UNSUPERVISED LEARNING

Kristina Lerman

USC Information Sciences Institute

DSCI 552 – Spring 2021

February 8, 2021

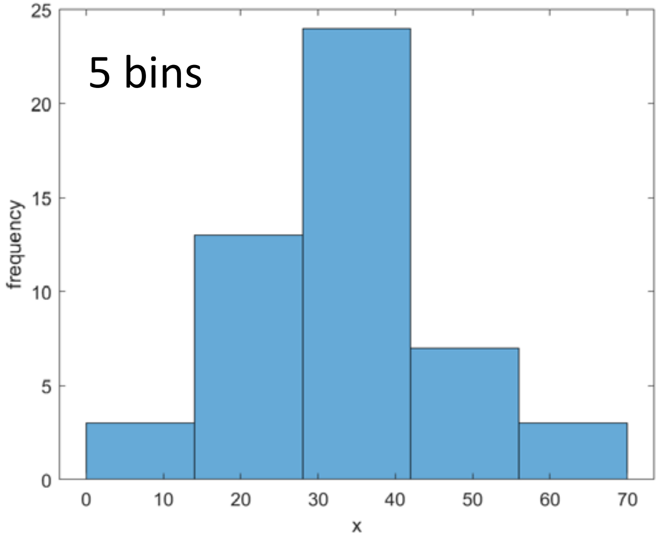
Topics this week

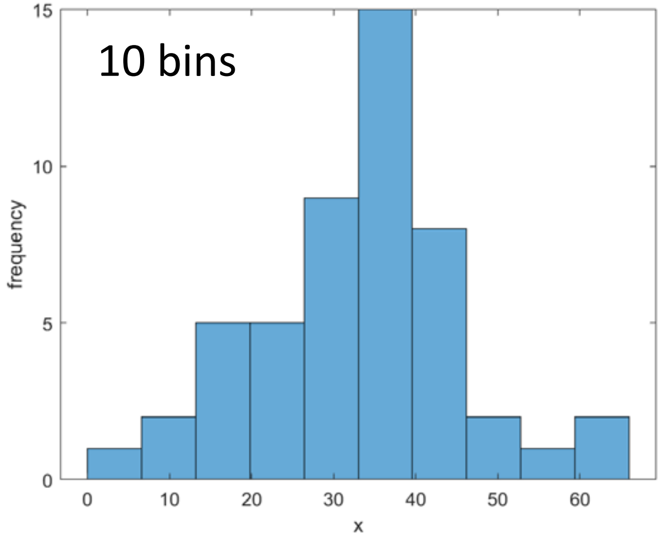
- Reminders:
  - Quiz 3 due today
  - Homework 1 due Thursday
- Unsupervised learning: working with unlabeled data
  - Non-parametric estimation and embedding
- Online office hours following the class (see link on BB or slides)

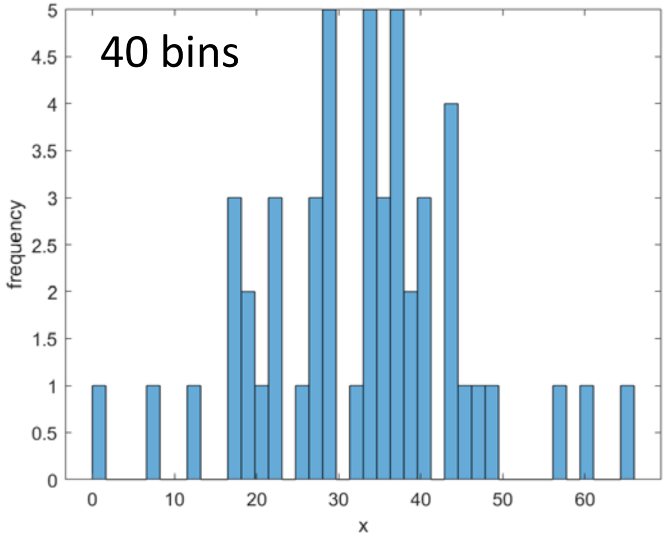
NON-PARAMETRIC ESTIMATION

Histogram

- Histogram is the most popular way to visualize data
- Bin size changes the shape of the histogram
- For normal data, use equal size bins

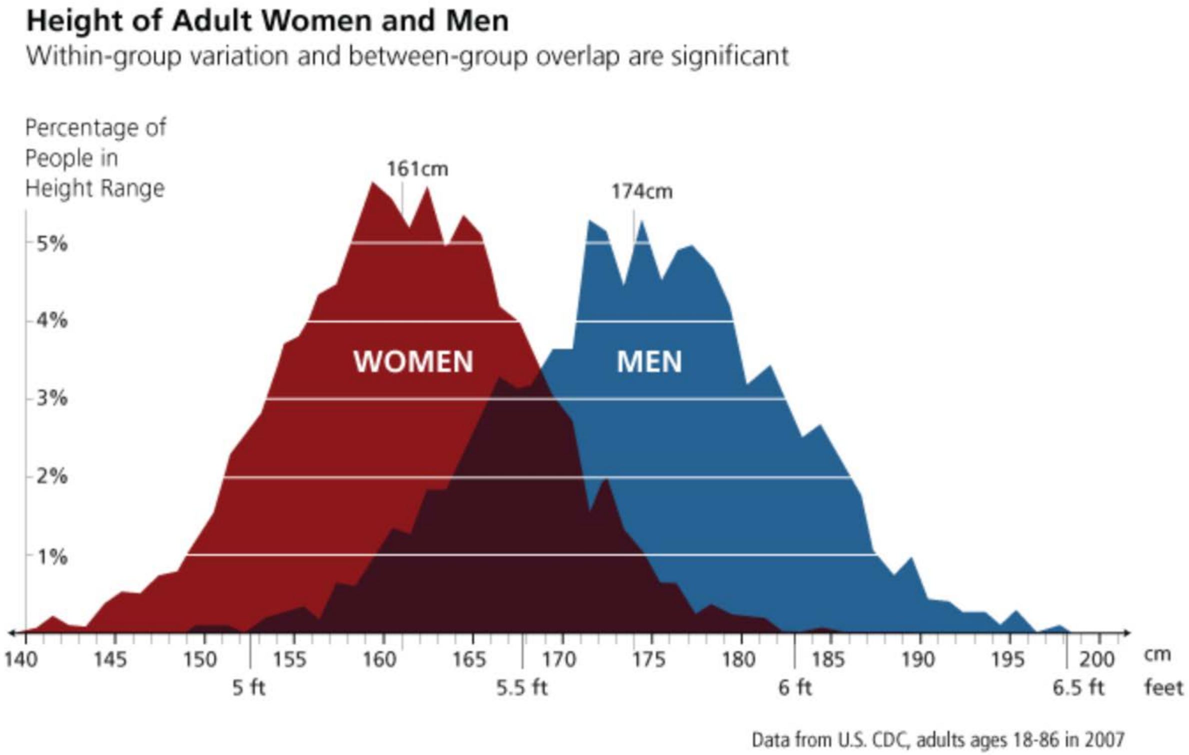






Probability density

- Probability density function *f* shows you where the data is: $P(a \leq X \leq b) = \int_a^b f(x) d_x$
- Note: y-value does not give the probability of observing data point x
- Instead, think of areas; Area under the pdf has to be 100%
- Must be estimated from data: parametric vs non-parameteric

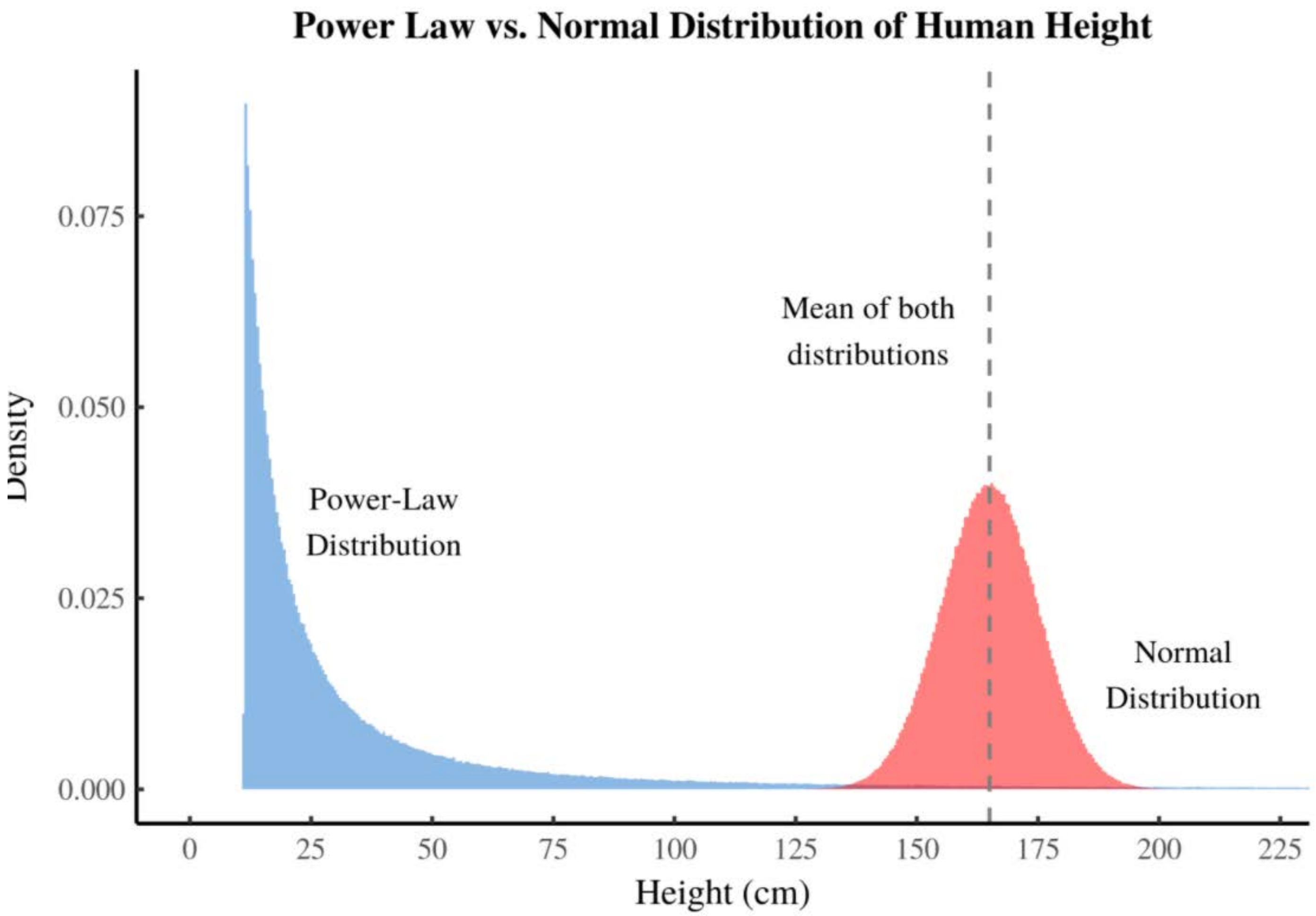


Height of Adult Women and Men

Within-group variation and between-group overlap are significant

Data from U.S. CDC, adults ages 18-86 in 2007

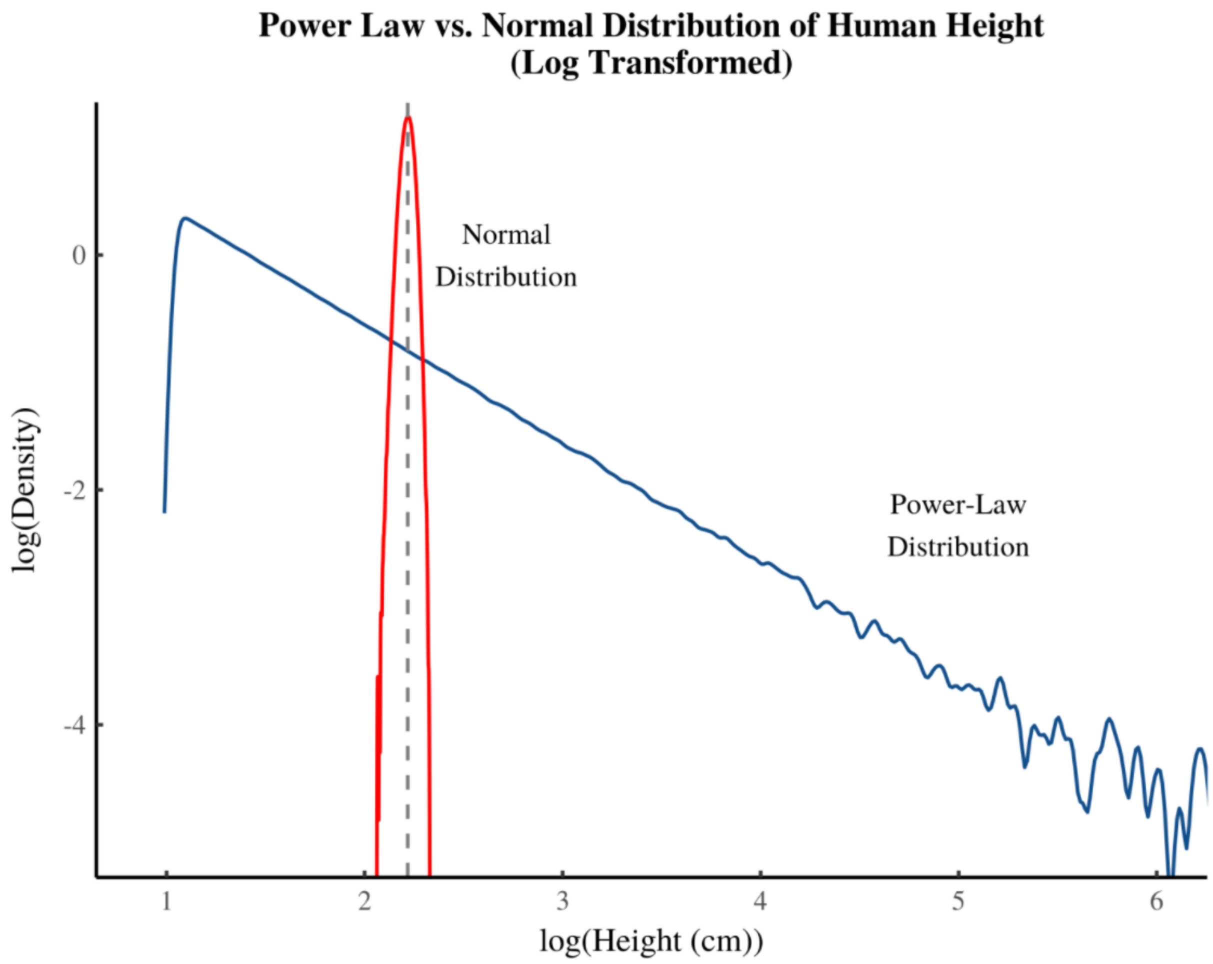
Normal vs skewed data



Equal size bins have sparse data in the tail; use log bins

Source: https://economicsfromthetopdown.com/2019/04/25/visualizing-power-law-distributions/

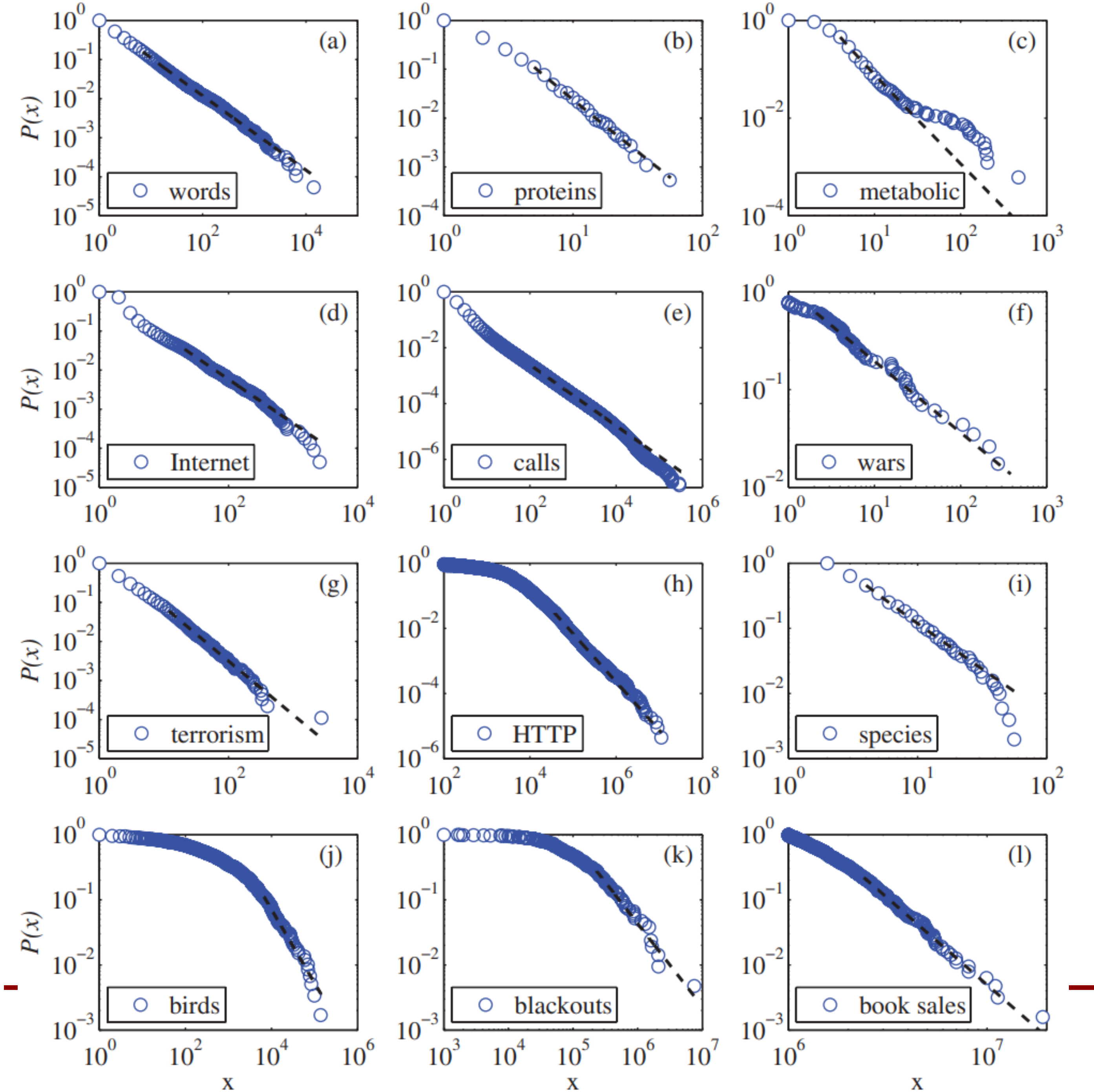
Normal vs skewed data



Source: https://economicsfromthetopdown.com/2019/04/25/visualizing-power-law-distributions/

Lots of data is highly skewed

(a) Frequency of unique words in Moby Dick by Herman Melville. (b) Degree distribution of protein interaction network. (c) and metabolic network of E. coli. (d) Degree distribution of autonomous systems on the Internet. (e) Number of long-distance phone calls received by AT&T customers. (f) Number of deaths in wars from 1816–1980 measured. (g) Deaths due to terrorist attacks worldwide 1968 -2006. (h) Number of bytes by HTTP (web) requests. (i) The number of species per genus of mammals during the late Quaternary period. (j) Frequency of sightings of bird species in the US. (k) Number of customers affected by blackouts in the US. (l) The sales volume of bestselling books in the US.



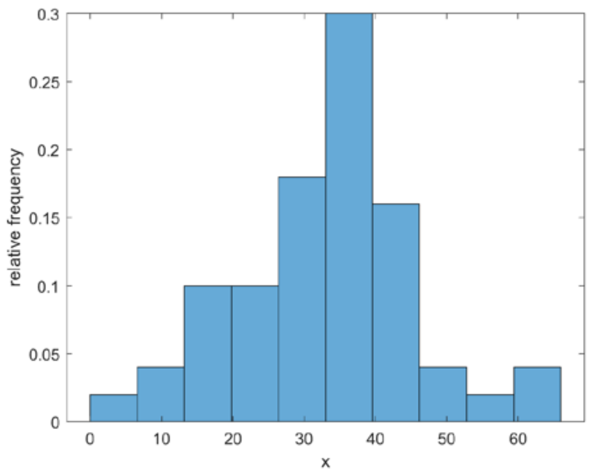
Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661-703.

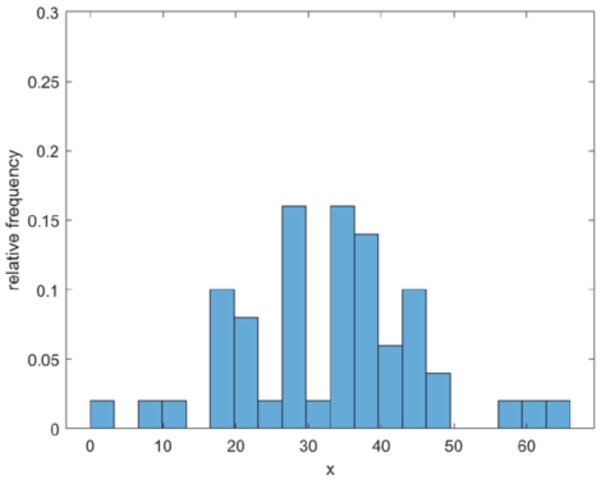
Density Estimation

- Given the training set $\boldsymbol{X}=\{x^t\}_t$ drawn iid from $p(x)$
- Divide data into bins of size h
- Histogram:

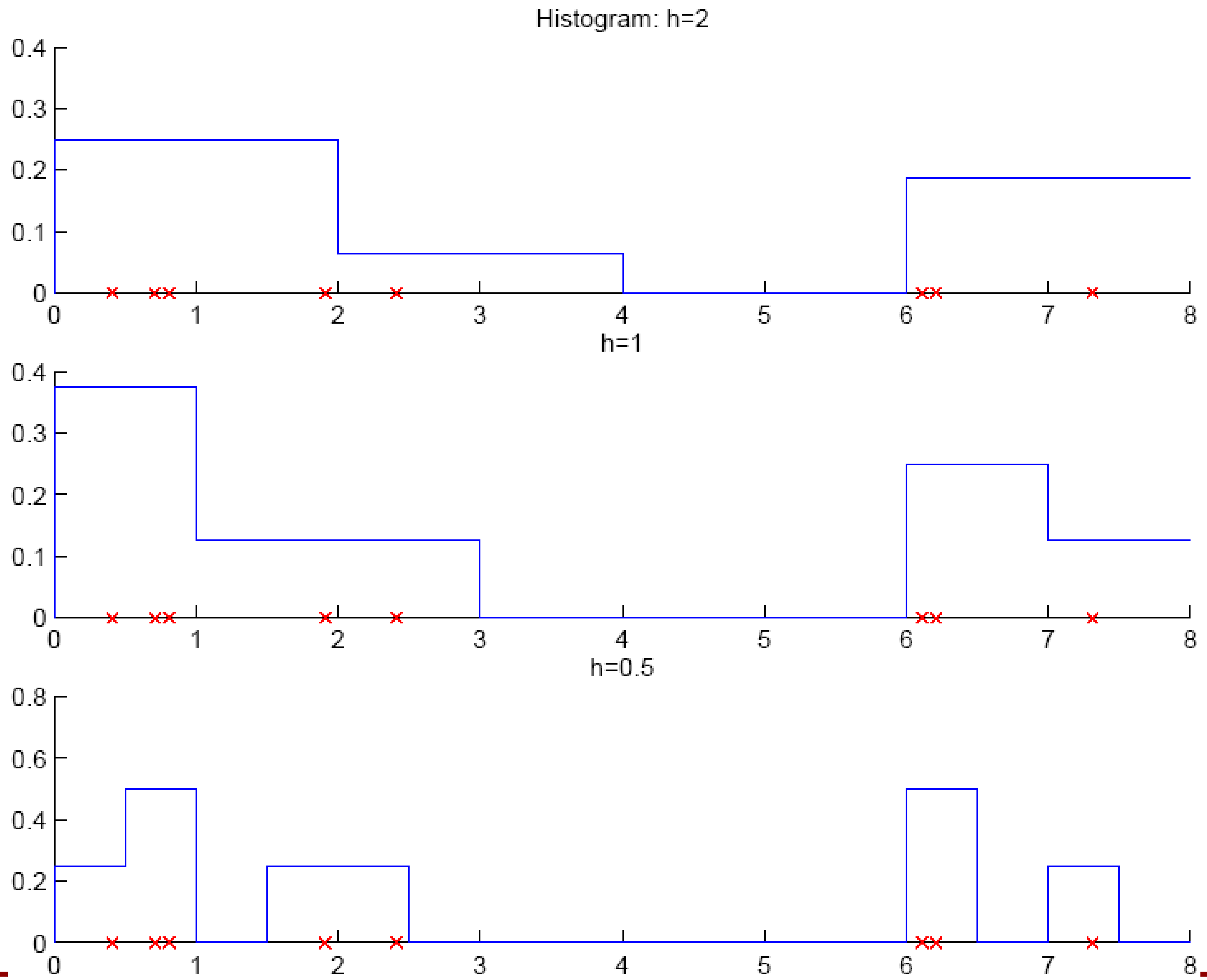
$$\hat{p}(x)=\frac{\#\{x^t \text{ in the same bin as } x\}}{Nh}$$

- Area equals 100%





Estimator often not smooth


Histogram: h=2
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
h=1
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
h=0.5
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8

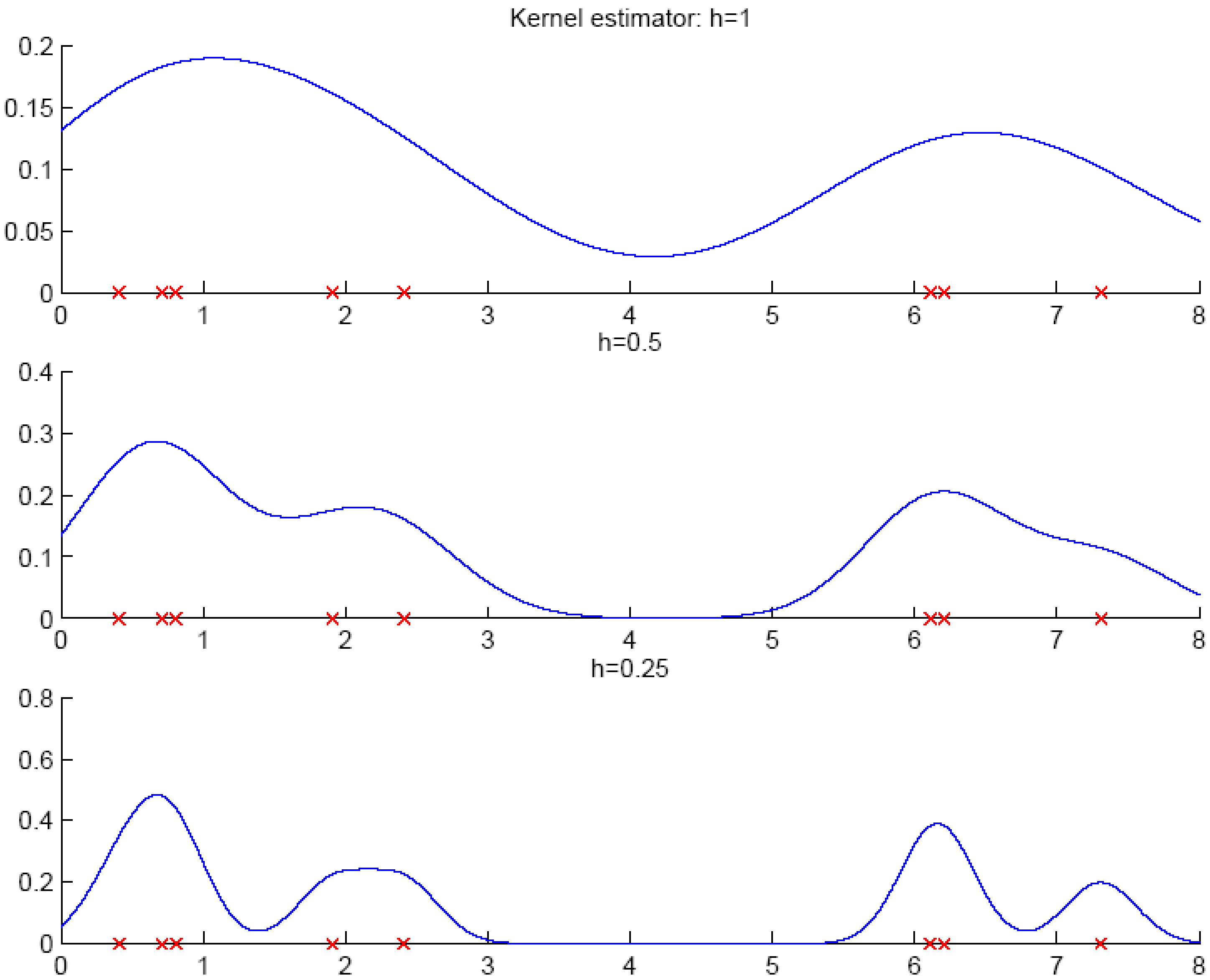
Kernel Estimator

- Smoothing the estimators
- Kernel function, e.g., Gaussian kernel:

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

- Kernel estimator (Parzen windows)

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)$$

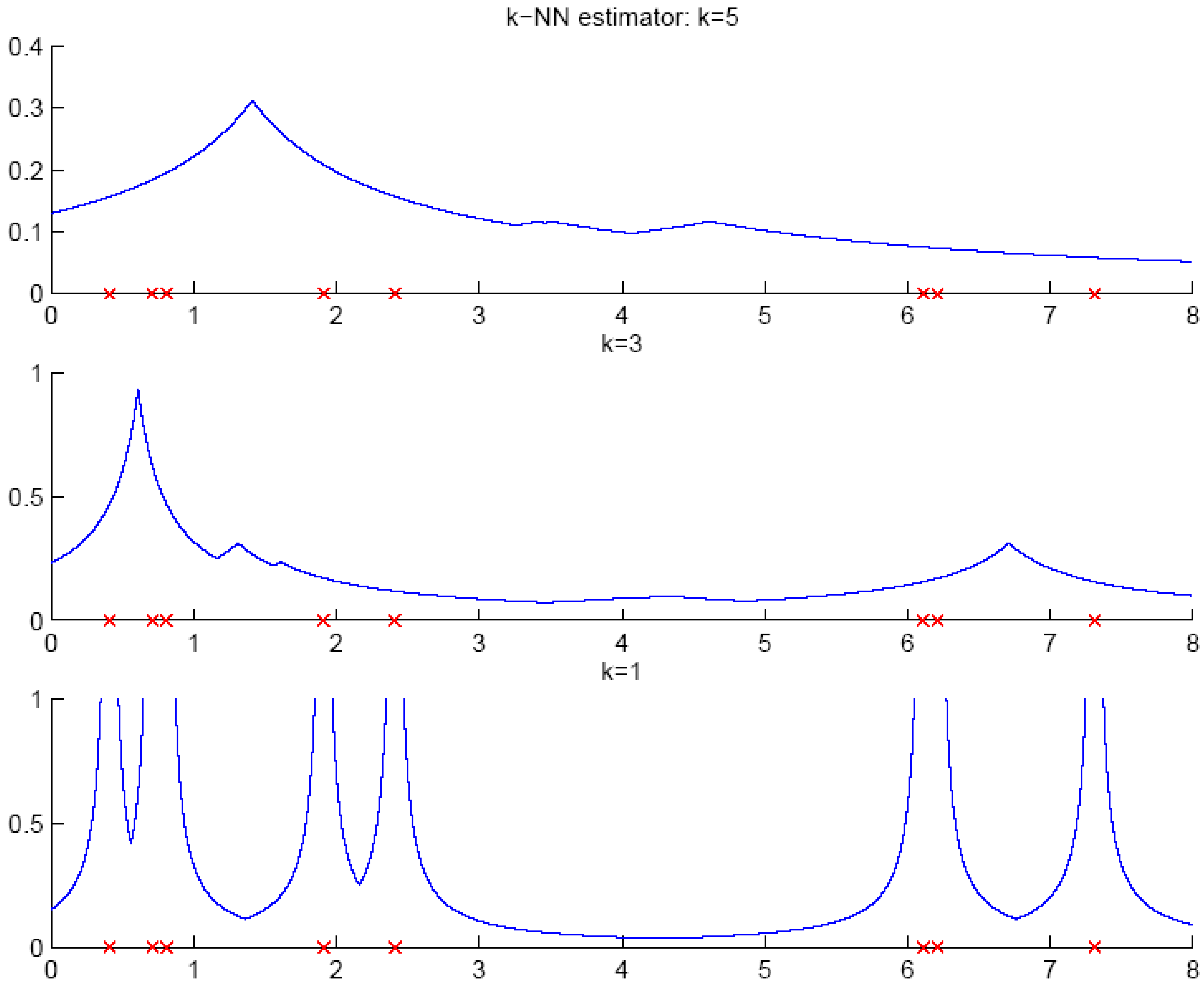

Kernel estimator: h=1
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
h=0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
h=0.25
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8

k-Nearest Neighbor Estimator

- Instead of fixing bin width *h* and counting the number of instances, fix the instances (neighbors) *k* and check bin width

$$\hat{p}(x) = \frac{k}{2Nd_k(x)}$$

$d_k(x)$, distance to *k*th closest instance to *x*


k−NN estimator: k=5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
k=3
1
0.5
0
0
1
2
3
4
5
6
7
8
k=1
1
0.5
0
0
1
2
3
4
5
6
7
8

Multivariate Data

- Kernel density estimator

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh^d} \sum_{t=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}^t}{h}\right)$$

Multivariate Gaussian kernel

spheric

$$K(\mathbf{u}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\mathbf{u}\|^2}{2}\right]$$

ellipsoid

$$K(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}|\mathbf{S}|^{1/2}} \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{S}^{-1}\mathbf{u}\right]$$

Multivariate Data

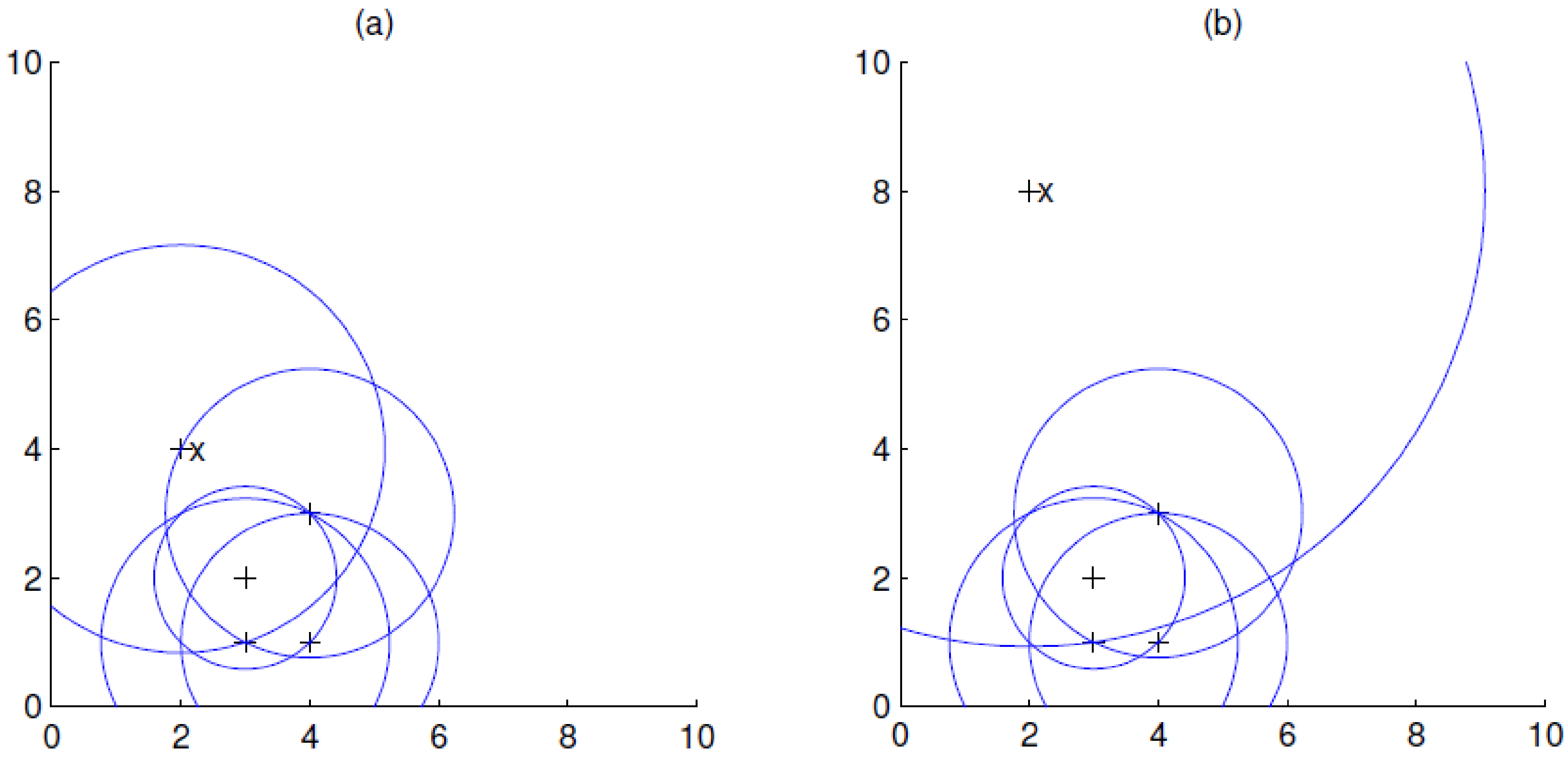
- Watch out for using non-parametric estimates with high-dimensional data: curse of dimensionality
- Example: dim(X)=8, and we use a histogram with 10 bins per dimension
  - 10^8 bins, most will be empty
  - Estimates will be mostly zero
- Choose bin size carefully

Outlier Detection

- Find outlier/novelty points
- Not a two-class problem because outliers are very few, of many types, and seldom labeled
- Instead, one-class classification problem: Find instances that have low probability
- In nonparametric case: Find instances far away from other instances

Local Outlier Factor

$$\mathrm{LOF}(\boldsymbol{x}) = \frac{d_k(\boldsymbol{x})}{\sum_{\boldsymbol{s}\in\mathcal{N}(\boldsymbol{x})} d_k(\boldsymbol{s})/|\mathcal{N}(\boldsymbol{x})|}$$



Nonparametric Regression

- Aka smoothing models
- In regression, g(x) is polynomial.
- Regressogram.

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} b(x, x^t) r^t}{\sum_{t=1}^{N} b(x, x^t)}$$

where

$$b(x, x^t) = \begin{cases} 1 & \text{if } x^t \text{ is in the same bin with } x \\ 0 & \text{otherwise} \end{cases}$$

Running Mean/Kernel Smoother

- Running mean smoother. Creates symmetric bins around x, averages points

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right)}$$

where

$$w(u) = \begin{cases} 1 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

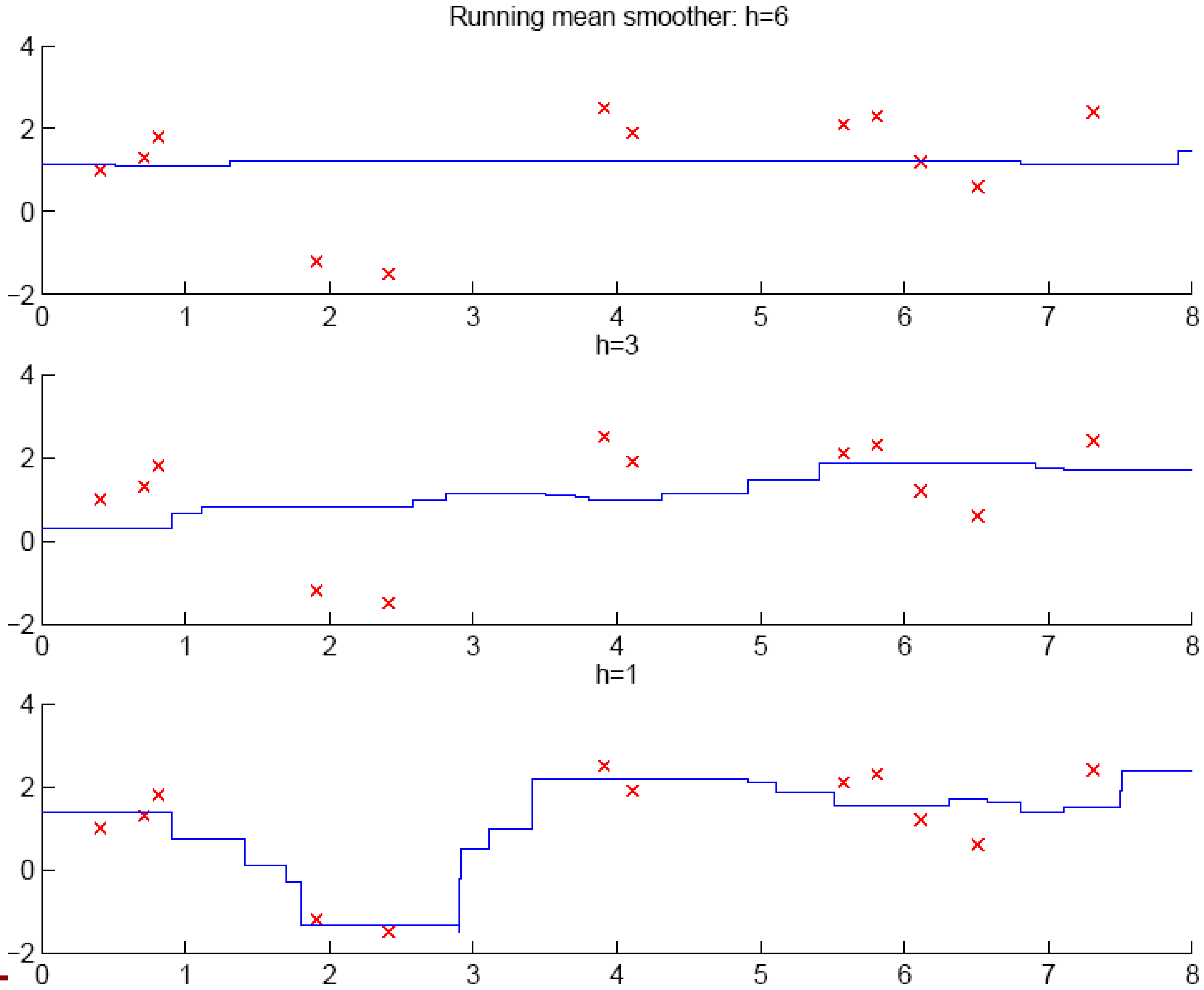
- Running line smoother

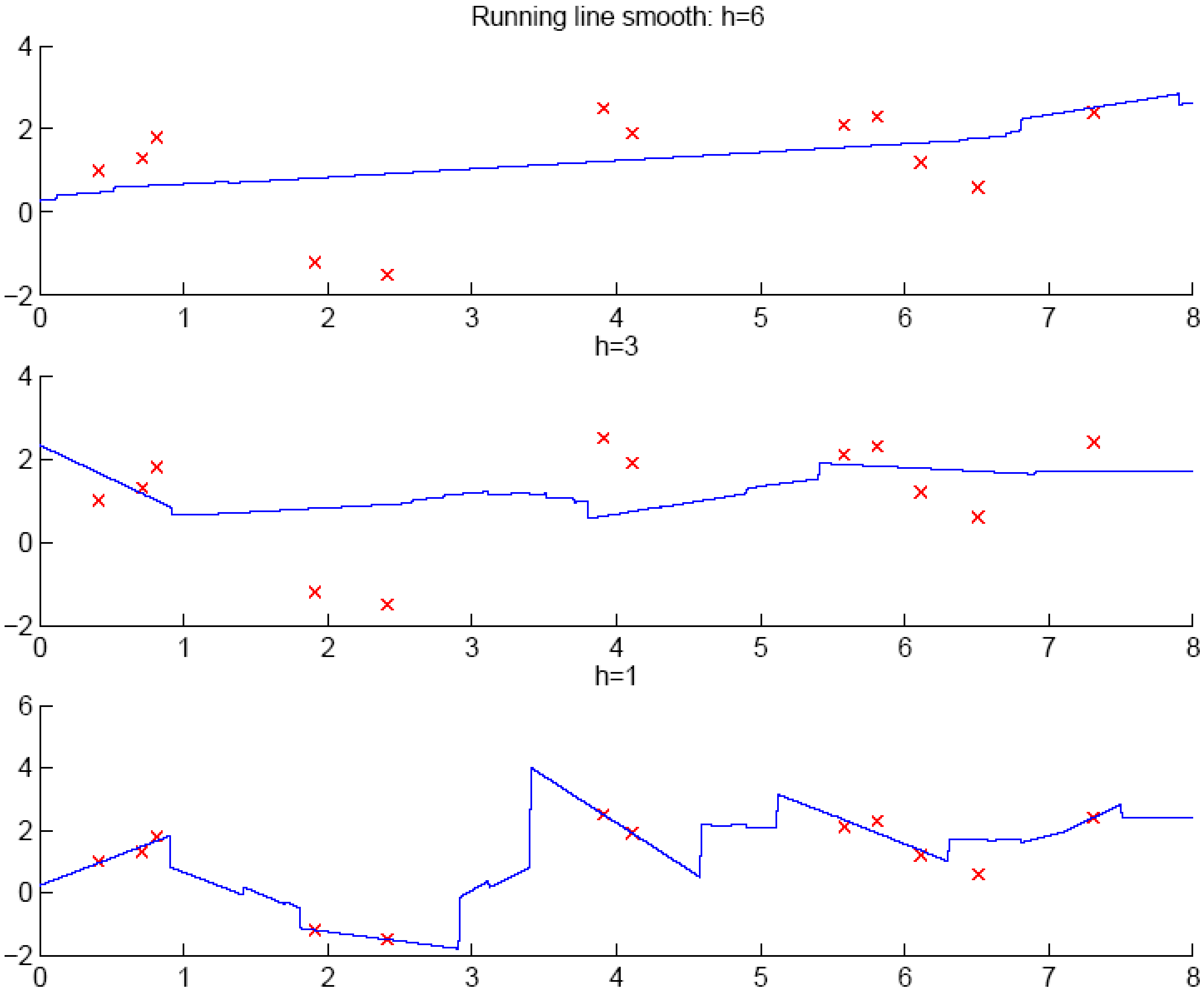
- Kernel smoother: like in density estimation, gives less weight to distant points

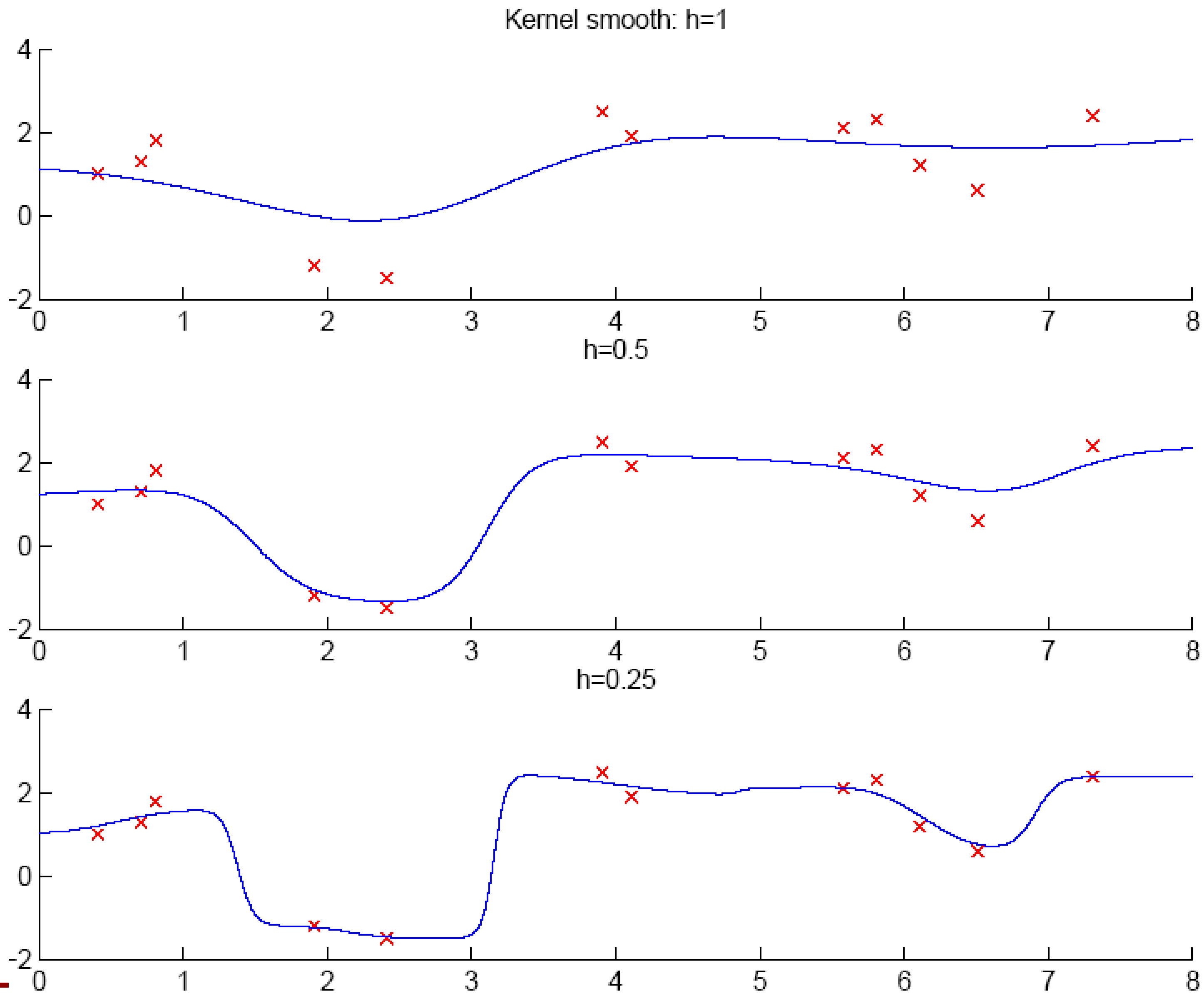
$$\hat{g}(x) = \frac{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)}$$

where *K*( ) is Gaussian

- Additive models (Hastie and Tibshirani, 1990)

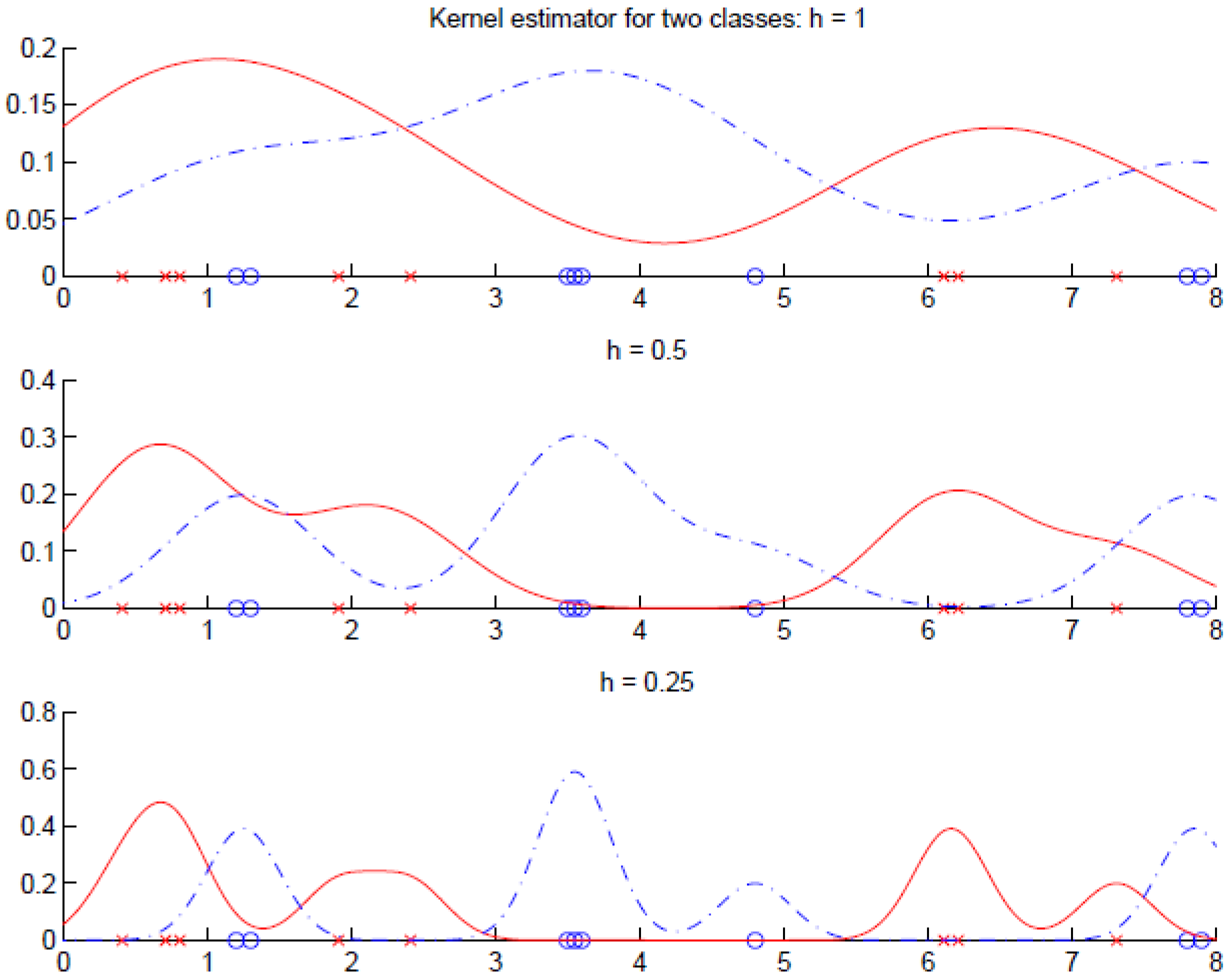

Running mean smoother: h=6
h=3
h=1
4
2
0
-2
0
1
2
3
4
5
6
7
8


Running line smooth: h=6
4
2
0
−2
0
1
2
3
4
5
6
7
8
h=3
4
2
0
−2
0
1
2
3
4
5
6
7
8
h=1
6
4
2
0
−2
0
1
2
3
4
5
6
7
8


Kernel smooth: h=1
h=0.5
h=0.25
4
2
0
-2
0
1
2
3
4
5
6
7
8

How to Choose *k* or *h* ?

- When *k* or *h* is small, single instances matter; bias is small, variance is large (undersmoothing): High complexity
- As *k* or *h* increases, we average over more instances and variance decreases but bias increases (oversmoothing): Low complexity
- Cross-validation is used to finetune *k* or *h*.


Kernel estimator for two classes: h = 1
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
h = 0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
h = 0.25
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8

VISUALIZING HIGH DIMENSIONAL DATA USING DIMENSIONALITY REDUCTION TECHNIQUES

Visualizing high dimensional data

- No visualization is available for high dimensional data (only partially; lose a lot of information)
- Methods to allow us to visualize/analyze data from higher dimensions in lower dimensions without losing a good part of the information
- Dimensionality reduction

Dimensionality reduction

- Given dataset with *n* rows(observations) and *m* columns(features).
- Create a new dataset with *n* rows and *m'* columns that best summarizes the original data
- *m'*<*m*, where *m* is a large number (*m*>>3)

- Two common solutions
  - PCA (Principal Component Analysis)
  - t-SNE (t-distributed Stochastic Neighbor Embedding)
    - Visualizing Data using t-SNE (original paper)

PCA (Principal Component Analysis)

Consider the original data set with n rows and m columns as a matrix X (columns of X should be standardized)

The goal is to find two **low rank** P (n rows, m’ columns) and Q (m’ rows, m columns)

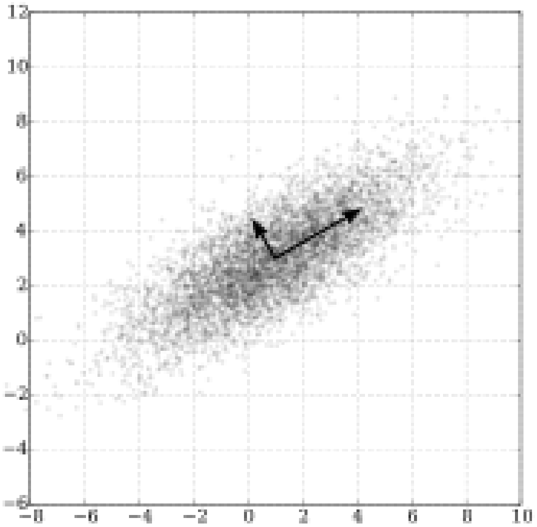
such that X’ = PQ best approximate X.

Then P is the low dimensional version of X. Rows of Q are the m’ features extracted, represented in original m dimensional space.

PCA (Principal Component Analysis)

How do we calculate P and Q?

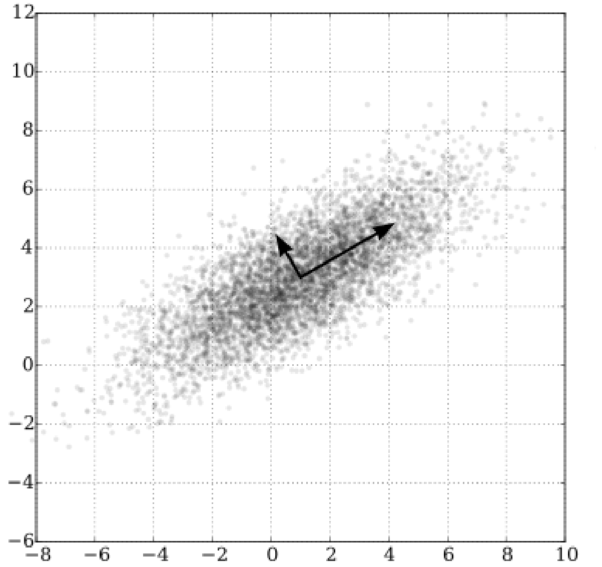
Intuition -- The n data points distributed along certain axes in the m dimensional space.



The axes are given by the eigenvectors of covariance matrix
$C = X^TX$.
Find only first m’ eigenvectors with largest absolute eigenvalues. Project data to those axes.

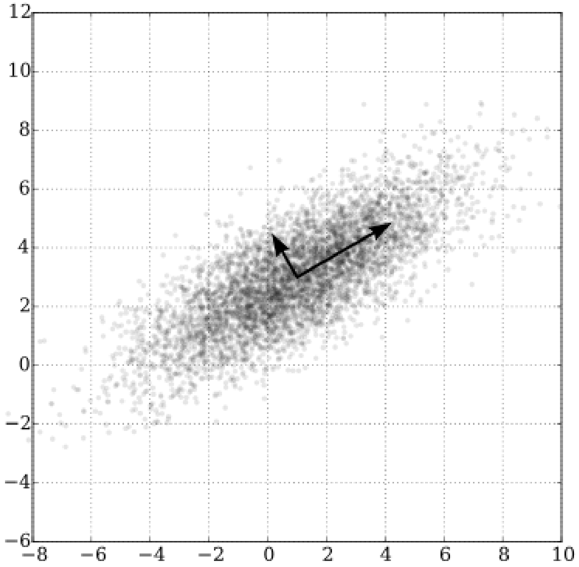
Principal Component Analysis - PCA

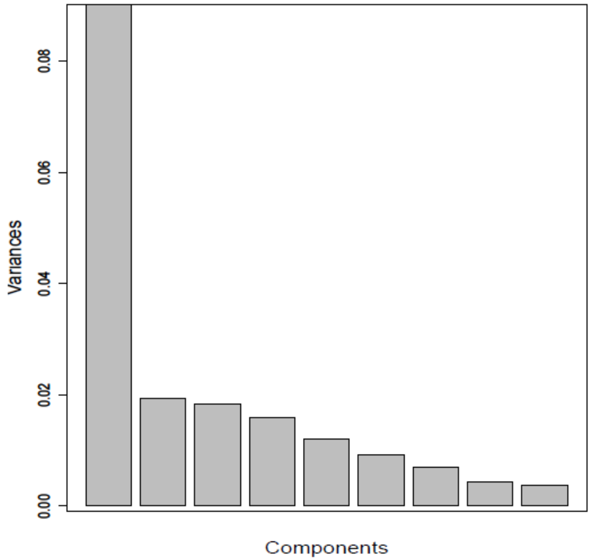
- Unsupervised method that identifies the internal structure of high-dimensional data that best explains its variance
- Embeds the data in a new lower-dimensional space, such that
  - The *first component* is that (normalized) linear combination of the variables with the largest variances.
  - The *second principal component* has largest variance, subject to being uncorrelated with the first….etc.
- Thus, with many correlated features, we replace them with a small set of principal components that capture their joint variation.



PCA (Principal Component Analysis)

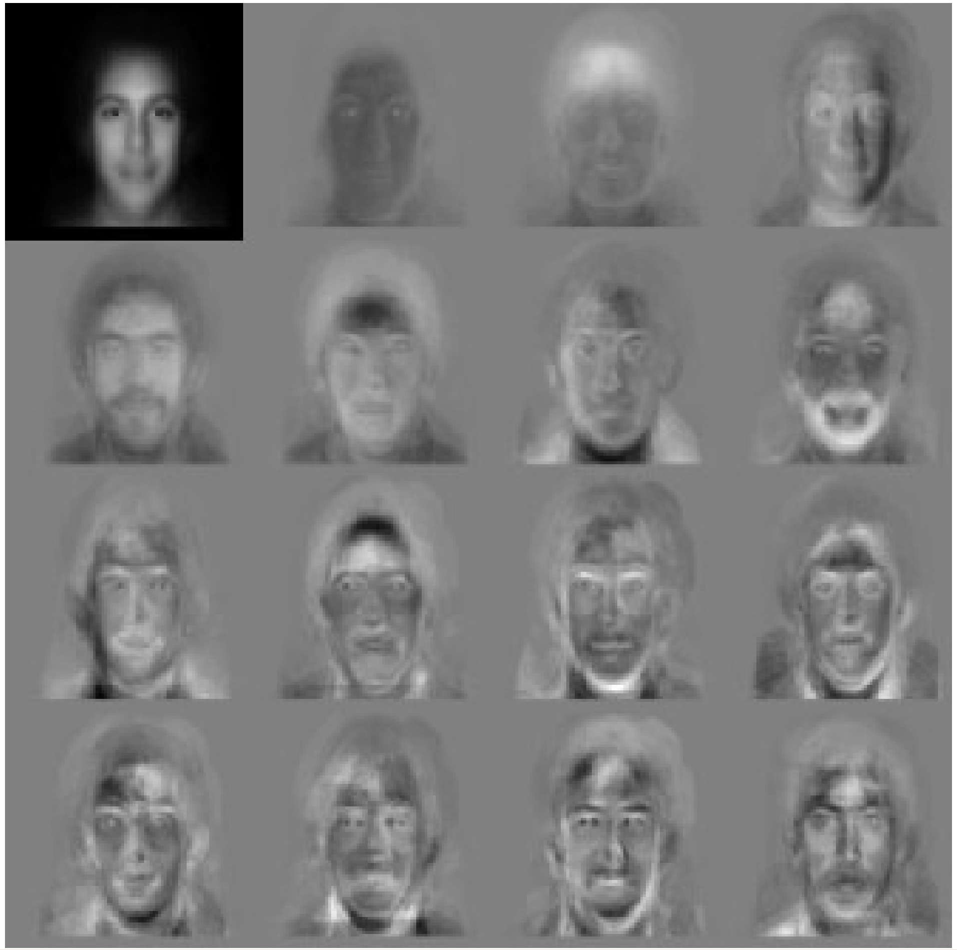
- Eigenvectors are orthogonal towards each other and have length one
- The first couple of eigenvectors explain the most of the variance observed in the data
- Low eigenvalues indicate little loss of information if omitted





Principal Component Analysis

- Use cases for PCA:
  - Data compression
  - Feature selection and feature reduction
  - Data visualization
  - Eigenfaces are the principal components of a large set of “faces”
    - Shape
    - Hair
    - Beard
    - Glasses
    - ...



t-distributed Stochastic Neighbor Embedding (t-SNE)

- Nonlinear dimensionality reduction technique for embedding high-dimensional data in a low-dimensional space (two or three dimensions)
- Used for visualizing high-dimensional data
- Models each object as a 2- or 3-dimensional point such that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

t-SNE (t-distributed Stochastic Neighbor Embedding)

For original data, define similarity measures

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Where $$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||^2 / 2\sigma_i^2)}$$

Define the similarity measures of transformed data (y) as

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}$$

t-SNE (t-distributed Stochastic Neighbor Embedding)

And set $p_{ii} = 0 \qquad q_{ii} = 0$

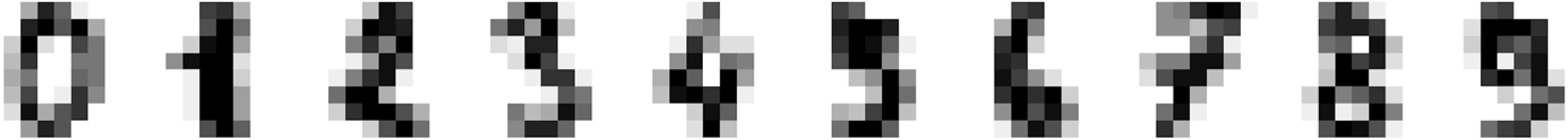
Minimize $\mathrm{KL}\left(P \parallel Q\right) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

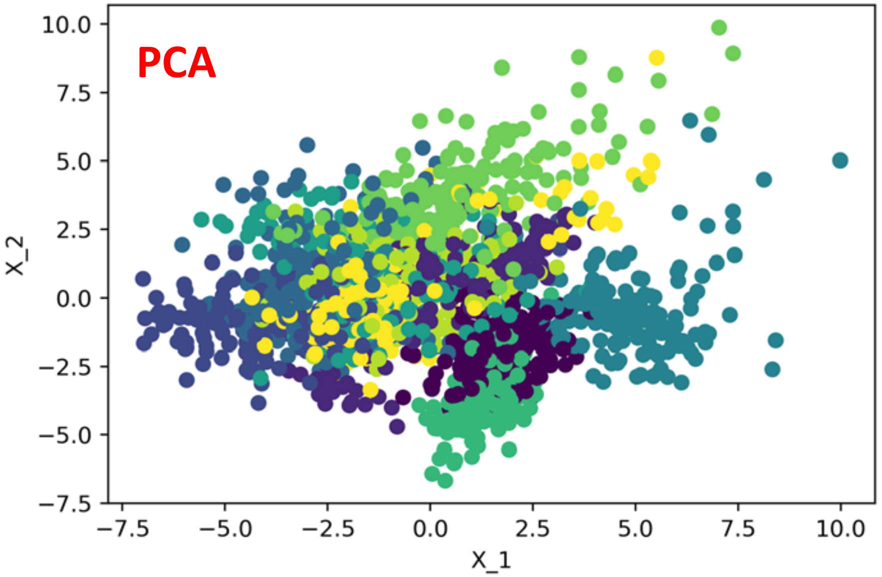
Using gradient descent.

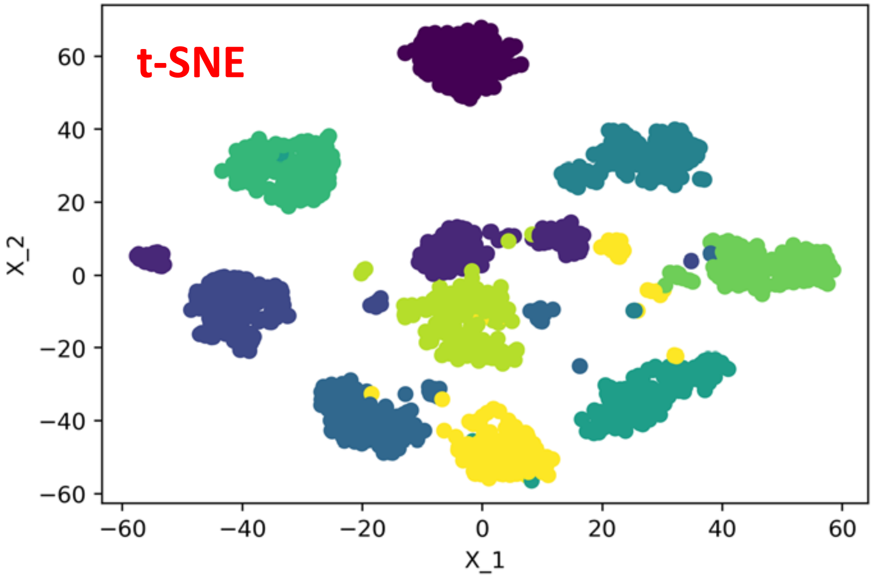
Intuition: Similar data observations are located closer in the transformed space.

Example -- Handwritten digits images

Original data (8*8 images, 64 dimensions)

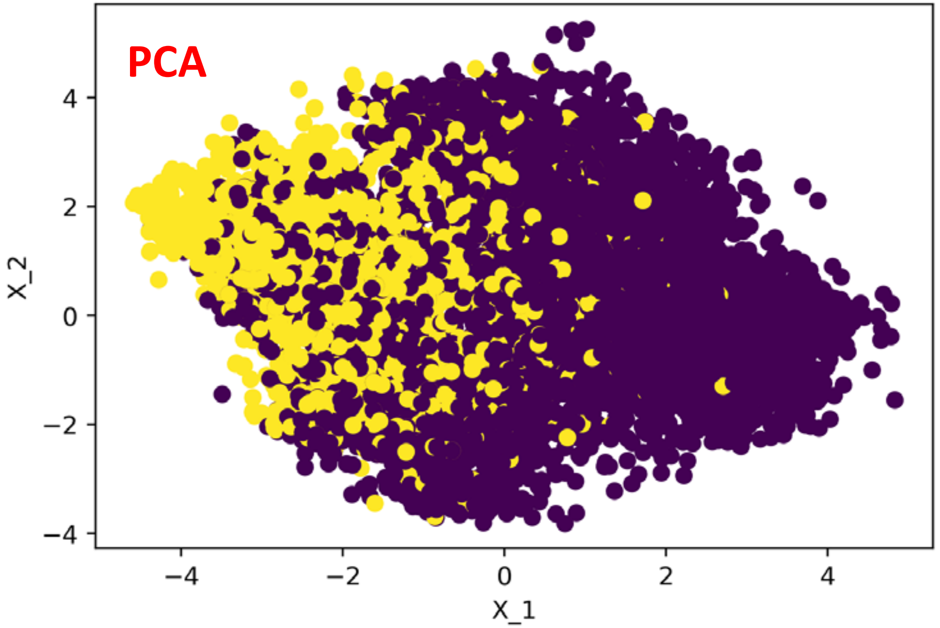


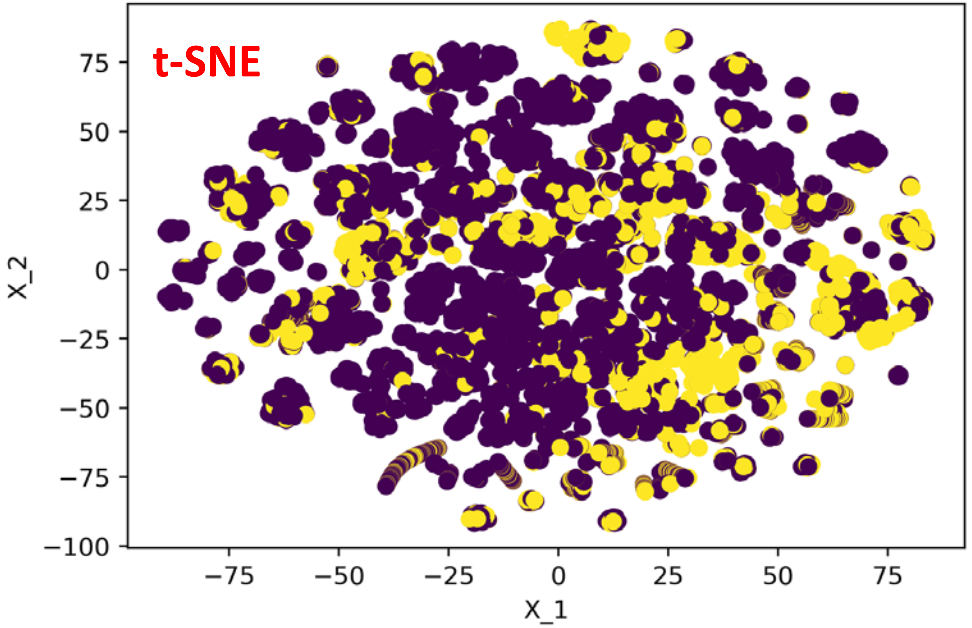




Example -- Adult Dataset (Tabular)

Adult dataset contains features of individuals. The outcome is whether the person makes $50K annually. (shown as yellow)





Summary

PCA

- Linear
- Global structure
- Affected by outliers
- Deterministic
- Fitting data to an ellipsoid

t-SNE

- Nonlinear
- Local structure
- Not sensitive to outliers
- Stochastic
- Based on data similarity
- Usually the best method for image data

- Questions?

- Virtual office hour
- https://usc.zoom.us/j/95136500603?pwd=VEJhblhWK25lT2N3RC9FNWk3eTJKQT09