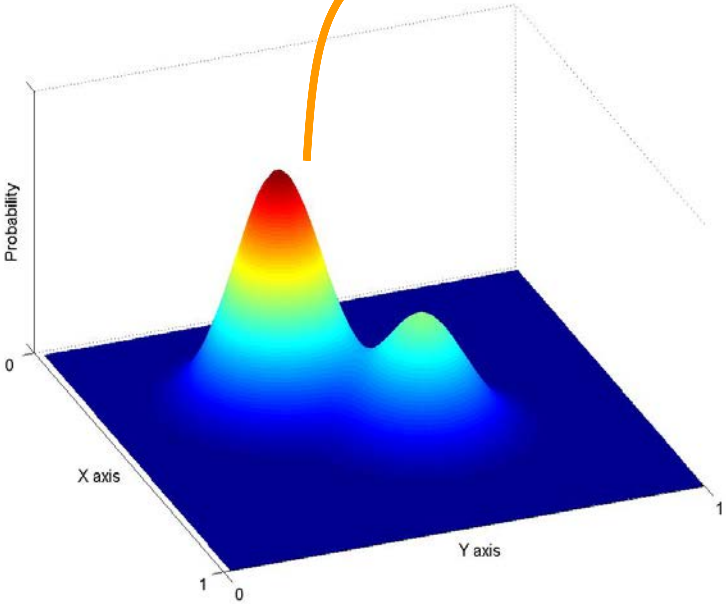
Assumed Underlying PDF

Data Samples

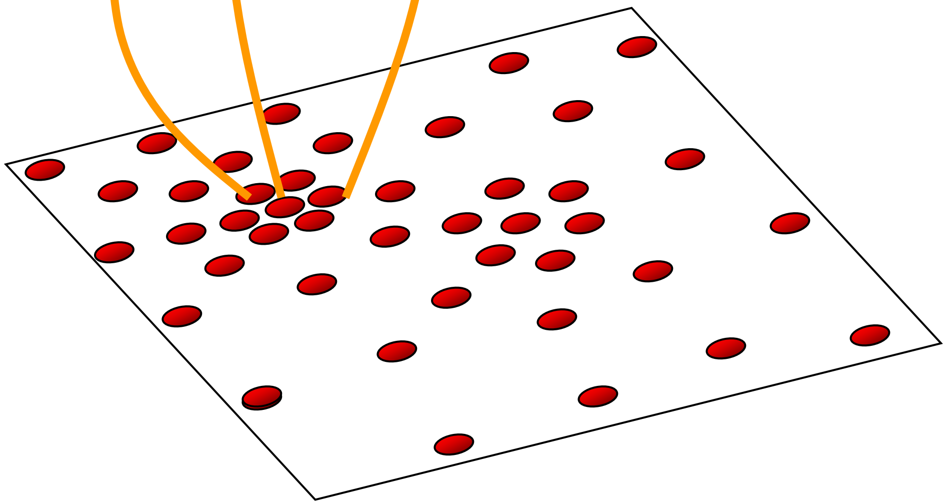
Parametric Density Estimation

Mixture of Gaussians or Gaussian Mixture Model (GMM)

$$p(\mathbf{x}) = \sum_{i=1}^{n} \pi_i e^{-\frac{(\mathbf{x}-\mu_i)^2}{2\sigma_i^2}}$$



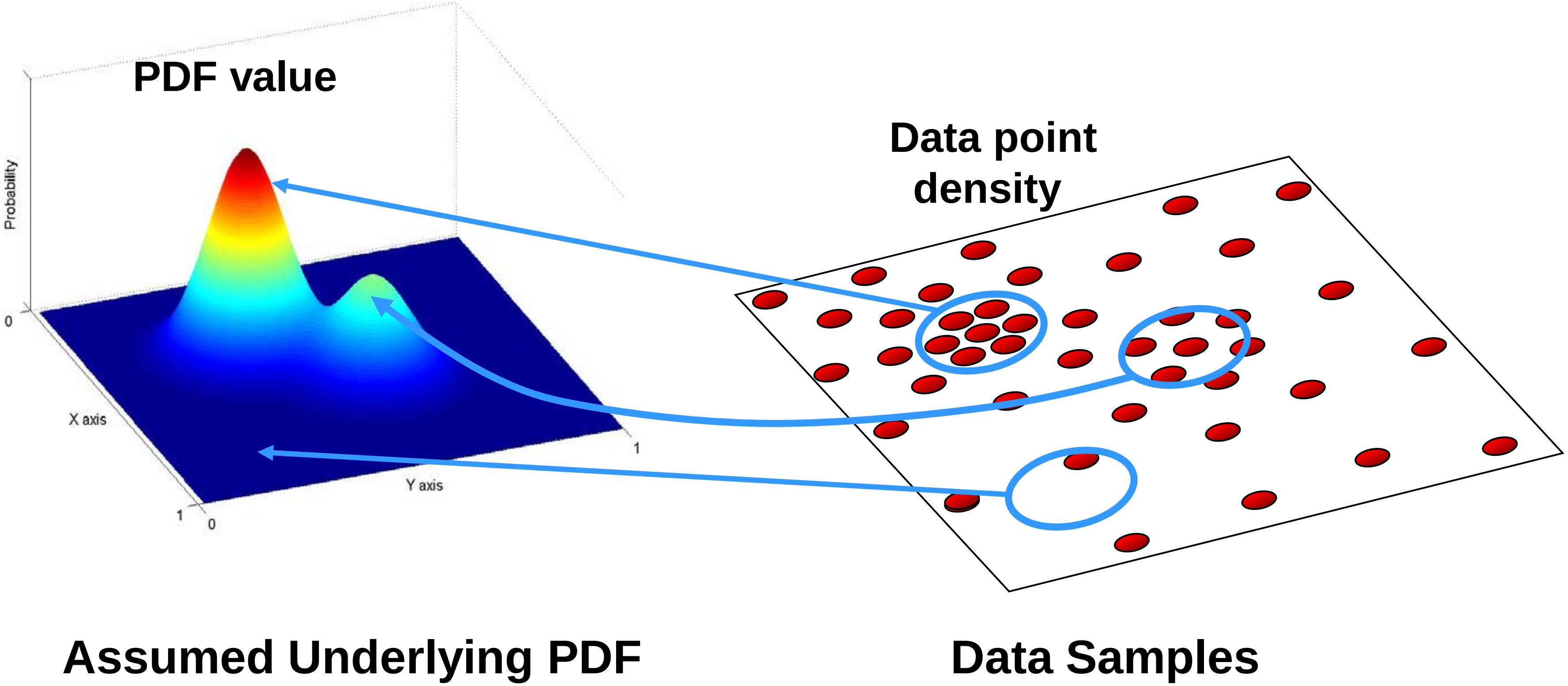
Estimate from data



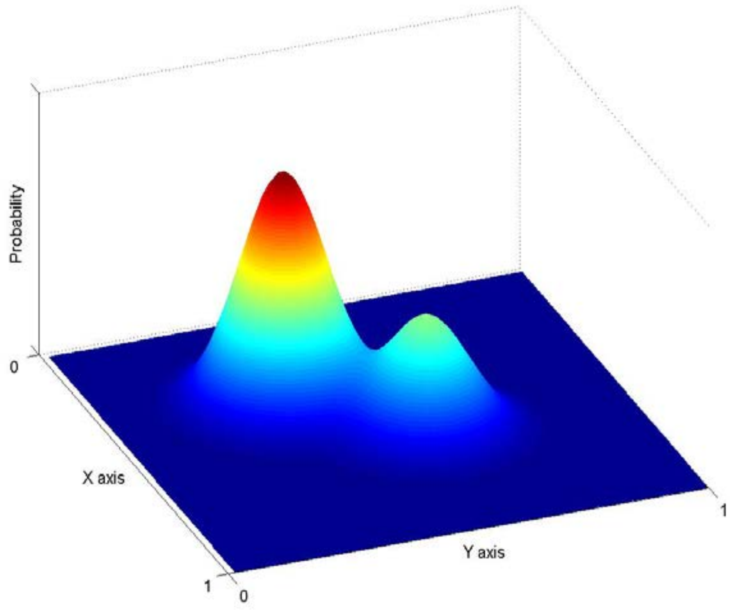
Assumed Underlying PDF

Data Samples

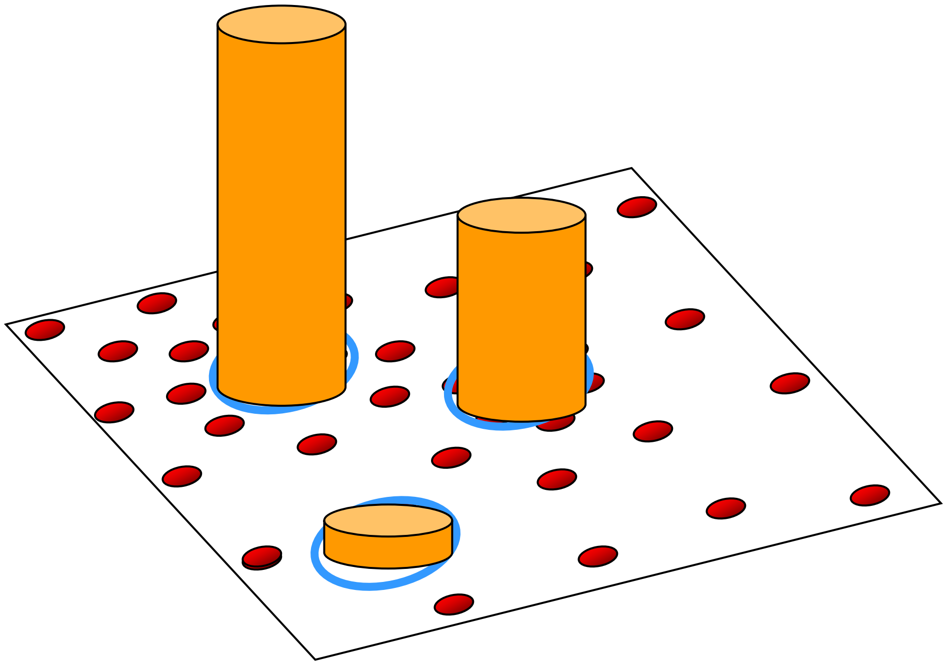
Non-parametric Density Estimation



Non-parametric Density Estimation



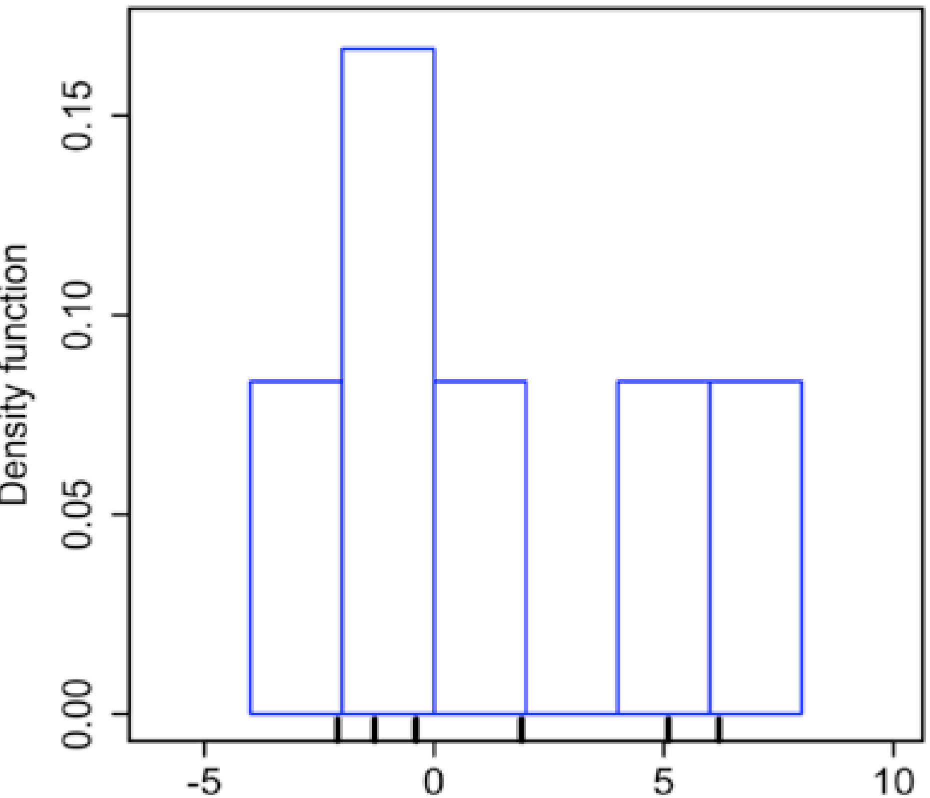
Assumed Underlying PDF

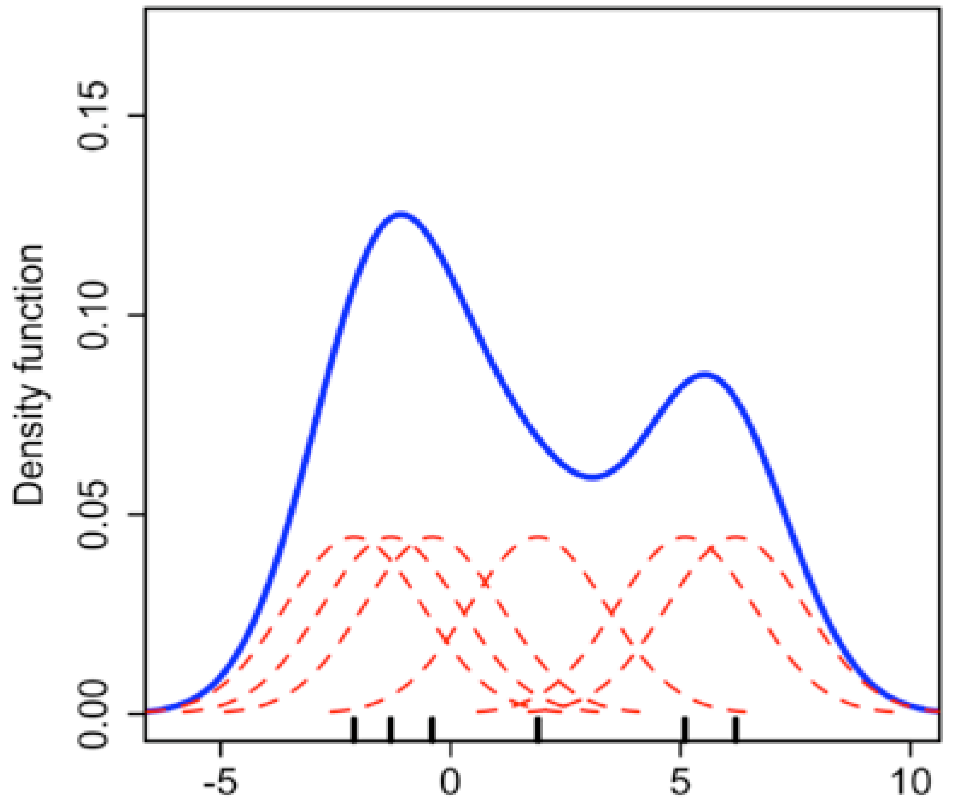


Data Samples

Non-parametric Density Estimation

- 1. Histogram
- 2. Kernel Density Estimation (KDE)



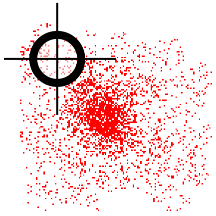


Kernel Density Estimation (KDE)

Parzen Windows - General Framework

$$P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{x} - \mathbf{x}_i)$$

A function of some finite number of data points $x_1 \dots x_n$



Kernel Properties:

- Normalized

$$\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1$$

- Symmetric

$$\int_{R^d} \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0$$

- Exponential weight decay

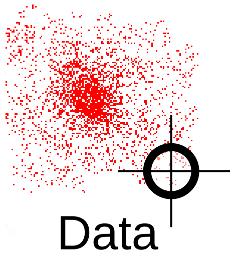
$$\lim_{\|\mathbf{x}\|\to\infty} \|\mathbf{x}\|^d K(\mathbf{x}) = 0$$

Kernel Density Estimation

Various Kernels

$$P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{x} - \mathbf{x}_i)$$

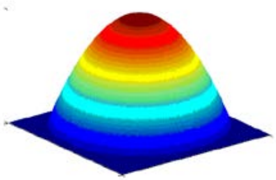
A function of some finite number of data points $x_1 \ldots x_n$



Examples:

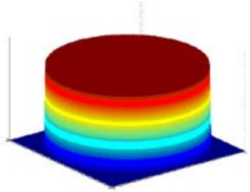
- Epanechnikov Kernel

$$K_E(\mathbf{x}) = \begin{cases} c\left(1 - \|\mathbf{x}\|^2\right) & \|\mathbf{x}\| \le 1 \\ 0 & \text{otherwise} \end{cases}$$



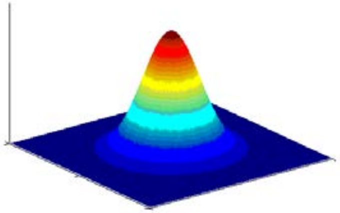
- Uniform Kernel

$$K_U(\mathbf{x}) = \begin{cases} c & \|\mathbf{x}\| \le 1 \\ 0 & \text{otherwise} \end{cases}$$



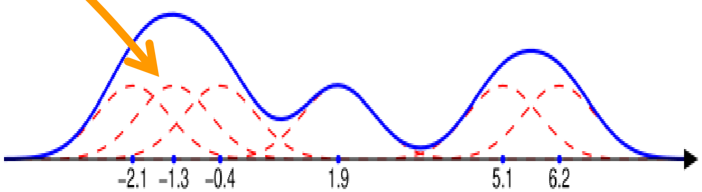
- Normal Kernel

$$K_N(\mathbf{x}) = c \cdot \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)$$



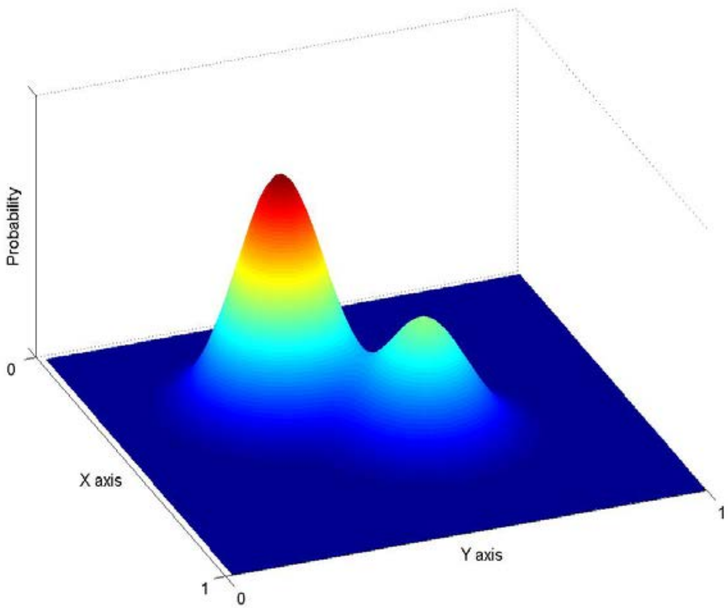
Bandwidth

$$\hat{f}(\mathbf{x}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)$$

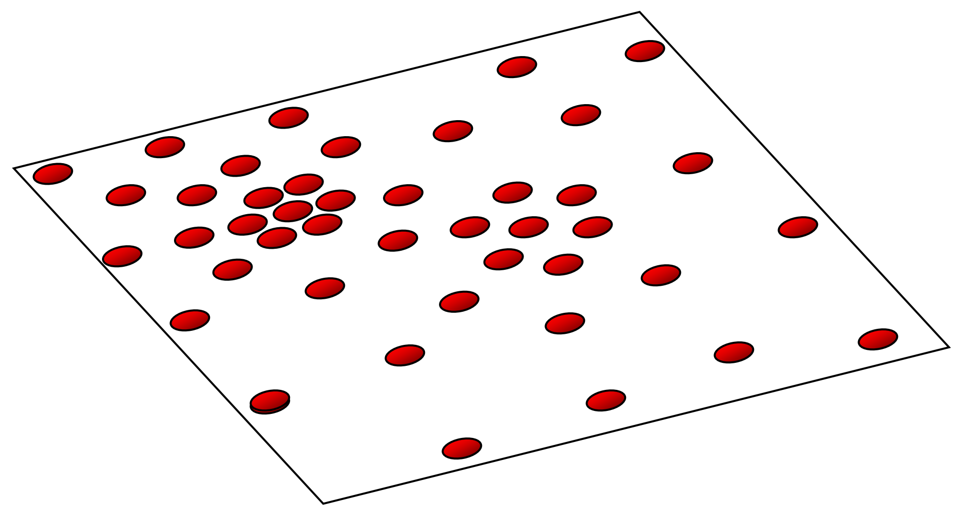


$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

Mode Seeking or “Bump Finding”



Assumed Underlying PDF



Data Samples

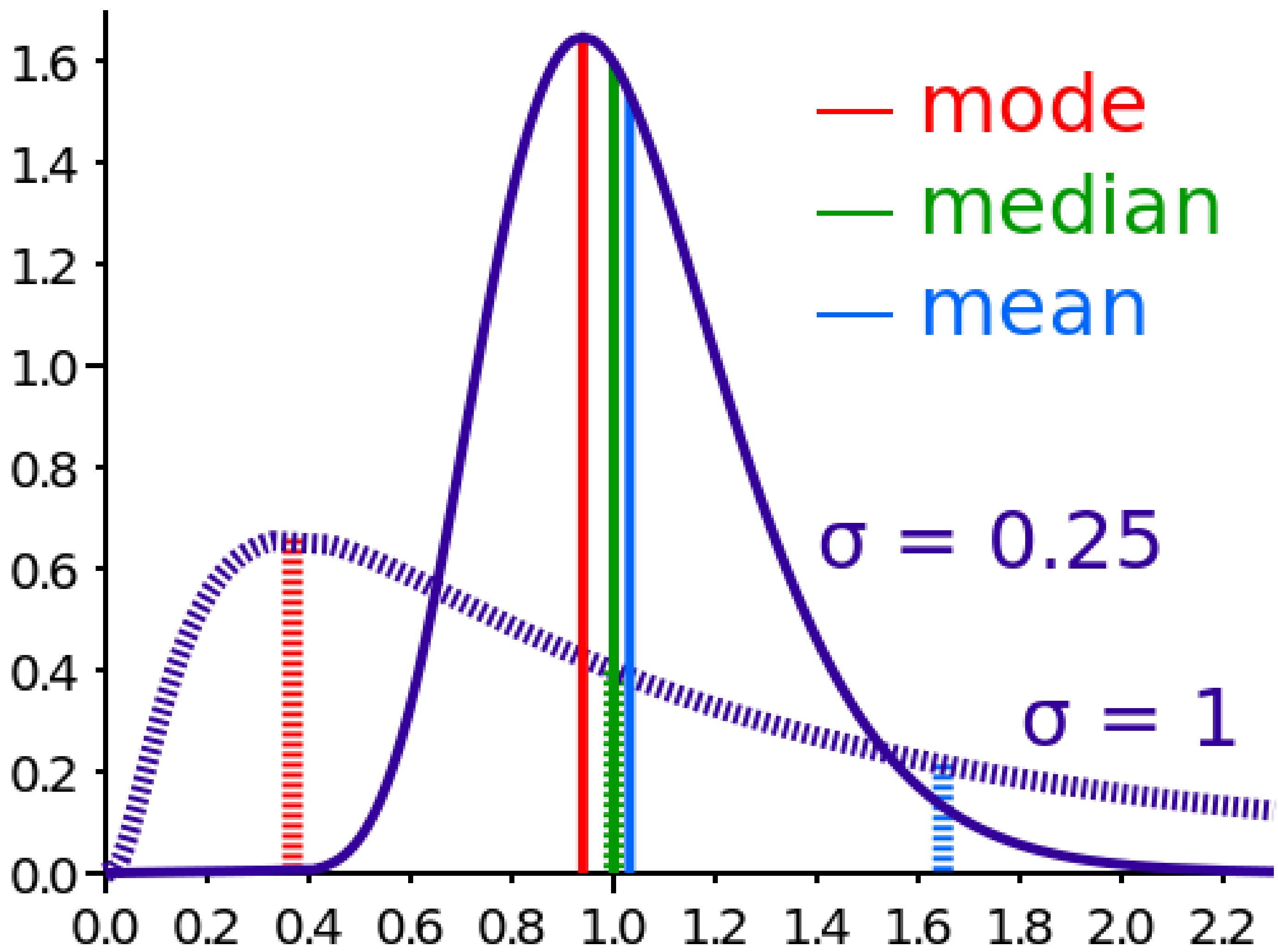
Definition of "Mode"

The **mode** is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value *x* at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled. The mode of a continuous probability distribution is the value *x* at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

When a probability density function has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution. Such a continuous distribution is called multimodal (as opposed to unimodal).

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**



mode
median
mean
σ = 0.25
σ = 1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

Mode Seeking in Ordinal Data

- Lots of work in database community:
  - “Association Rules”
  - “Frequent Itemtsets”
  - “Basket Analysis”
- Sometimes called “Data Mining”

- Basic Idea: discover “interesting” modes in the data

Association Rules

- Rule X→Y
  - Rule form: "Body → Head [support, confidence]"
  - e.g. {butter,bread} → {milk}
- support
  - supp(X) = frequency, i.e. P(X)
    - supp({milk,bread,butter})=20%
- Confidence
  - conf(X→Y) = $\frac{supp(X \cup Y)}{supp(X)}$, i.e. P(Y|X)
    - conf({butter,bread}→{milk}) = 100%
- Lift
  - Lift(X→Y) = $\frac{supp(X \cup Y)}{supp(X) * supp(Y)}$

Example database with 4 items and 5 transactions

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Examples

- Rule form: "Body $\rightarrow$ Head [support, confidence]".
- buys(x, "diapers") $\rightarrow$ buys(x, "beers") [0.5%, 60%]
- major(x, "EECS") ^ takes(x, "ML") $\rightarrow$ GPA(x, "A-") [5%, 75%]

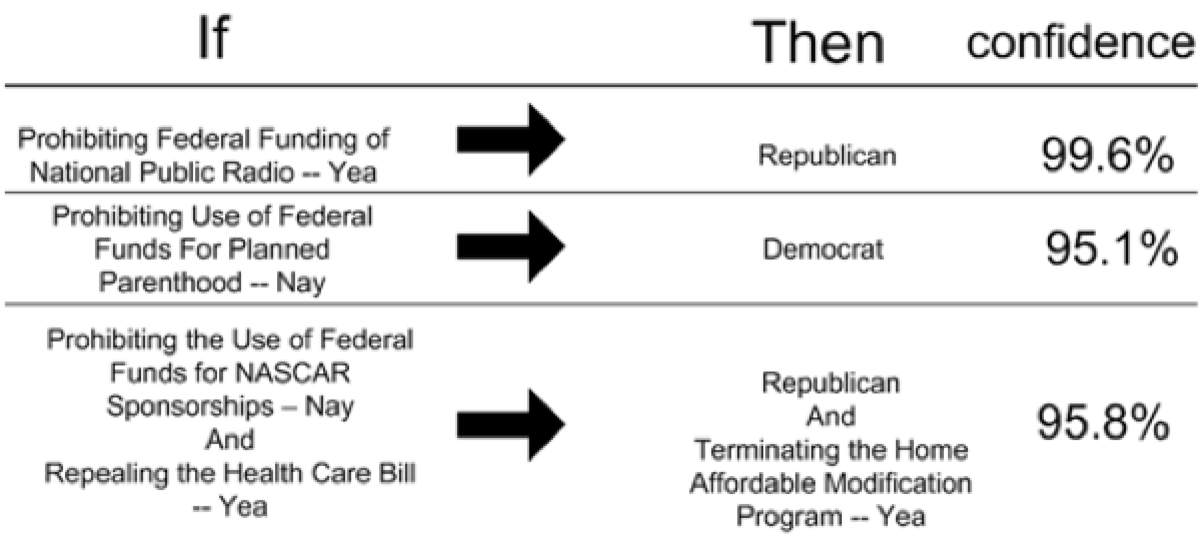


Figure 11.6 Association rules {3} → {0}, {22} → {1}, and {9,26} → {0,7} with their meanings and confidence levels

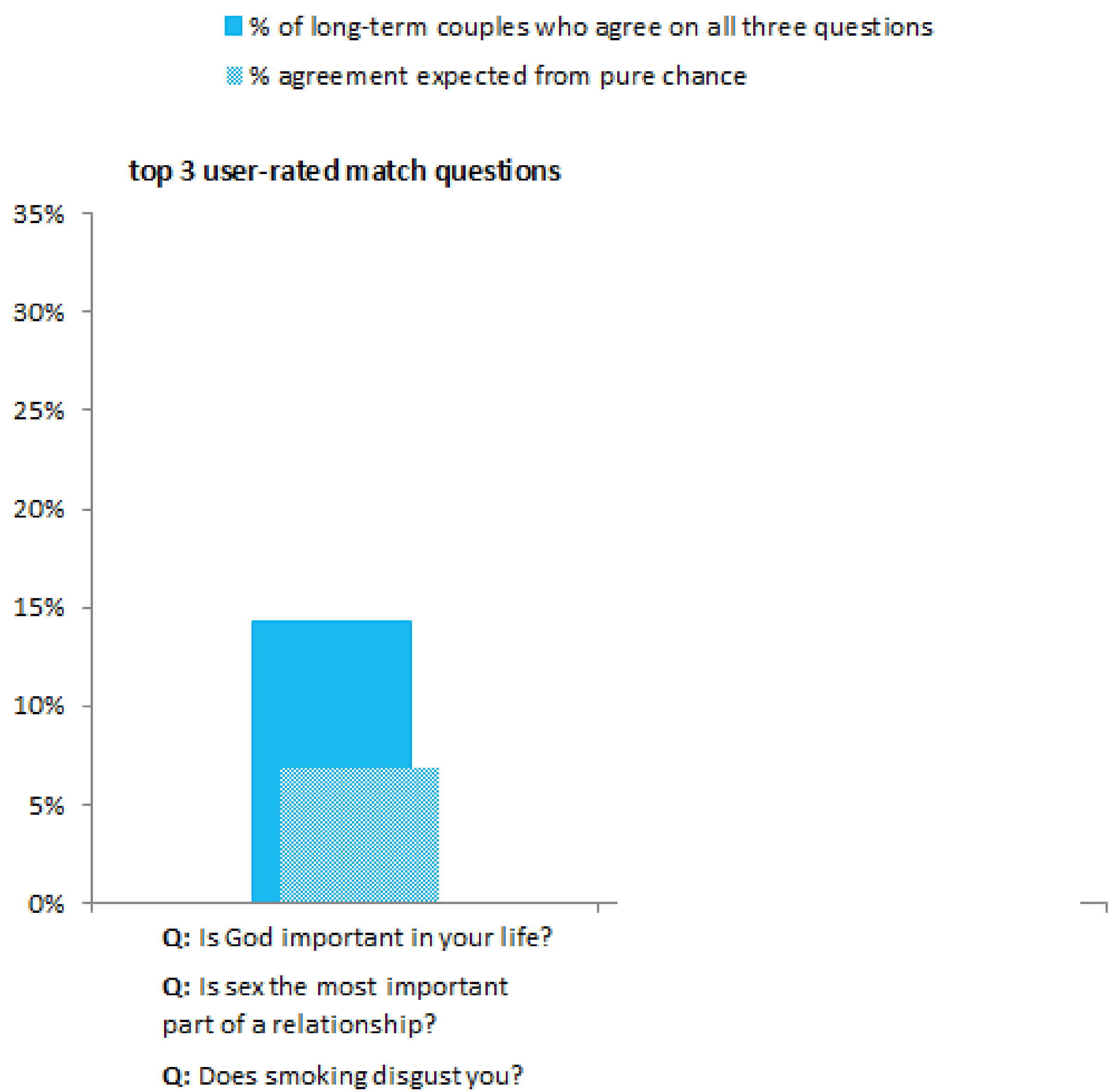
Data from Project Vote Smart (http://www.votesmart.org)

[Harrington]

Real-world Example from OKCupid

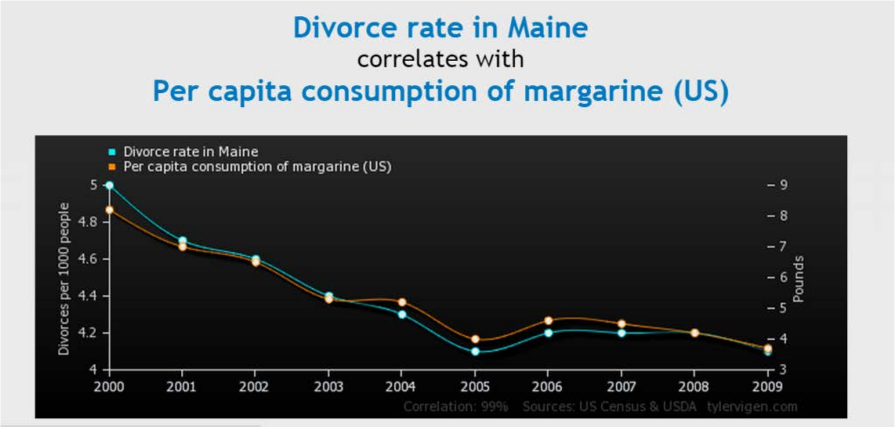
http://oktrends.okcupid.com/

Does a date have long-term potential?


% of long-term couples who agree on all three questions
% agreement expected from pure chance
top 3 user-rated match questions
35%
30%
25%
20%
15%
10%
5%
0%
Q: Is God important in your life?
Q: Is sex the most important part of a relationship?
Q: Does smoking disgust you?

Spurious Rules

- For 10,000 items, there are ~10^12 *“(a,b)=>c”* rules
  - For p-value 0.05 (5%), we expect 10^10 spurious rules!



http://www.tylervigen.com/