School of Computing
UNIVERSITY OF GEORGIA

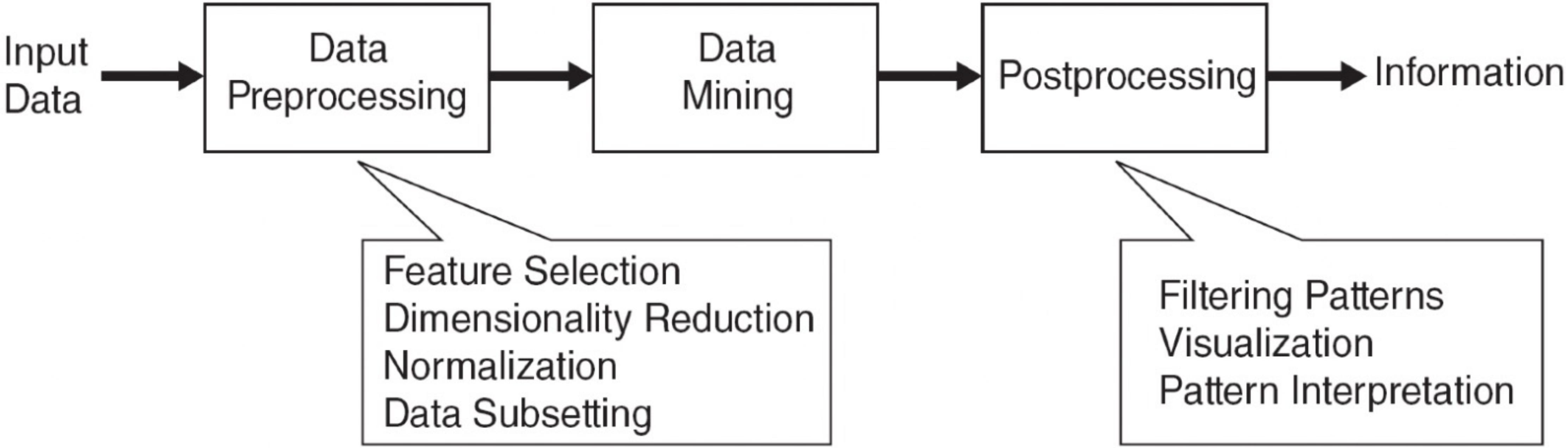
CSCI 4380/6380 Data Mining

Fei Dou

Assistant Professor
School of Computing
University of Georgia

September 12,13 2023

Recap: Data Mining Process



Recap: Data Preprocessing

- **Data cleaning**:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Recap: PCA

Algorithm

- Normalize the data to be zero mean. (m data objects with n features)

- Calculate the sample covariance matrix

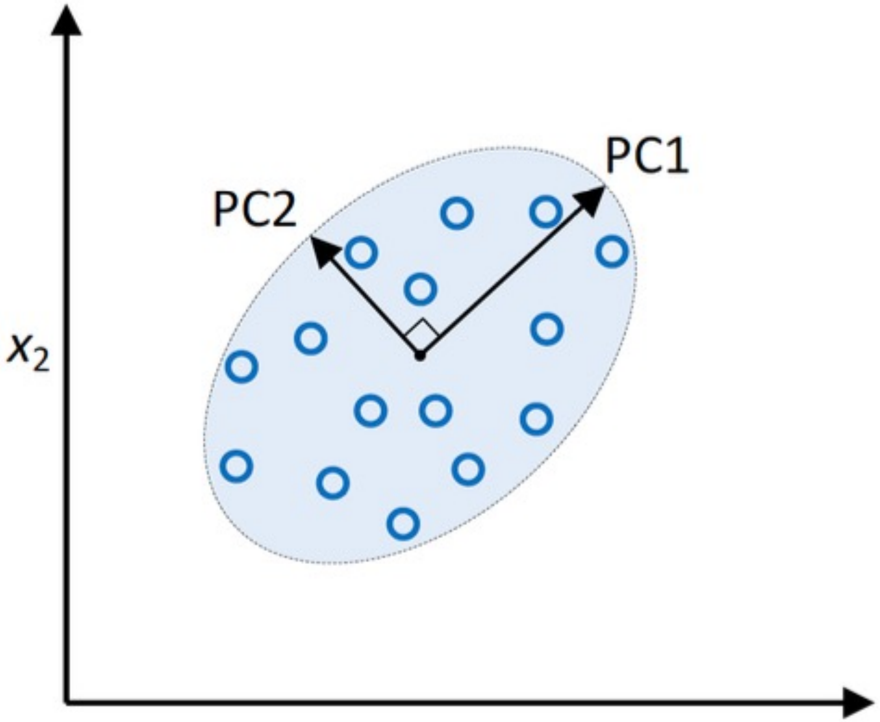
- Find the n eigenvector -eigenvalue pairs of the sample covariance matrix
  - PCA basis vectors = the eigenvector
  - Larger eigenvalue $\Longrightarrow$ more important eigenvectors

- Choose the top k eigenvectors corresponding to the highest eigenvalues

- Project the data to the lower dimensional space.

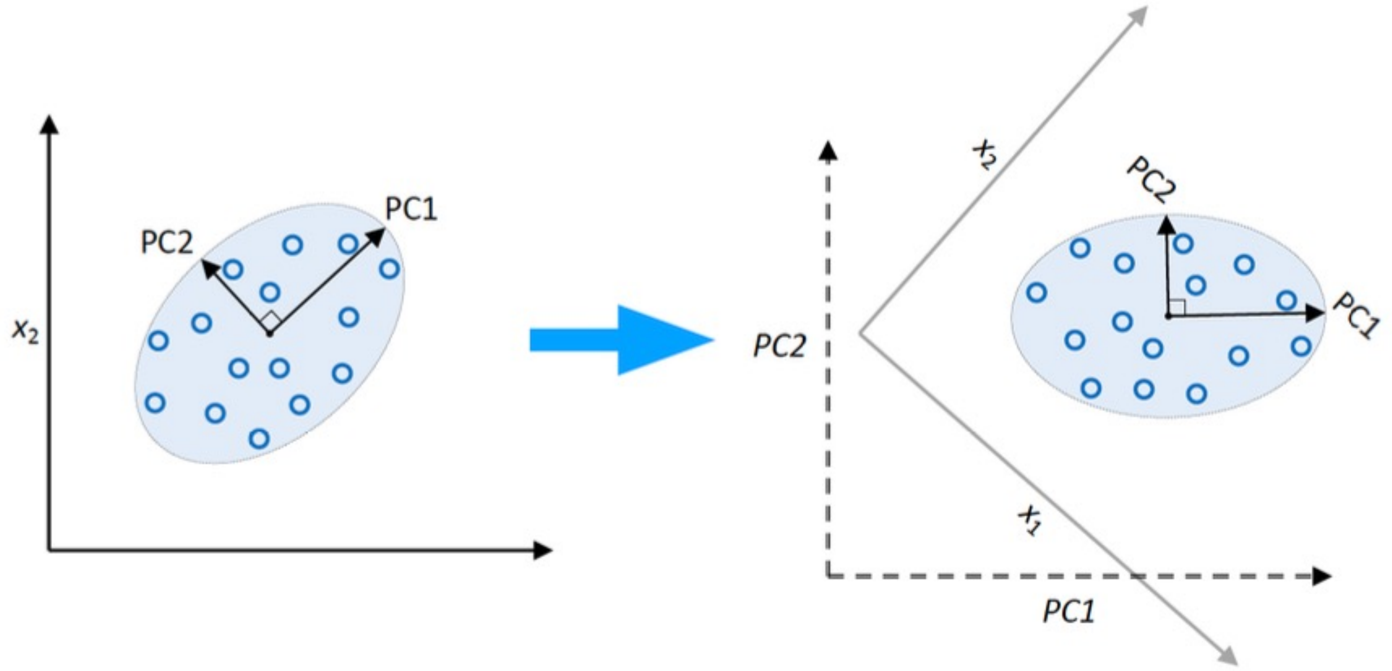
PCA

- **Intuition**
  - Step 1: Find directions of maximum variance



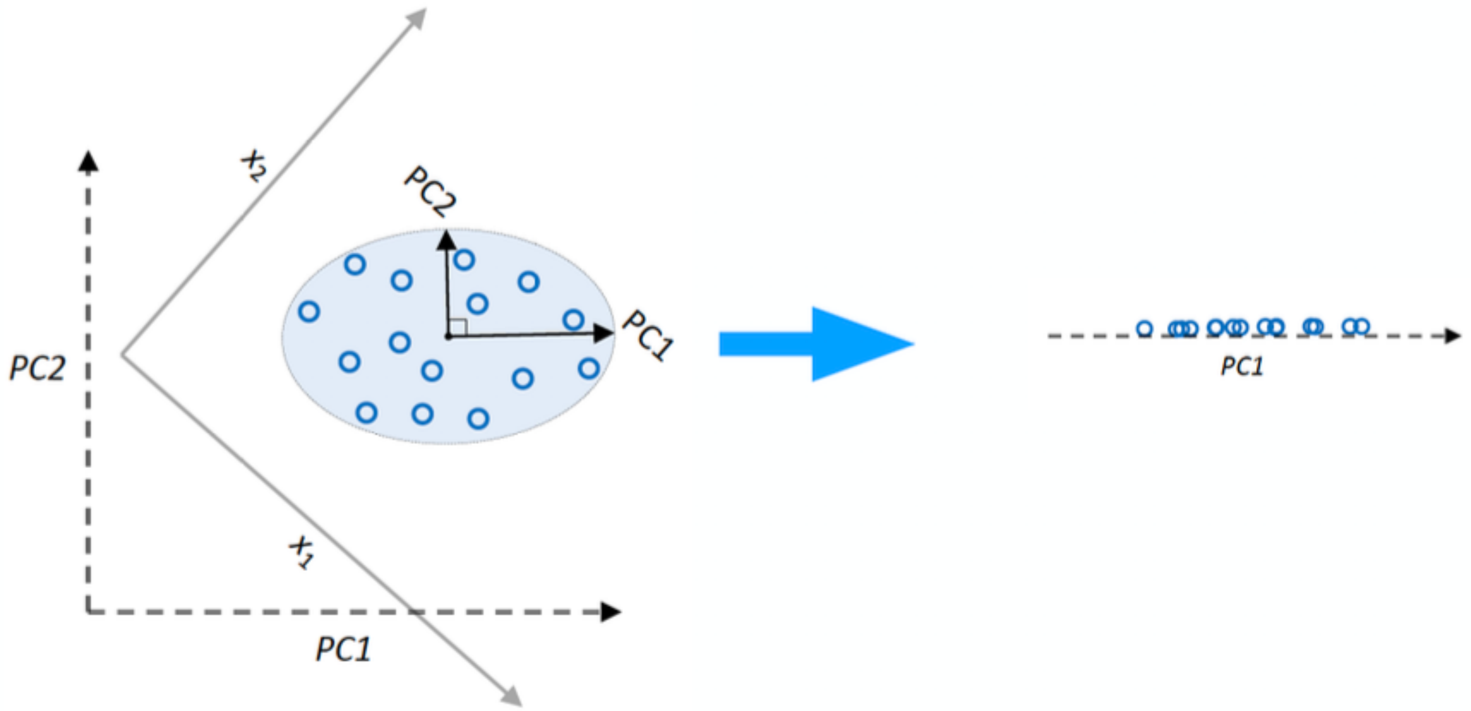
PCA

- **Intuition**
  - Step 2: Transform features onto directions of maximum variance



PCA

- **Intuition**
  - Step 3: Usually consider a subset of vectors of most variance (DR)



PCA - Interpretation

Maximum Variance Direction: 1^{st} PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

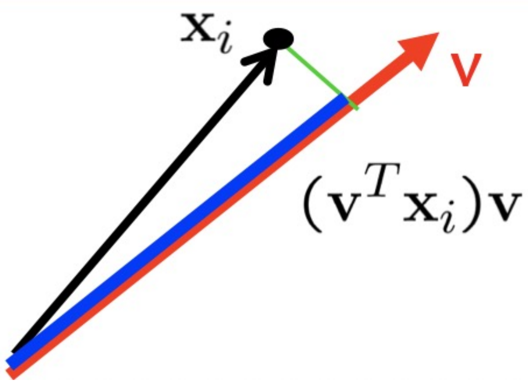
Minimum Reconstruction Error: 1^{st} PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$

blue² + green² = black²

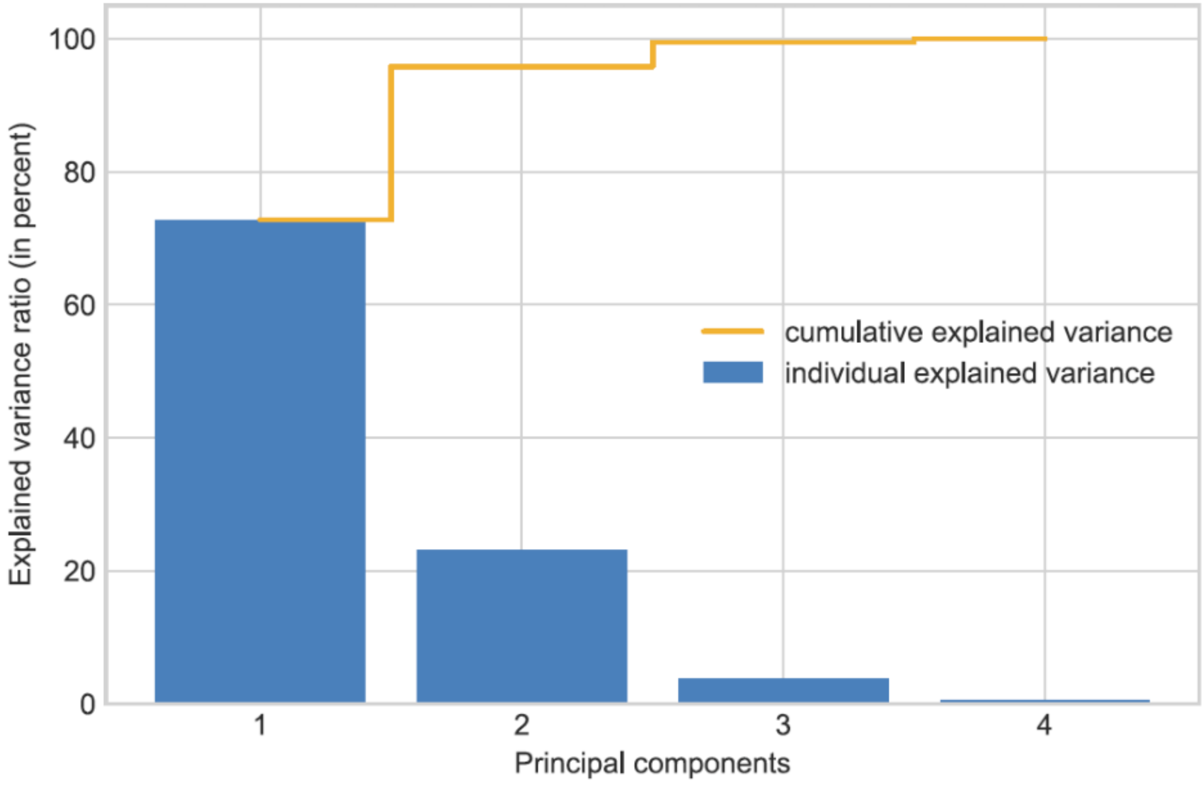
black² is fixed (it's just the data)

So, maximizing blue² is equivalent to minimizing green²



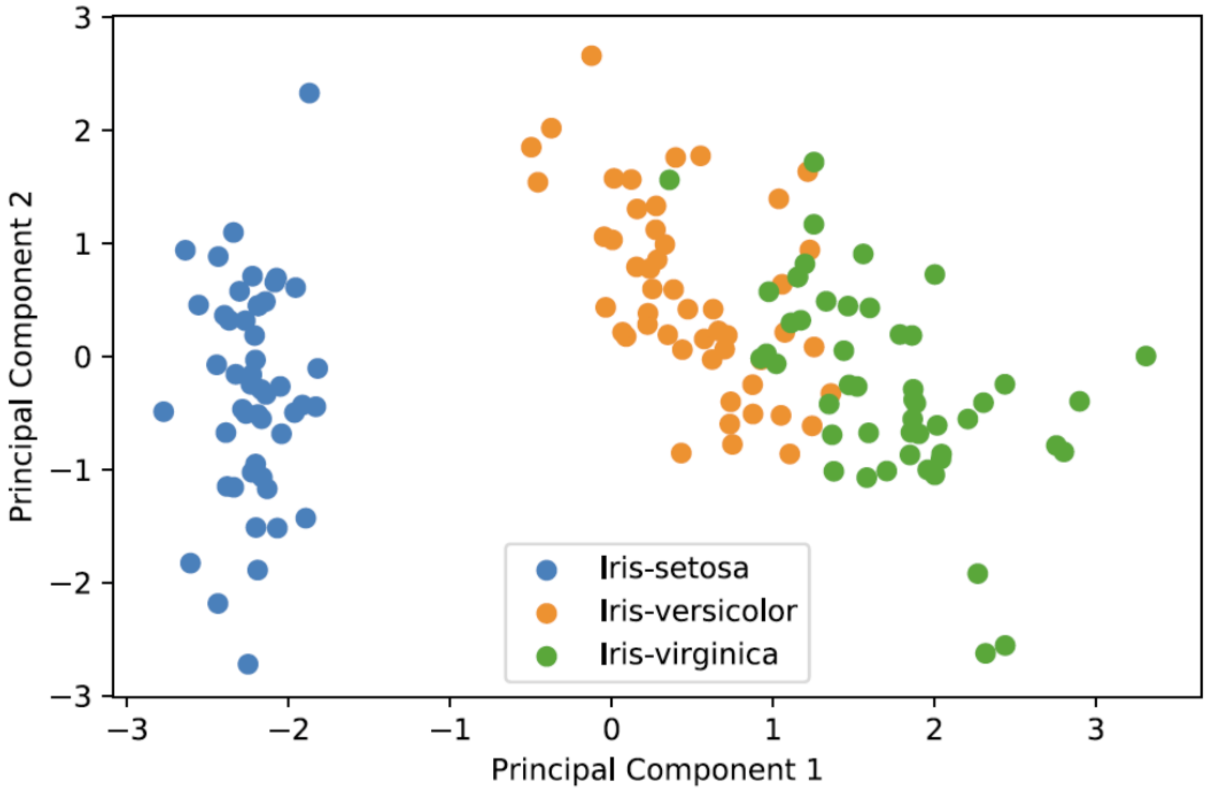
PCA

- Usually useful to plot the explained variance (normalized eigenvalues)
- Ration = $\frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$



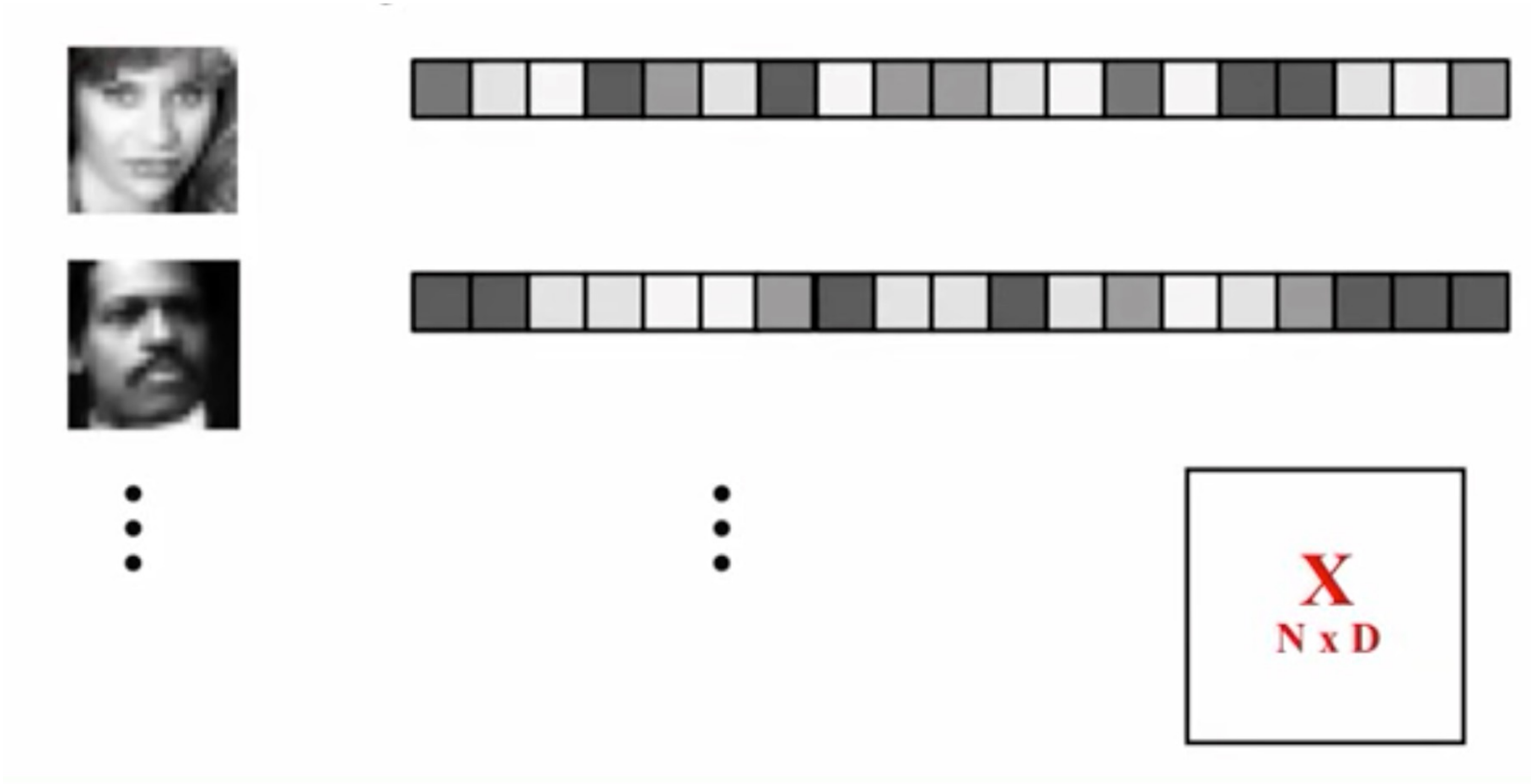
PCA

- PCA is an unsupervised method.
- Similarity measure can be conducted on the low dimensional space.
- Visualization can be performed with 2D or 3D space.



PCA Example: Eigenfaces

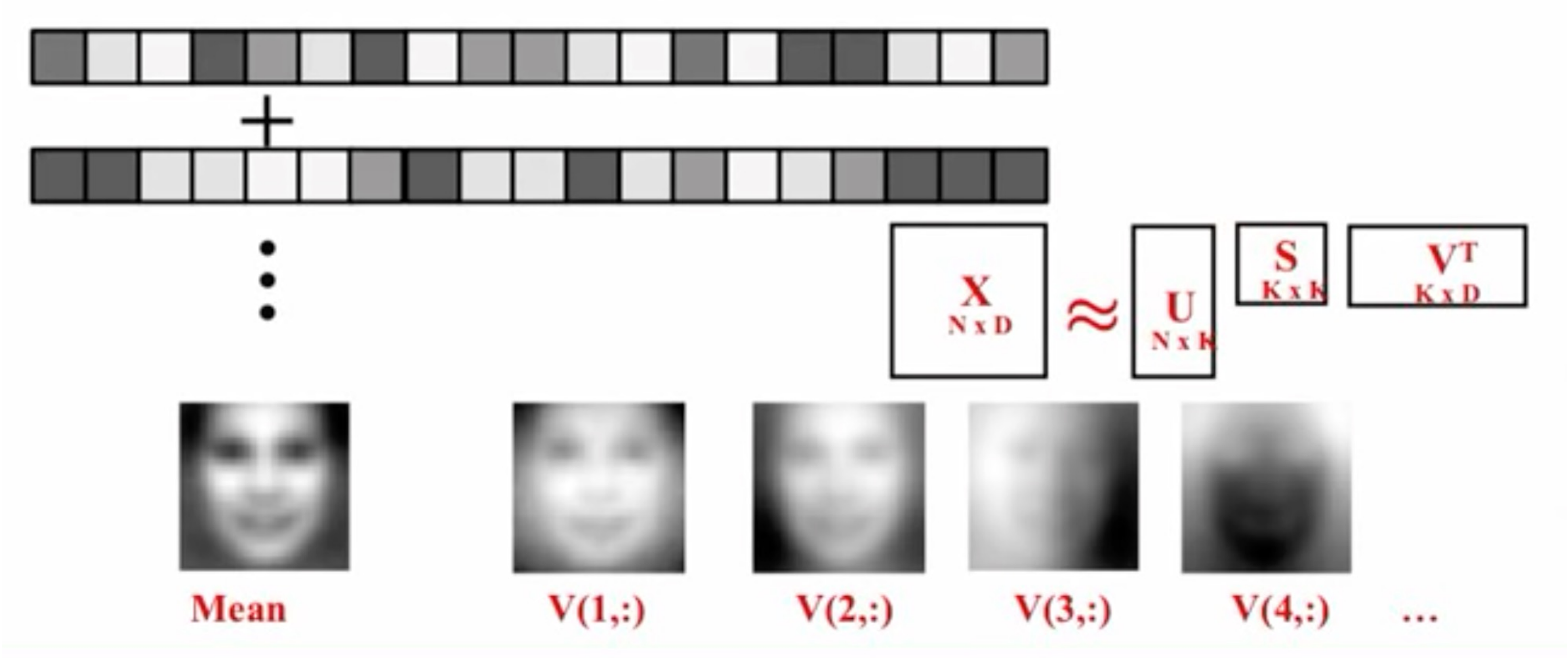
- Eigen-X = represent X using PCA
- Viola Jones data set: 24*24 images of faces = 576 dimensional measurments



PCA Example: Eigenfaces

- Eigen-X = represent X using PCA

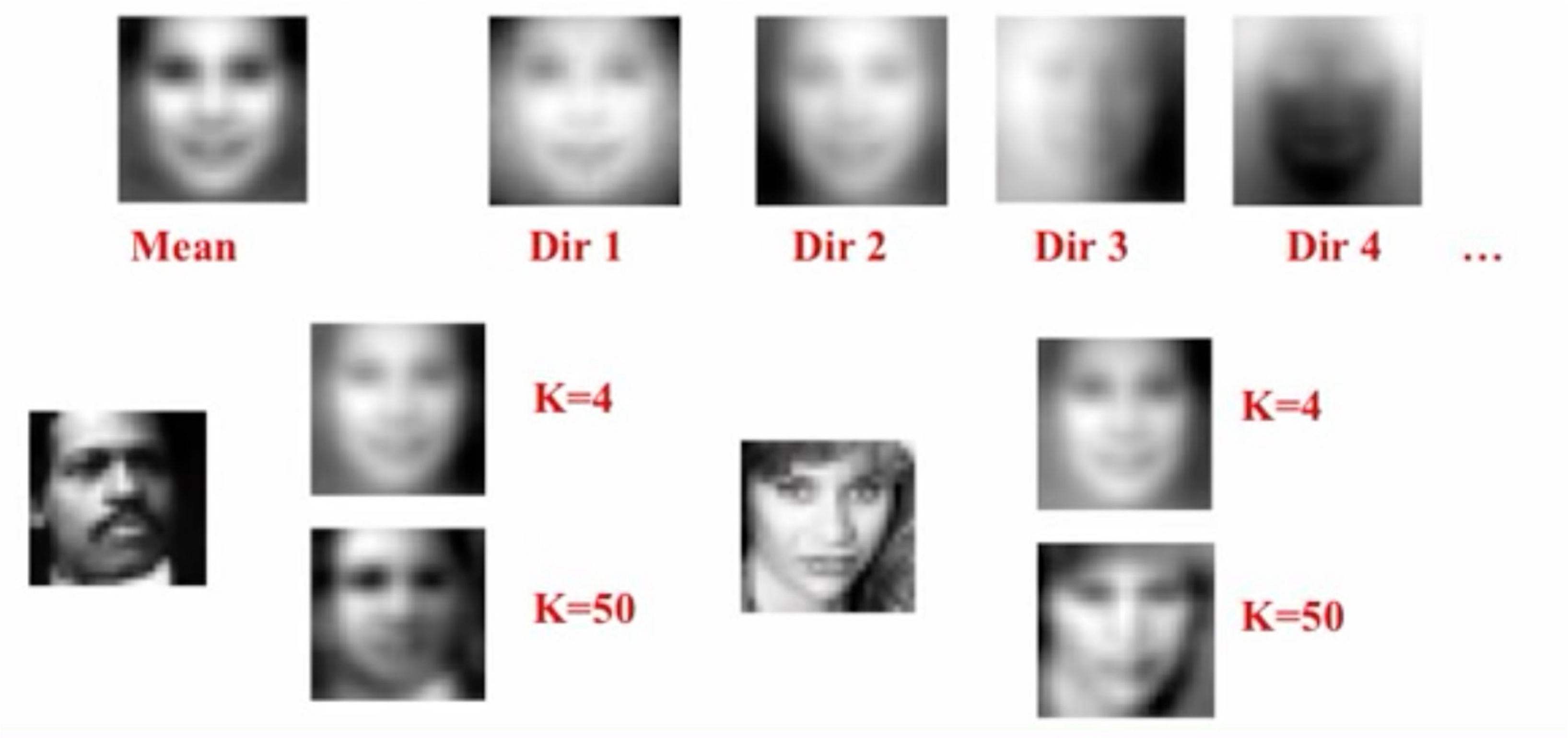
- Viola Jones data set: 24*24 images of faces = 576 dimensional measurements
- -Take first k PCA components



PCA Example: Eigenfaces

- Eigen-X = represent X using PCA

- Viola Jones data set: 24*24 images of faces = 576 dimensional measurements
- -Take first k PCA components



Dimensionality Reduction - Supervised

Supervised DR

- We aim to find the following mapping

$$\mathbf{x} \in \mathbb{R}^d \longrightarrow f(\mathbf{x}) \in \mathbb{R}^k \text{or}$$

$$\mathbf{X} \in \mathbb{R}^{m \times d} \longrightarrow f(\mathbf{X}) \in \mathbb{R}^{m \times k}$$

  when label information y is available.

- y can be the class, or similarity score of two data objects.
- $f(\cdot)$ can be a linear mapping or nonlinear mapping (e.g., kernel functions, neural networks).
- The similarity measure is conducted on the embedding space $f(\cdot)$ using Euclidean distance.

Feature Subset Selection

- Another way to reduce dimensionality of data

- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to New Space - DFT

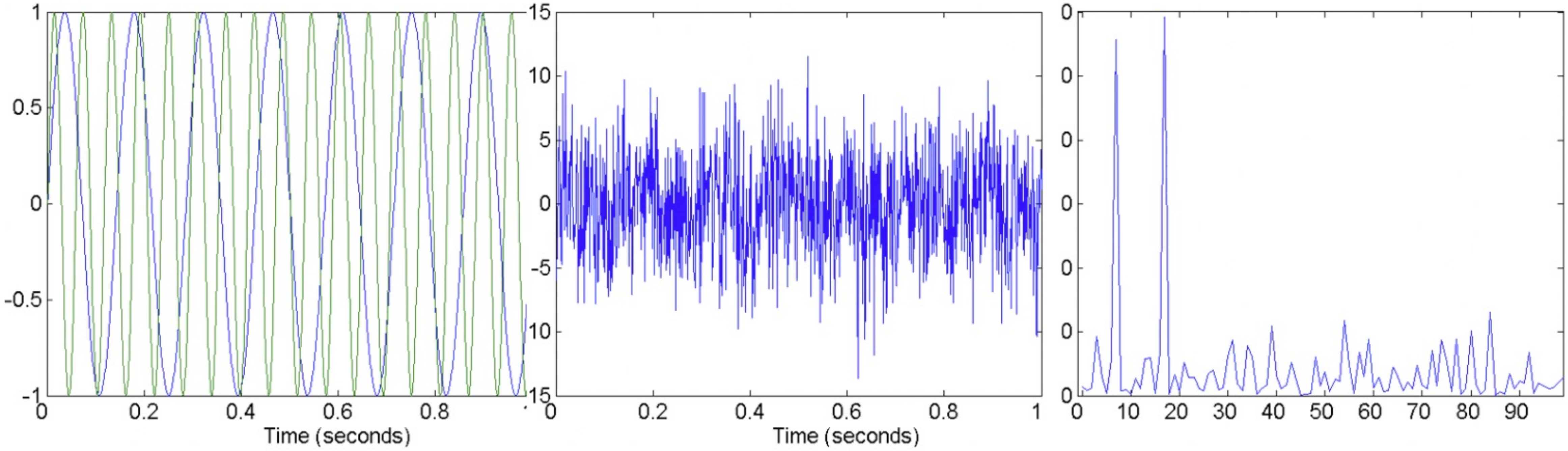
Discrete Fourier Transform (DFT)

- Why DFT?

- Many sequence are periodic.
  - sales patterns follow seasons
  - economy follows a *n* year circle
  - traffic and temperature follow daily circles

- Many real-time series follows multiple circles.

DFT



Two Sine Waves **Two Sine Waves + Noise** **Frequency**

Numerosity Reduction (Data Reduction)

- Reduce data volume by choosing alternative, smaller forms of data representation

- **Parametric** methods (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)

- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, . . .

Regress Analysis and Log-Linear Models

- Linear regression: $y = wx + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
- Multiple regression: $y = b_0 + b_1 x_1 + b_2 x_2$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - Approximate discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations

Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth later

Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a **representative** subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:

Types of Sampling

- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Data Transformation

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values

- Methods
  - Smoothing: Remove noise from data
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Binning, Concept hierarchy climbing

Normalization

Min-Max normalization

- A set of data objects with one numeric attribute: $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^{\boldsymbol{T}} \in \mathbb{R}^{\boldsymbol{n}}$

- Min-Max normalization

$$\widehat{x_i} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

- x_{max} and x_{min} are the maximum and the minimum values of the feature respectively.

Normalization

Z-normalization

- A set of data objects with one numeric attribute: $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^{\boldsymbol{T}} \in \mathbb{R}^{\boldsymbol{n}}$

- Z-normalization (standardization)

$$\widehat{x_i} = \frac{x_i - \mu_x}{\sigma_x}$$

- μ_x is the mean of population, σ_x is the standard deviation.

Normalization

Unit vector normalization

- A data object with multi-dimensional attribute: $\boldsymbol{x} = [x_1, x_2, \cdots, x_d] \in \mathbb{R}^{\boldsymbol{d}}$
- Unit vector normalization

$$\widehat{x_i} = \frac{x_i}{\|\boldsymbol{x}\|}$$

Normalization

Midrange and Decimal Scaling

- **Midrange** = $\frac{\min x \; + \; \max x}{2}$
- **Decimal scaling** $\boldsymbol{v_i}' = \frac{\boldsymbol{v_i}}{\boldsymbol{10^j}}$

- Example:
  - When you have a range of numbers like 50, 250, 400
  - Take the maximum number of digits. Here it is 3 (400 has 3 digits)
  - Calculate power of 10. 103 = 1000.
  - Divide each number by 1000.
  - The results would be 0.05, 0.25 and 0.4.

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
  - Equal-depth (frequency) partitioning
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Equi-width binning – for bin width of e.g., 10:
  - Bin 1: 4, 8, 9 [4,14) bin
  - Bin 2: 15, 21, 21 [14,24) bin
  - Bin 3: 24, 25, 26, 28, 29, 34 [24,+) bin

- Equi-frequency binning – for bin density of e.g., 3:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Equi-frequency binning – for bin density of e.g., 3:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smooth by Bin Means:
  - Bin 1 (Smoothing by Mean): 9 (average of 4, 8, 9, 15)
  - Bin 2 (Smoothing by Mean): 22.75 (average of 21, 21, 24, 25)
  - Bin 3 (Smoothing by Mean): 29.25 (average of 26, 28, 29, 34)
- Smoothing by Bin Boundaries:
  - Bin 1 (Smoothing by Boundaries): 4, 4, 4, 15
  - Bin 2 (Smoothing by Boundaries): 21, 21, 25, 25
  - Bin 3 (Smoothing by Boundaries): 26, 26, 26, 34