Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

Associate Professor

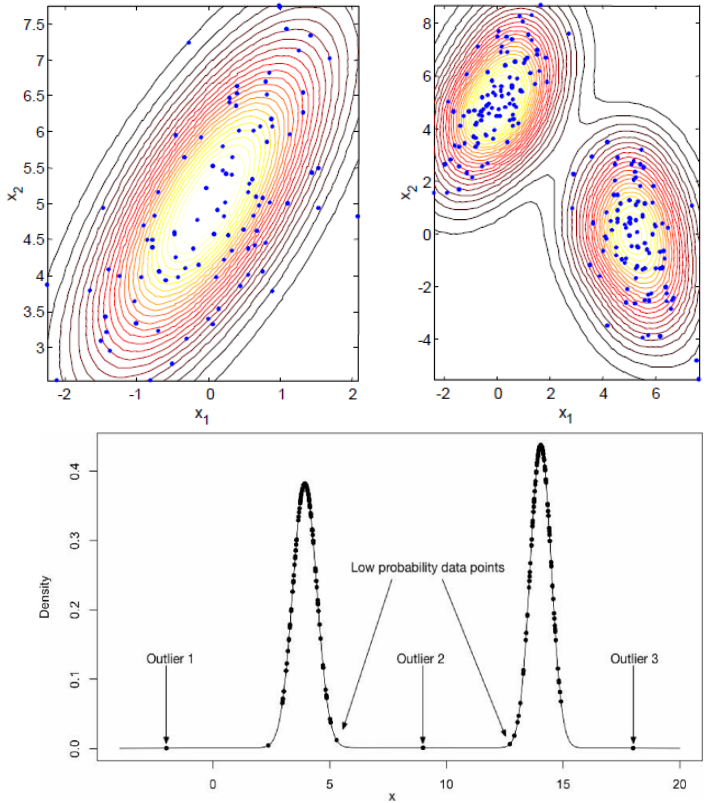
Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering

Density Estimation

Why do we need density estimation?

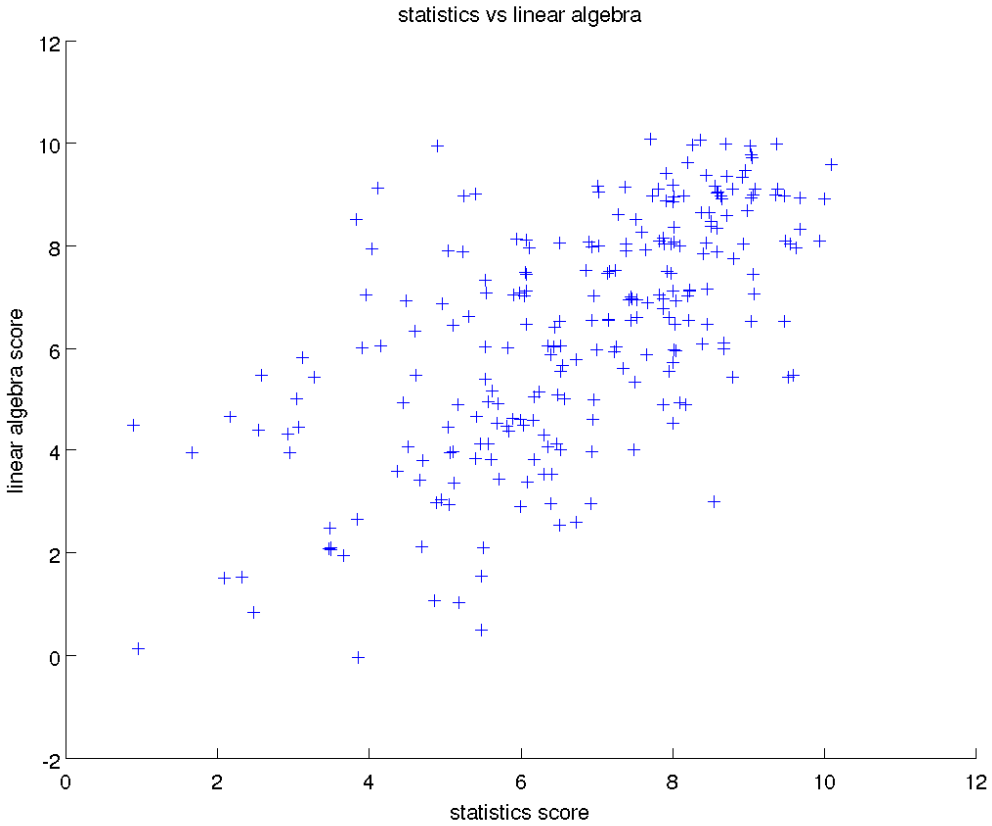
- Learn more about the “shape” of the data cloud
- Assess the likelihood of a particular data point
  - Is this a typical data point? (high density value)
  - Is this an abnormal data point / outlier? (low density value)
- Probabilistic model
  - Making prediction and inference
- Building block for more sophisticated learning algorithms
  - Classification, regression, graphical models …
  - A simple recommendation system
  - Anomaly detection



A. Reddy et al., "Using Gaussian Mixture Models to Detect Outliers in Seasonal Univariate Network Traffic," 2017 IEEE Security and Privacy Workshops (SPW), San Jose, CA, 2017, pp. 229-234, doi: 10.1109/SPW.2017.9.

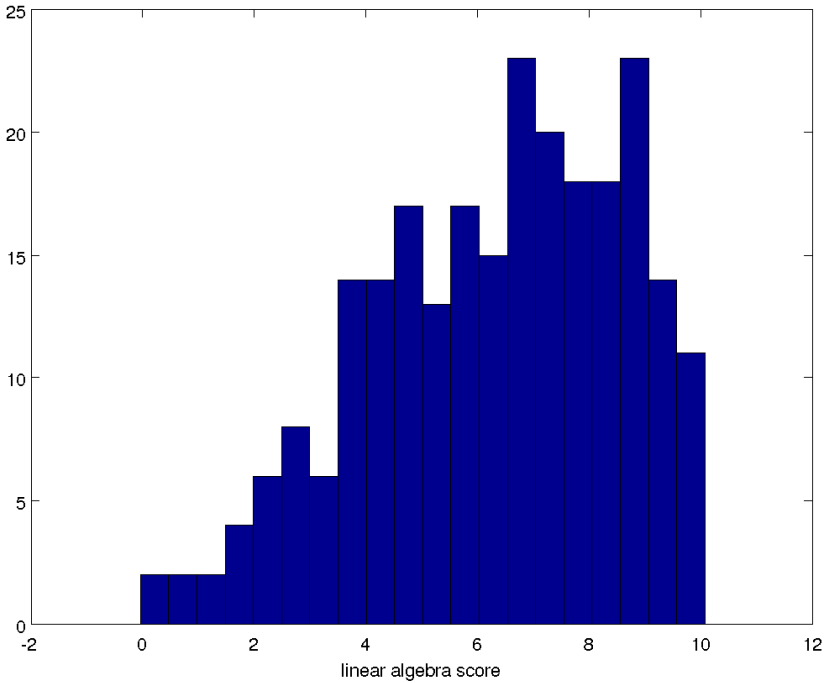
http://www.datawrangling.org/python-montage-code-for-displaying-arrays/

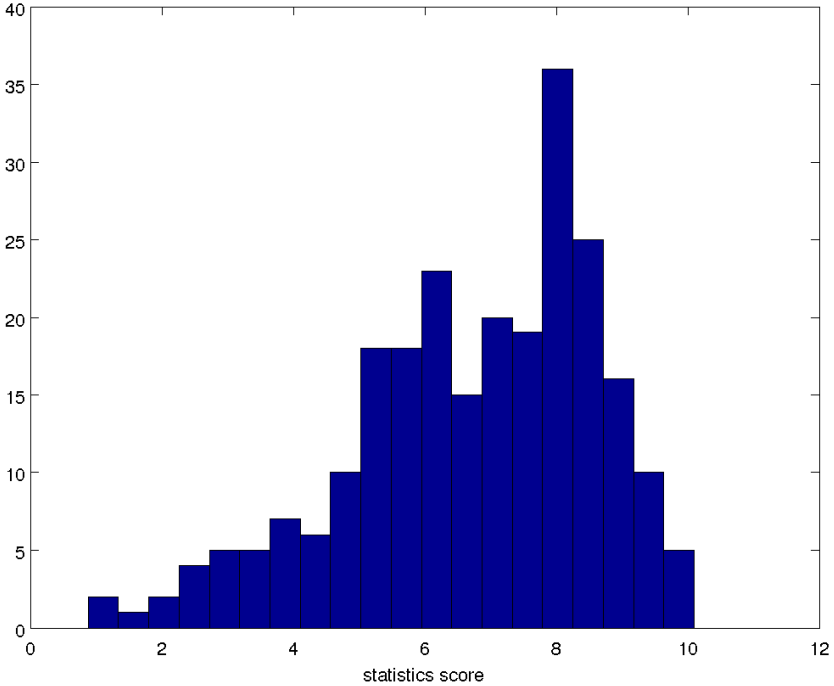
Example: test scores in a class



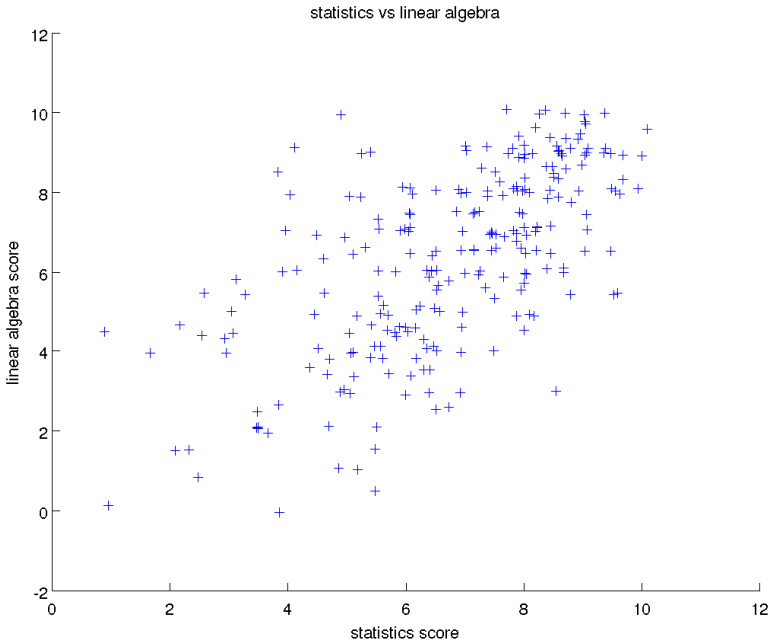
https://library.law.yale.edu/news/library-tip-exam-prep-resources-1

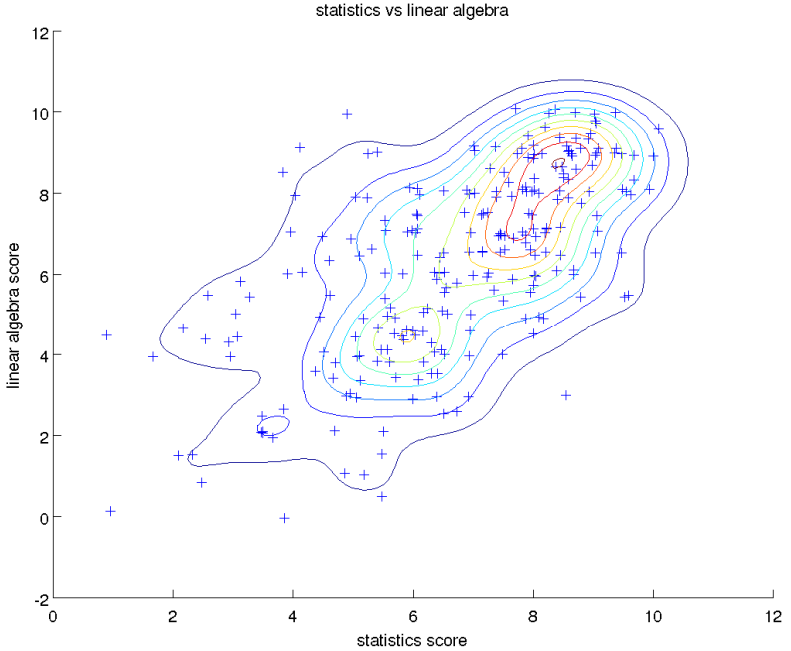
Example: test scores in a class





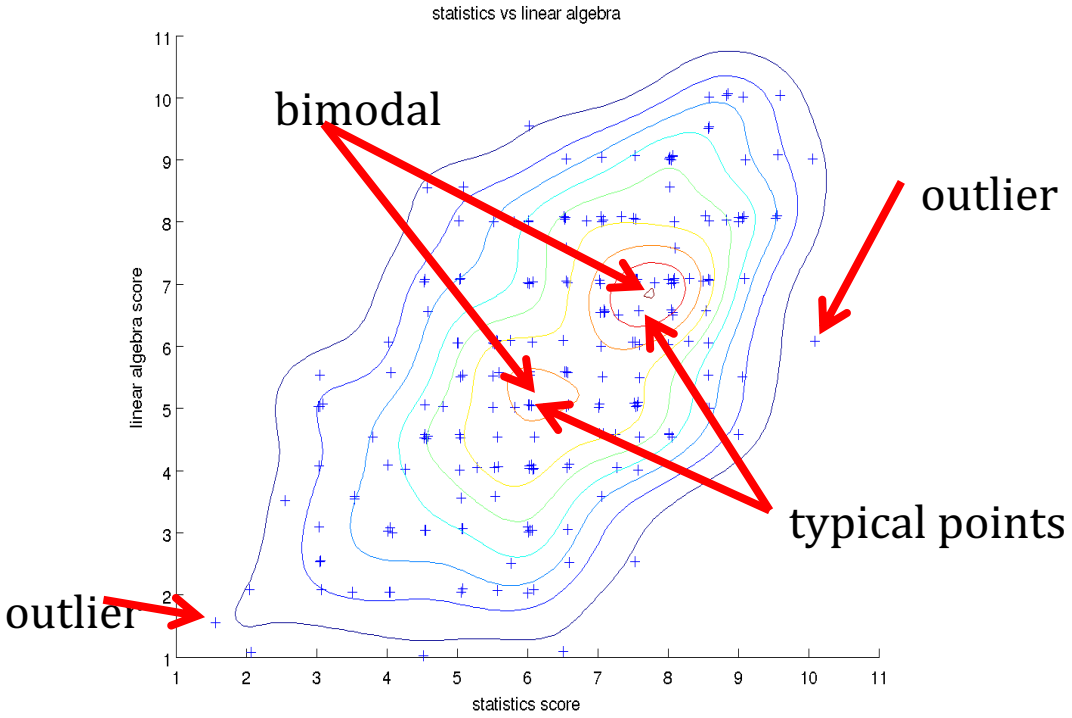
Example: test scores in a class (cont.)





Example: Understand density estimation results

Scatter plot of statistic vs. linear algebra scores, overlaid with contour plot of the density



Parametric versus non-parametric models

- Parametric: e.g. Gaussian distribution in R^n

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

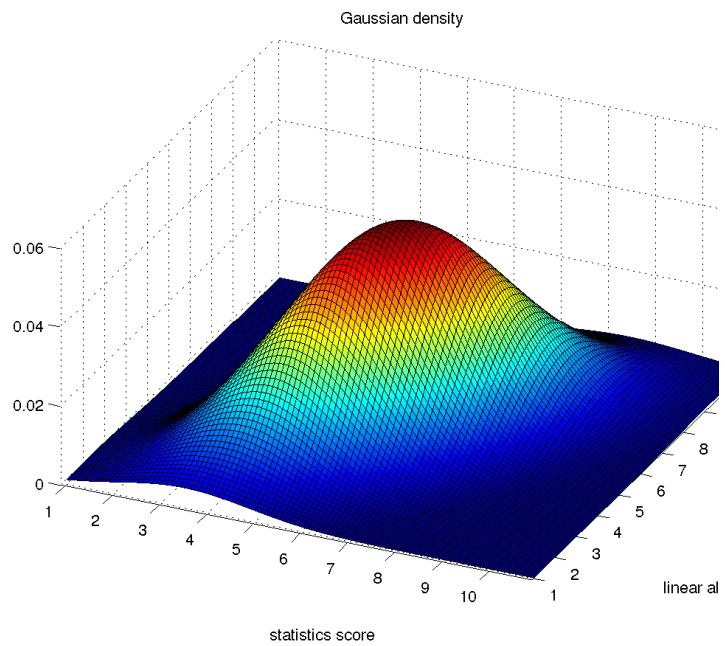
  Two sets of parameters $\{\mu, \Sigma\}$, which again generate a family of models

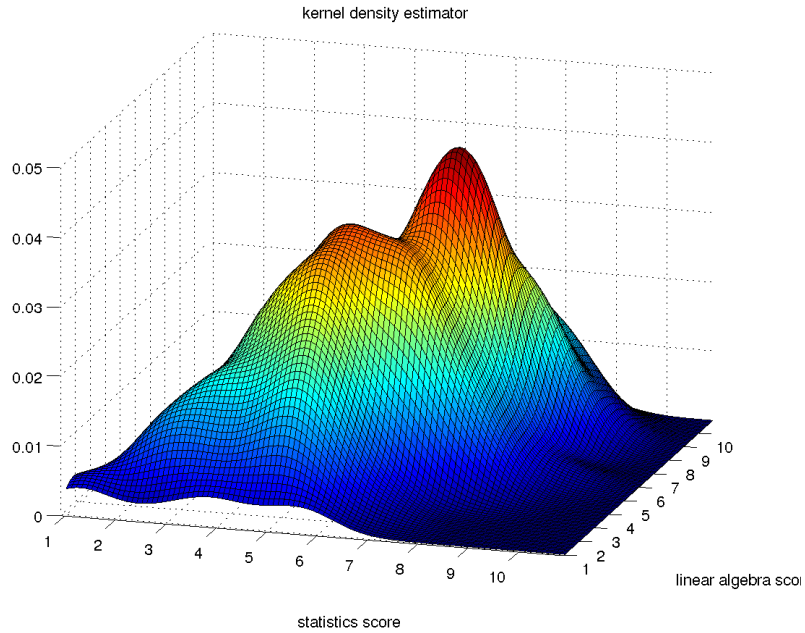
$$\mathcal{F} = \{p(x|\mu,\Sigma) \mid \mu \in R^n, \Sigma \in R^{n\times n} \; and \; PSD\}$$

- Nonparametric e.g. Histogram, Kernel density estimator

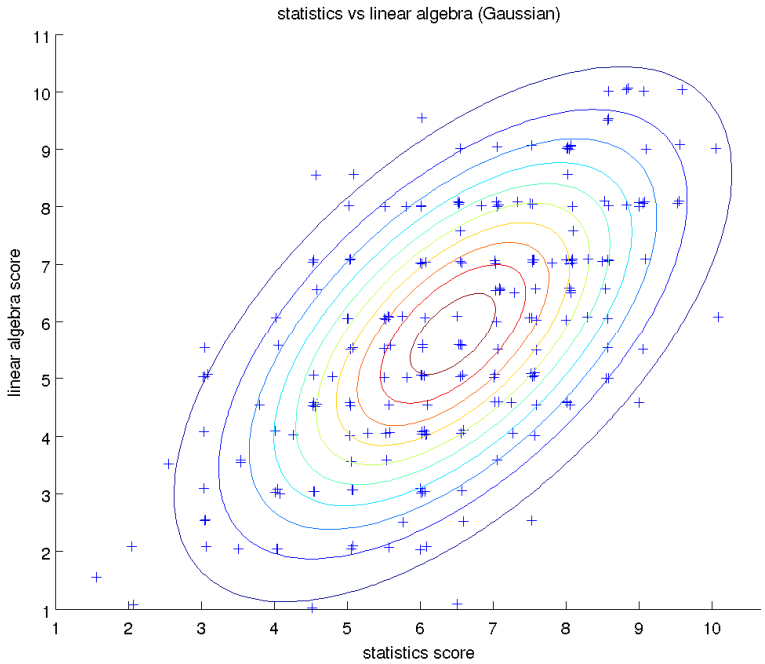
Parametric vs. nonparametric

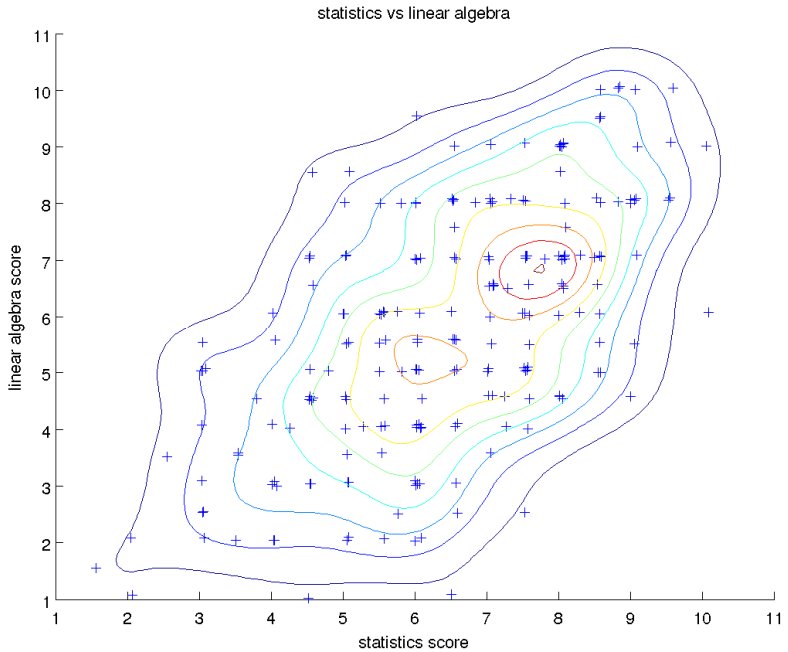
- Surface plot of density for statistic and linear algebra scores





Parametric vs. nonparametric





Parametric models

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

- Discrete case: Bernoulli distribution

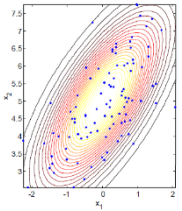
$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

one parameter, $\theta \in [0,1]$, which generate a family of models

$\mathcal{F} = \{P(x|\theta) \mid \theta \in [0,1]\}$,

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

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



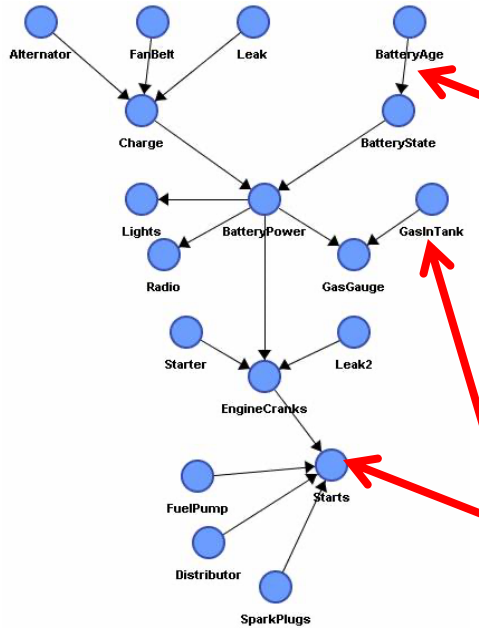
Two sets of parameters $\{\mu, \Sigma\}$, which again generate a family of models,

$\mathcal{F} = \{p(x|\mu,\Sigma) \mid \mu \in R^n, \Sigma \in R^{n\times n} \; and \; PSD\}$,

Probabilistic graphical models

- Use graph to describe the relation of many variables
- Reason about some unknown variable given known variables

Graphical models for your car



Each set of parent-child relation in graph corresponds to a density (or distribution)

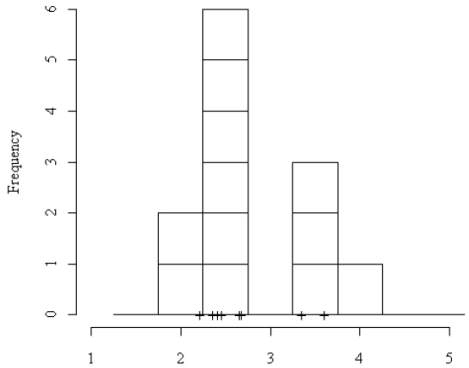
Given the car does not start, and the gas tank 1/10 full, what is the likelihood of leaking?

Nonparametric models

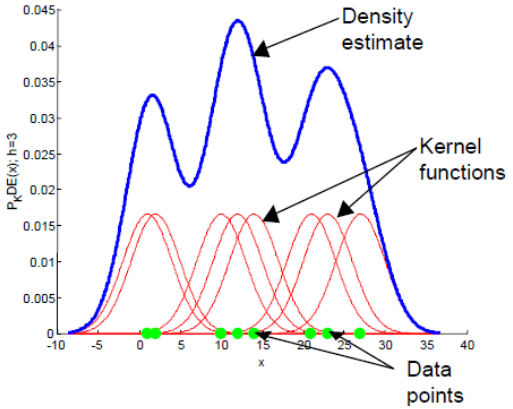
- Smooth density

$$\mathcal{F} = \left\{ p(x) \middle| \, p(x) \geq 0, \int_{\Omega} p(x)dx = 1, \int_{\Omega} \left(p''(x)\right)^2 dx < \infty \right\}$$

- Histogram



- Kernel density estimator



- What are nonparametric models?
  - "nonparametric" does not mean there are no parameters
  - can not be described by a fixed number of parameters

Estimation of 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 $\mathcal{D} = \{x^1, x^2, \ldots x^m\}$ drawn independent and identically distributed (iid) from some unknown distribution $P^*(x)$
- Want to fit the data with a model $P(x|\theta)$ with parameter θ

$$\hat{\theta}_{MLE} = argmax_\theta \log P(\mathcal{D}|\theta)$$
$$= argmax_\theta \log \prod_{i=1}^{m} P(x^i|\theta)$$

Maximum likelihood estimate

- Estimate the probability θ of landing in heads for biased coin
- Given a sequence of m *i.i.d.* flips

$$\mathcal{D} = \{x^1, x^2, \dots x^m\} = \{1,0,1, \dots, 0\}, x^i \in \{0,1\}$$

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

$$P(x|\theta\ ) = \begin{cases} 1-\theta, & x = 0 \\ \theta, & x = 1 \end{cases}$$

- Likelihood of a single observation x_i

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

MLE for biased coin

- Objective function: log likelihood

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

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

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

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

Maximum likelihood estimators for Gaussian distribution

- Gaussian distribution in R

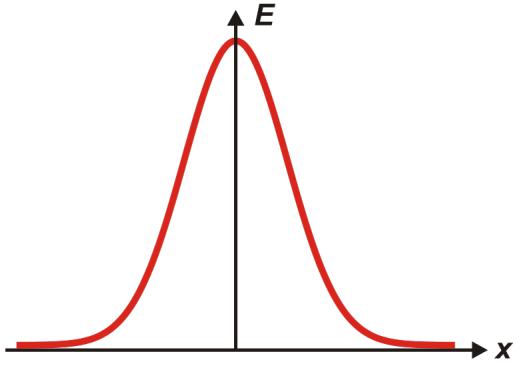
$$p(x|\mu,\sigma) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Need to estimate two sets of parameters μ, σ
- Given m *i.i.d.* samples, $\mathcal{D} = \{x^1, x^2, \dots x^m\}, x^i \in R$ Likelihood of one data point:

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

- Parameter estimate

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



Gaussian MLE

- Objective function, log likelihood

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

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

- Maximize $l(\mu, \sigma; \mathcal{D})$ with respect to μ, σ
- Take derivatives w.r.t. μ, σ^2

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

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

1-D Histogram

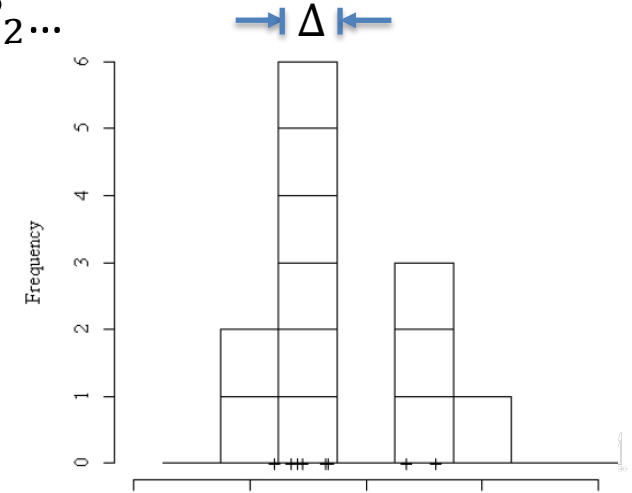
Given m *i.i.d.* samples $\mathcal{D} = \{x^1, x^2, \dots x^m\}, x^i \in [0,1)$ from $P^*(x)$

Split $[0,1)$ evenly into bins of size Δ (assume $1/\Delta$ is integer)

$$B_1 = [0, \Delta), B_2 = [\Delta, 2\Delta), \dots B_{1/\Delta} = [1 - \Delta, 1)$$

Count the number of points, c_1 within B_1, c_2 within B_2…

For a new test point x $p(x) = \frac{1}{\Delta} \sum_{j=1}^{1/\Delta} \frac{c_j}{m} I(x \in B_j)$



Why is histogram valid?

Requirement for probability density function $p(x)$

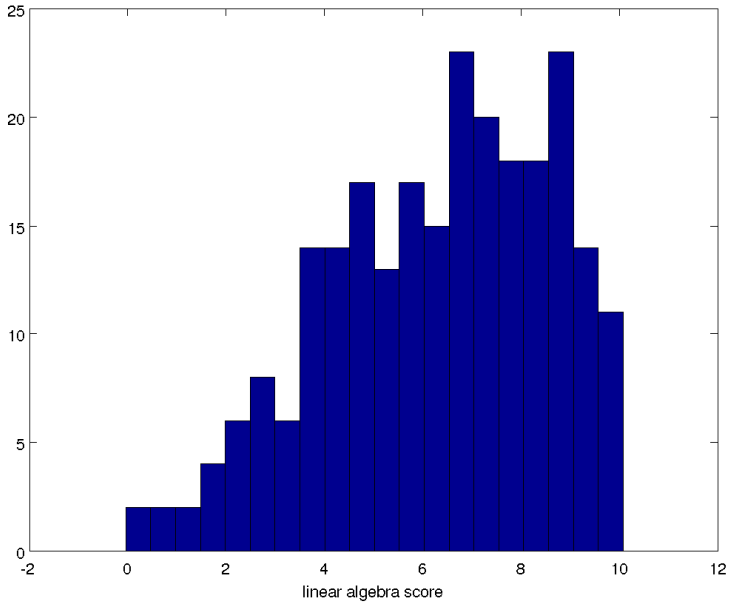
$$p(x) \geq 0, \int_{\Omega} p(x)dx = 1$$

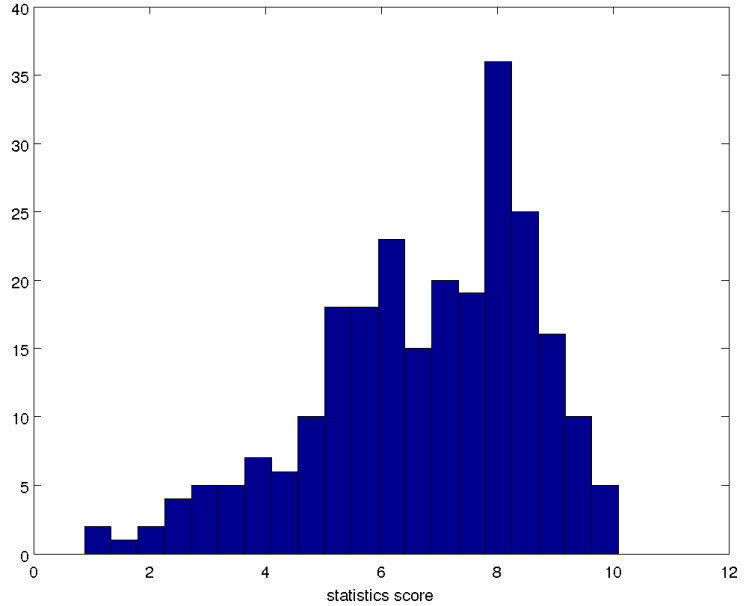
Verify this is satisfied for our histogram,

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

$$= \sum_{j=1}^{1/\Delta} \frac{1}{\Delta} \int_{[(j-1)\Delta, j\Delta)} \frac{c_j}{m} dx = \sum_{j=1}^{1/\Delta} \frac{c_j}{m} = 1$$

Example: test scores





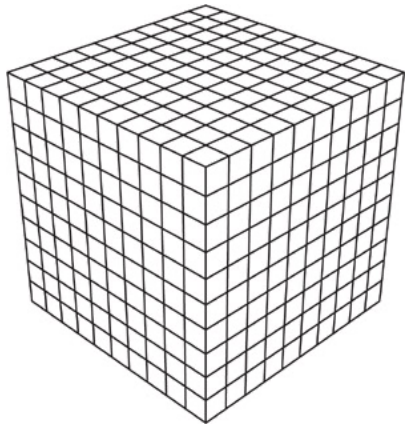
Higher dimensional histogram

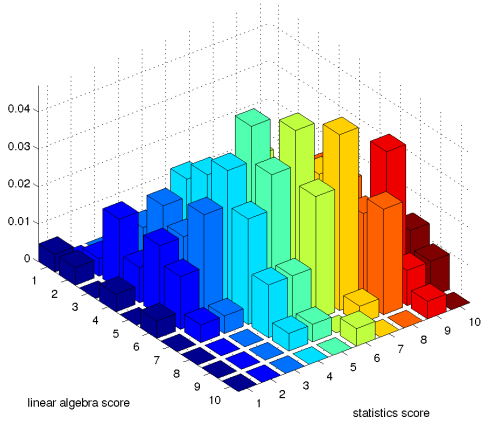
Assume data are n dimensional

Split $[0,1)^n$ evenly into $(1/\Delta)^n$ bins

$$B_1 = [0,\Delta)\times[0,\Delta)\ldots\times[0,\Delta),$$
$$B_2 = [\Delta,2\Delta)\times[0,\Delta)\ldots\times[0,\Delta),$$

$$B_{(1/\Delta)^n} = [1-\Delta,1)\times[1-\Delta,1)\ldots\times[1-\Delta,1)$$



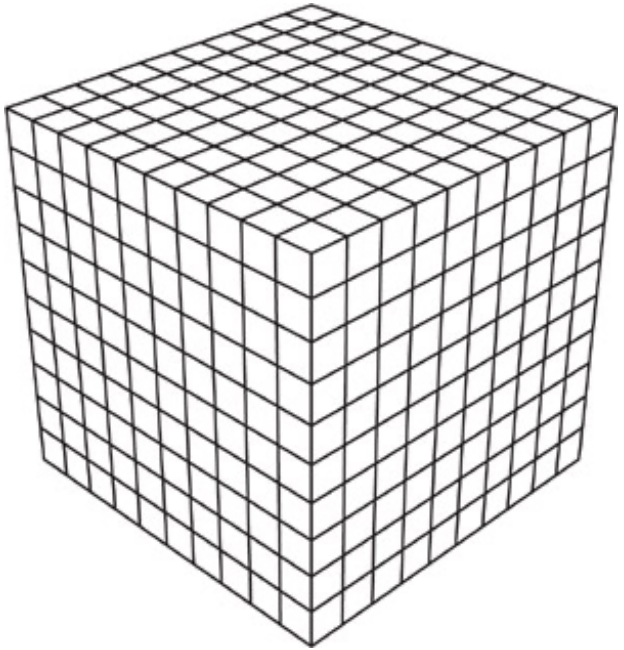


Histogram is not sample efficient for high-dimensional data

For high-dimensional data, too many bins!

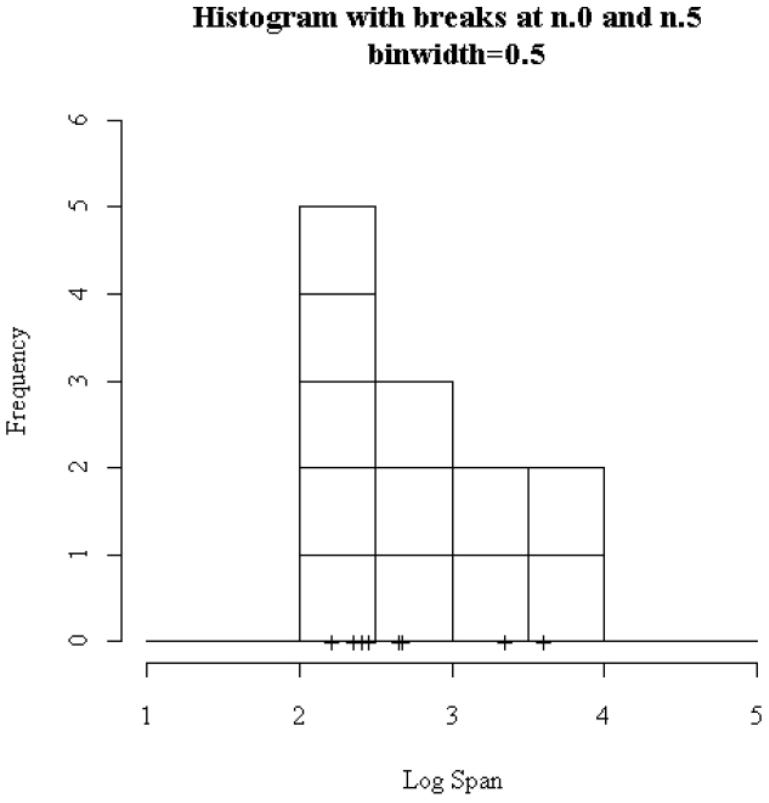
$\Delta = 0.1, n = 6$, need ~ 1 million bins

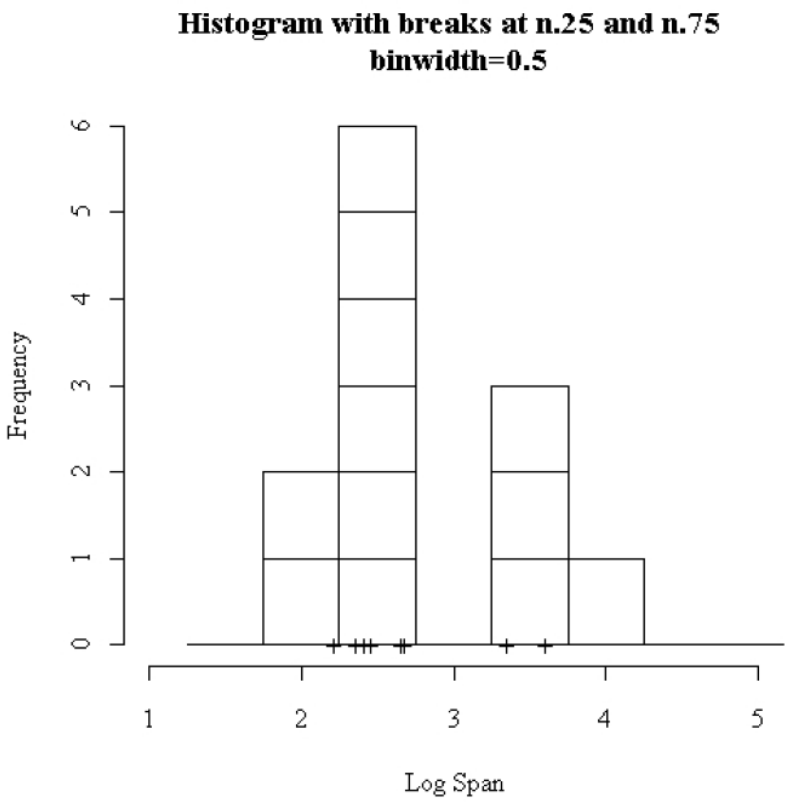
If the number of bins $(\frac{1}{\Delta})^n$ is larger than m (sample size), most bins are empty.



Histogram has some problems

Output depends on where you put the bins: estimates is **noisy**





Kernel density estimation

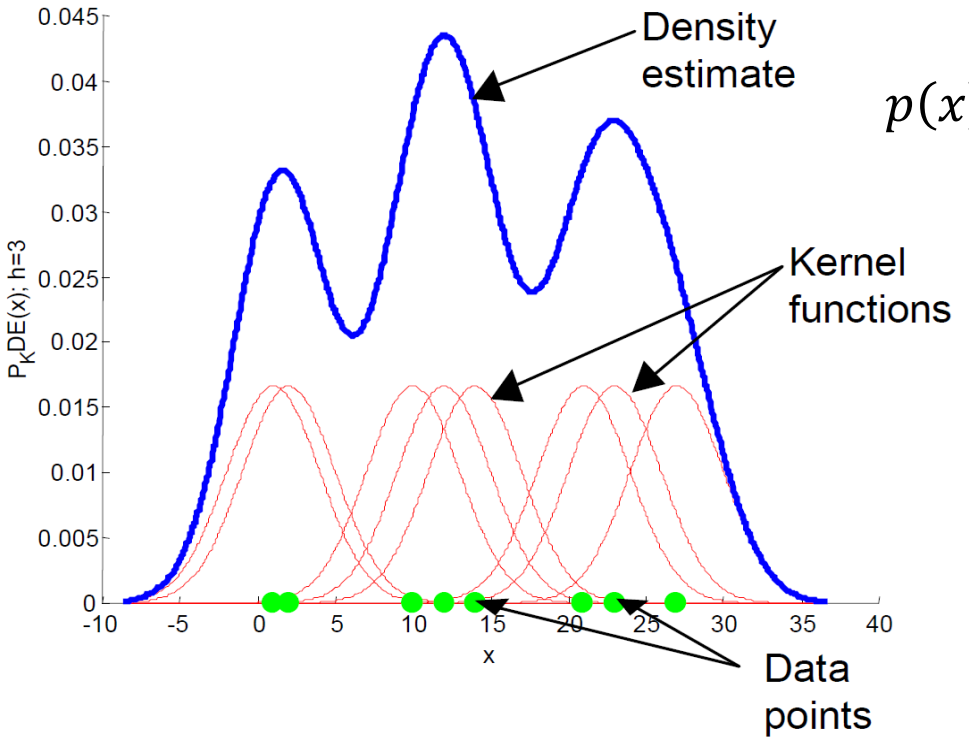
- Kernel density estimator

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

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

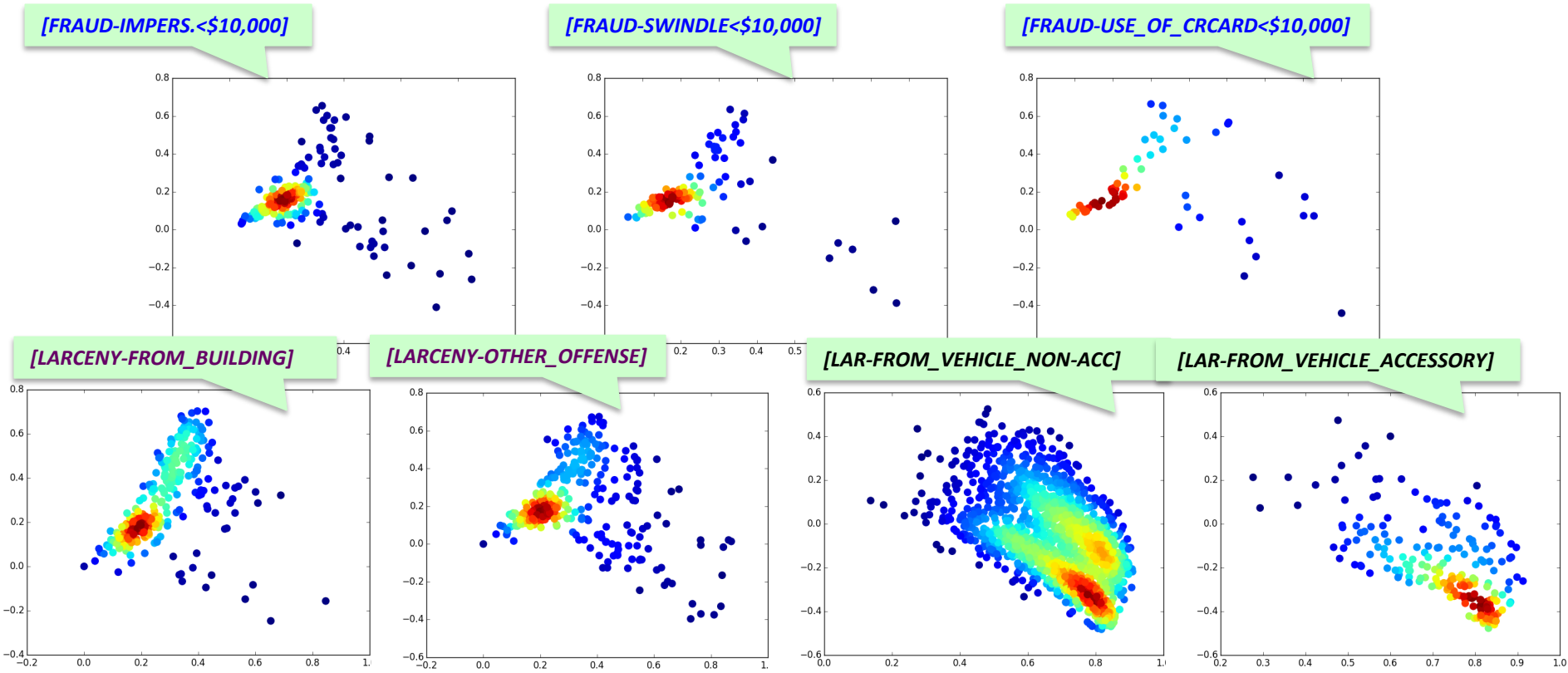
Example



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

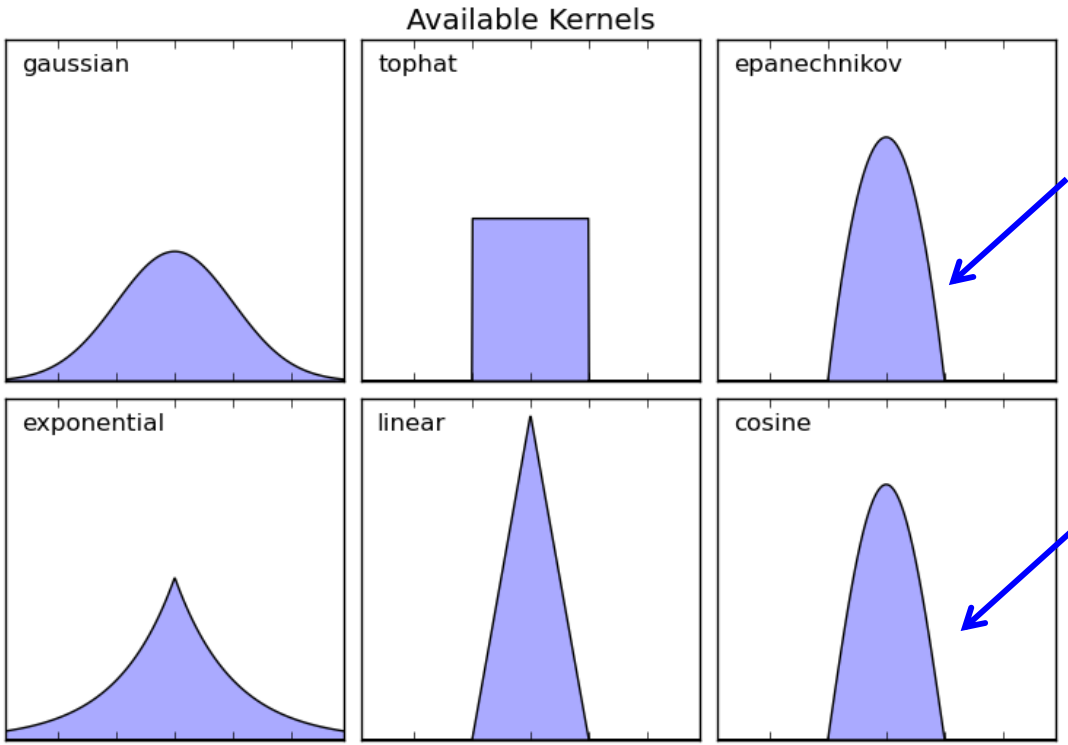
Example: using PCA for data visualization (revisit)

Atlanta police data, 20000 police reports, 7200 keywords (bi-gram), map into 2 principle components; shown 2d density estimation.



Smoothing kernel functions

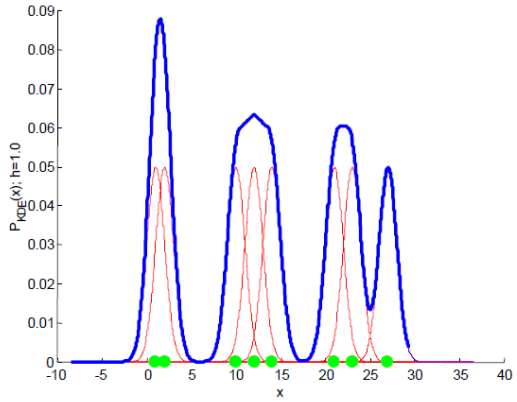
- An example: Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

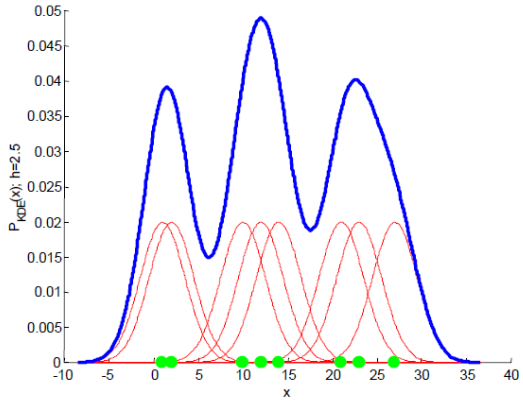


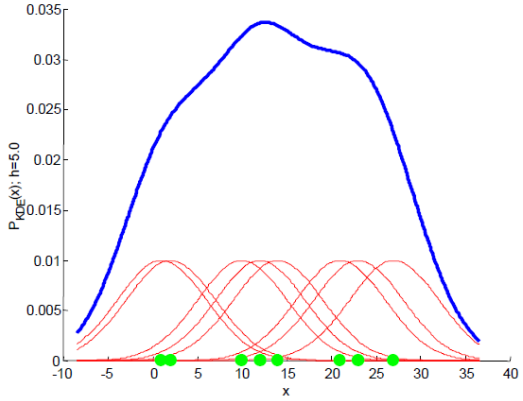
$$K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$$

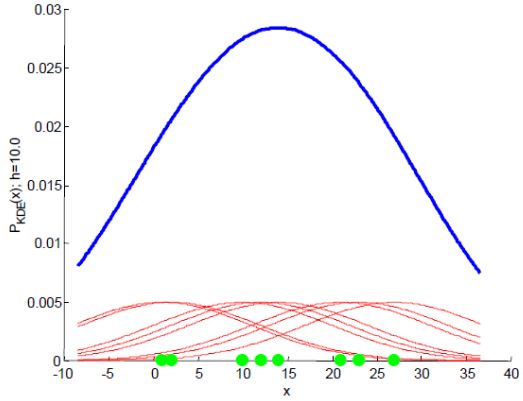
$$K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)I(|u| \leq 1)$$

Effect of the kernel bandwidth h



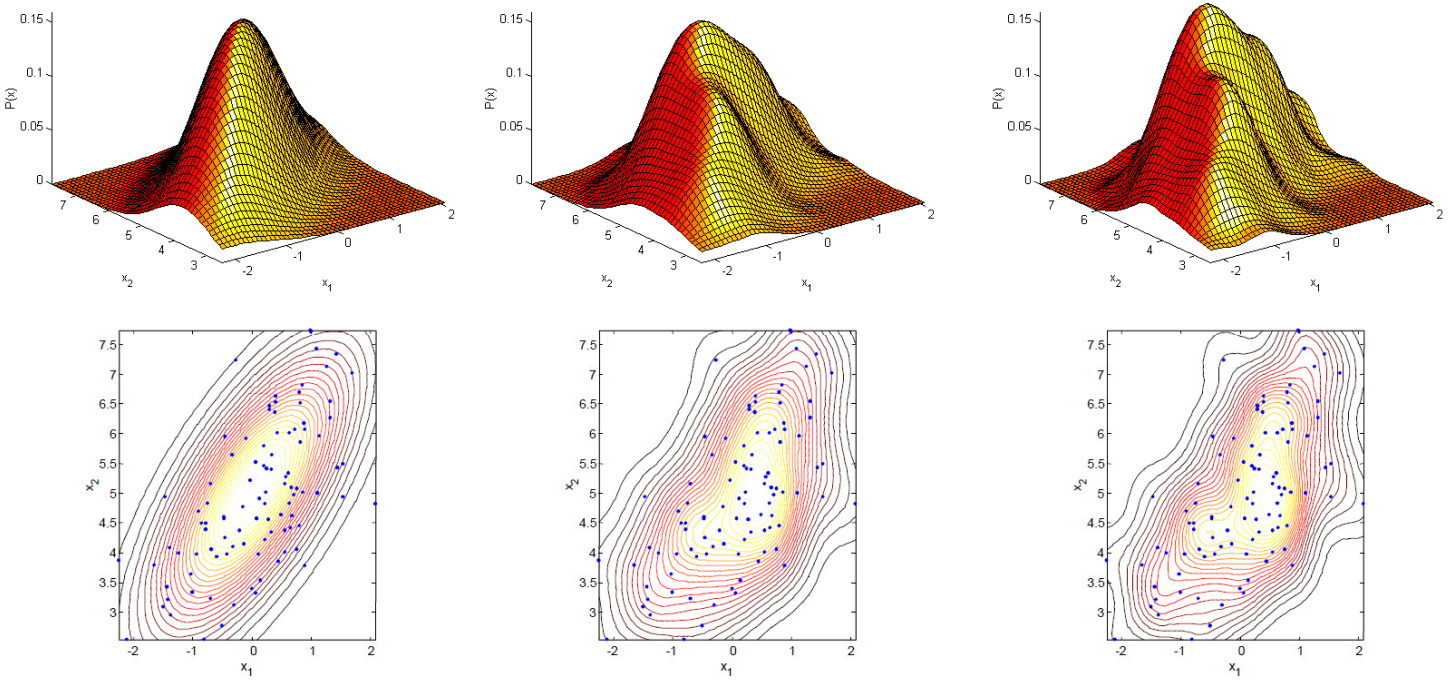






$$p(x) = \frac{1}{m}\sum_{i}^{m}\frac{1}{h}K\left(\frac{x^i - x}{h}\right)$$

Multi-dimensional example



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

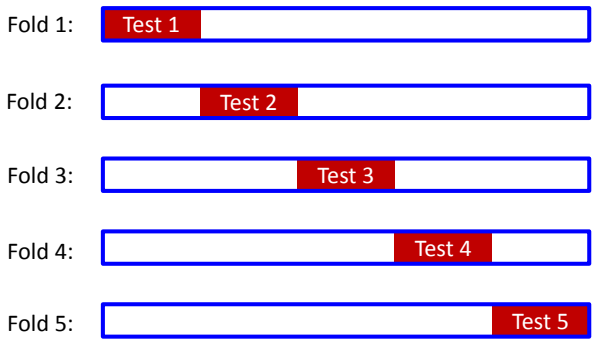
What is the best kernel bandwidth?

- 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 $\hat{\sigma}$ the standard deviation of the samples

- A better but more computationally intensive approach (cross-validation)
  - Randomly split the data into two sets
  - Obtain kernel density estimation using one set
  - Measure the likelihood of the second set
  - Repeat over many splits and average



Wine data example

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) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

http://www.iitaly.org/magazine/dining-in-out/articles-reviews/article/barolo-king-wines-and-wine-kings
http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Some theoretical results

- Consider estimated density function $\hat{p}(x)$ (random, since it depends on data)
- Assume true density function is $p(x)$
- Performance measure: integrated risk

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

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

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

- Kernel density estimator (with bandwidth $h \sim m^{-1/5}$)

$$r(\hat{p}, p) \sim \frac{C}{m^{4/5}}$$

Smaller errors

Difference even big for high dimensional data

Comparison

- Data $x \in R^n$ with fixed dimension n
- Given m training data points $\{x^1, x^2 \ldots, x^m\}$
- Partition into bin of (small) size Δ

Aspects	Gaussian	Histogram	KDE
Flexible	Not	Yes	Yes
Assumption	Strong	Not	Not
Parameter number	Fixed	Increase with $1/\Delta$	Increase with m
Memory requirement	$n + n^2$	$(1\backslash\Delta)^n$	mn
Modeling Learning	Closed form	Binning and Counting	Nothing
Test computation	Plug in formula	Find the bin	Evaluate m functions
Statistical guarantee	only Gaussian case	Arbitrary (worse)	Arbitrary (better)

Gaussian mixture model

A density model $p(X)$ may be multi-modal: model it as a mixture of uni-modal distributions (e.g. Gaussians)

$$\mathcal{N}(X|\mu_k, \Sigma_k) := \frac{1}{|\Sigma_k|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} exp\left(-\frac{1}{2}(X-\mu_k)^{\top}{\Sigma_k}^{-1}(X-\mu_k)\right)$$

Consider a mixture of K Gaussians

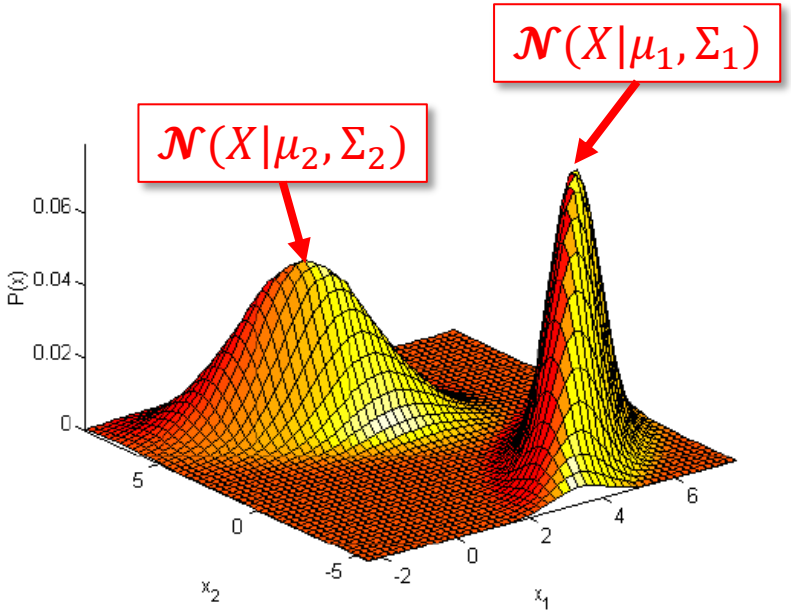
– $p(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu_k, \Sigma_k)$

mixing proportion

mixture Component

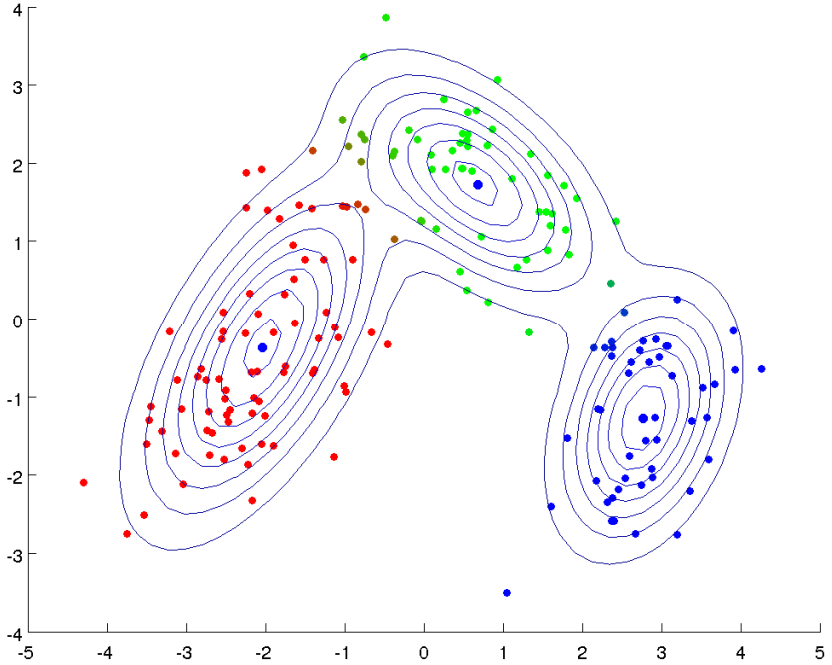
Parametric or nonparametric?

Learn $\pi_k \in (0,1), \mu_k, \Sigma_k$, $k = 1, \dots, K$



Demo: Wine data

- Clear cluster structure, can we fit 3 Gaussians?




SUMMARY