DATA ANALYTICS

Unit 1: Data Integration, Cleaning and Reduction

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1: Data Integration

Mamatha H R

Department of Computer Science and Engineering

Data Integration

- Data analysis often requires data integration—the merging of data from multiple data stores into a coherent store.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data analysis process.
- The semantic heterogeneity and structure of data pose great challenges in data integration.
- How can we match schema and objects from different sources?
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources

Data Integration

- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes can be detected using *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (for Categorical/ Nominal Data)

- **χ^2 (chi-square) test for independence of two variables in a contingency table**
- Null hypothesis: the two variables are independent
- Alternative hypothesis: the two variables are not independent
- **χ^2 statistic**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- ‘Expected’= what would we ‘expect’ if the null hypothesis were true?
- Larger the χ^2 value, the more likely the variables are correlated
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
- Can be used for categorical variables where entries are numbers (counts) and not percentages or fractions (for example, 20% of 200 has to be entered as 40 in the table)
- Correlation does not imply causation
  - The number of hospitals and number of car-thefts in a city may *appear to be* correlated
  - Both are causally linked to a third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Sum(col.)	300	1200	1500

Actual distribution (observed)

	Play chess	Not play chess	Sum (row)
Like science fiction	90	360	450
Not like science fiction	210	840	1050
Sum(col.)	300	1200	1500

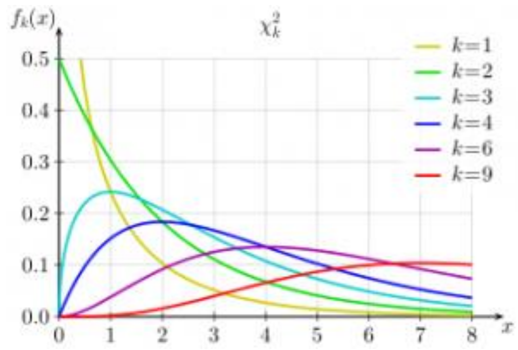
Expected distribution

$$e_{ij} = \frac{sum(A = a_i) * sum(B = b_j)}{N}$$

- X^2 (chi-square) calculation

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Degrees of freedom, k = (no_of_rows-1)(no_of_columns-1) = 1
- It shows that like_science_fiction and play_chess are correlated in the group



Correlation Analysis (Numeric Data)

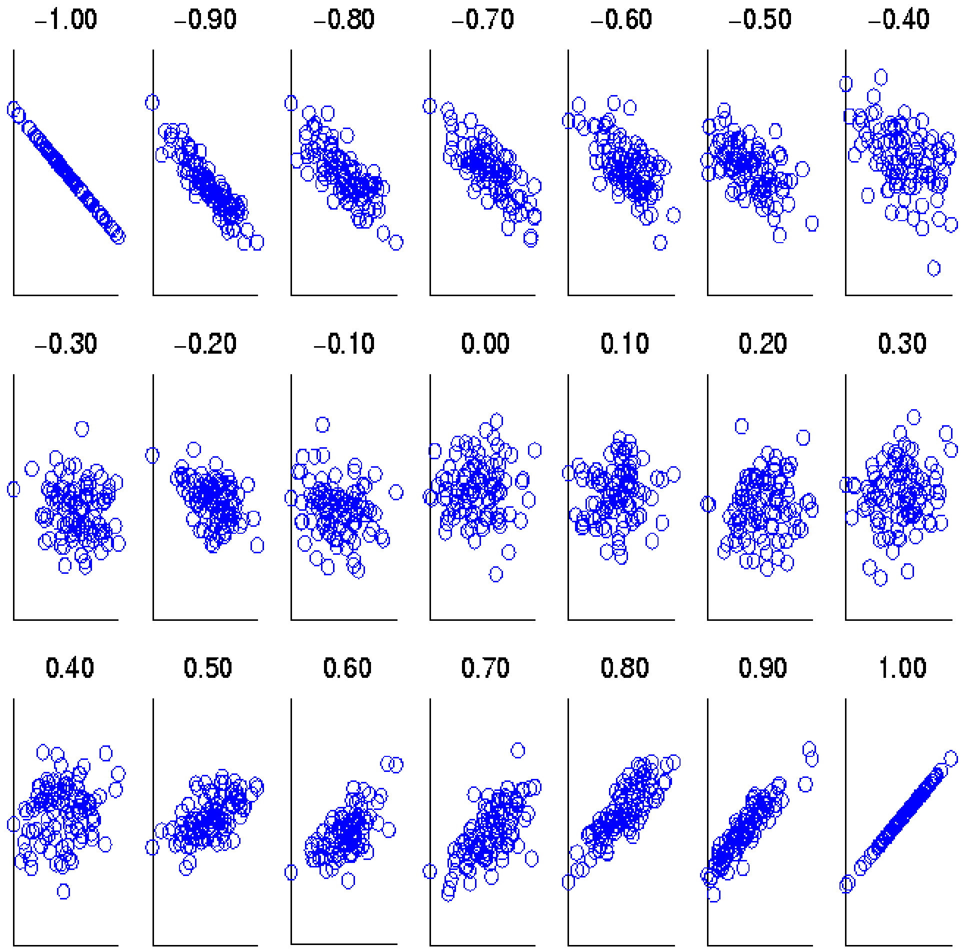
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the inner product $A^T B$ or sum of the point-wise product of *A* and *B*.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A^T B$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

Covariance (Numeric Data)

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$, but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply statistical independence

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

https://mathcs.clarku.edu/~djoyce/ma217/covar.pdf

Co-Variance: An Example

Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

$E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$

$E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$

$Cov(A,B) = (2\times5+3\times8+5\times10+4\times11+6\times14)/5 - 4 \times 9.6 = 4$

Thus, A and B rise together since Cov(A, B) > 0.

Tuple Duplication

- In addition to detecting redundancies between attributes, duplication should also be detected at the tuple level (e.g., where there are two or more identical tuples for a given unique data entry case).

- The use of denormalized tables (often done to improve performance by avoiding joins) is another source of data redundancy

Data Value Conflict Detection and Resolution

- Data integration also involves the detection and resolution of data value conflicts.

- For example, for the same real-world entity, attribute values from different sources may differ.

- This may be due to differences in representation, scaling, or encoding.

- For instance,a weight attribute may be stored in metric units in one system and British imperial units in another.

Exercise

- ❑ **Explain how redundancy is handled in data integration.**
- ❑ **Compare and contrast Correlation and Covariance.**

References

Text Book:

- **Data Mining: Concepts and Techniques** by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.

DATA ANALYTICS

Unit 1:Data Reduction

Mamatha H R, Gowri Srinivasa
Department of Computer Science and Engineering

Data Reduction

Data reduction: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

Data reduction strategies

- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation

- Numerosity reduction (some simply call it: Data Reduction)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation

- Data compression

Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

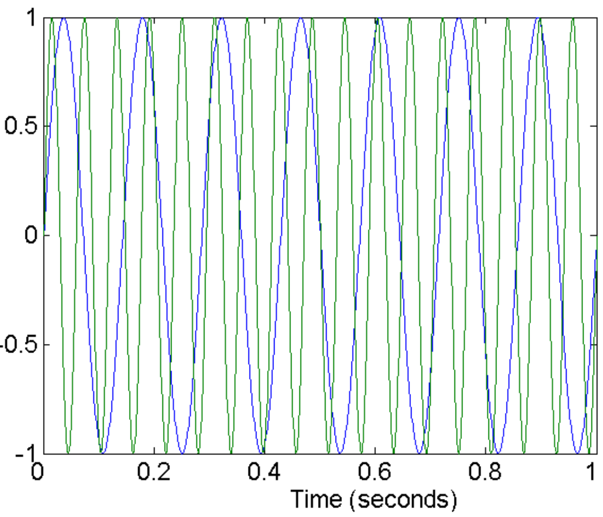
Data Reduction 1: Dimensionality Reduction

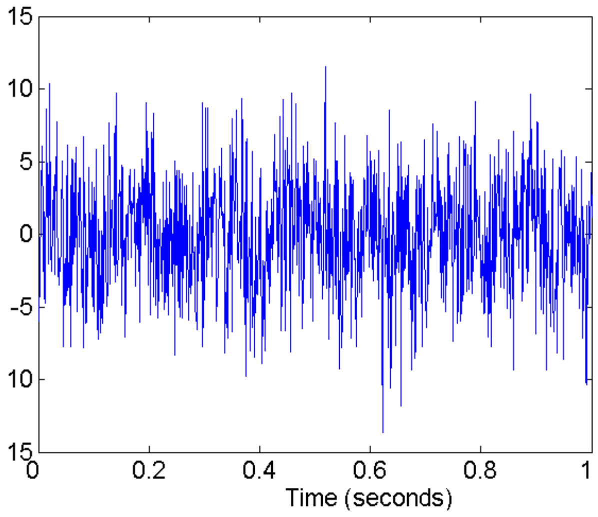
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

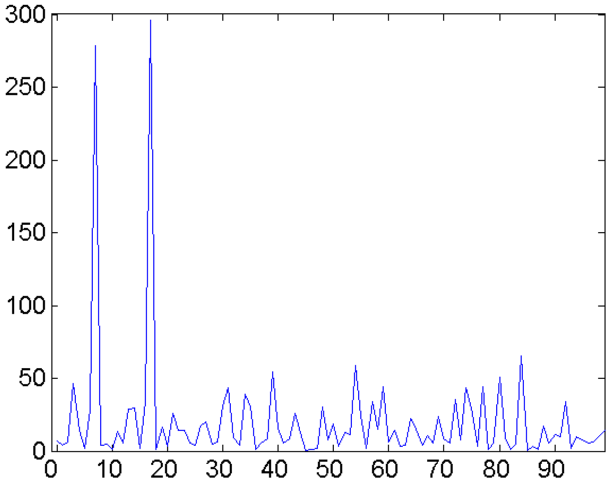
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

Mapping Data to a New Space

- **Fourier transform**
- **Wavelet transform**







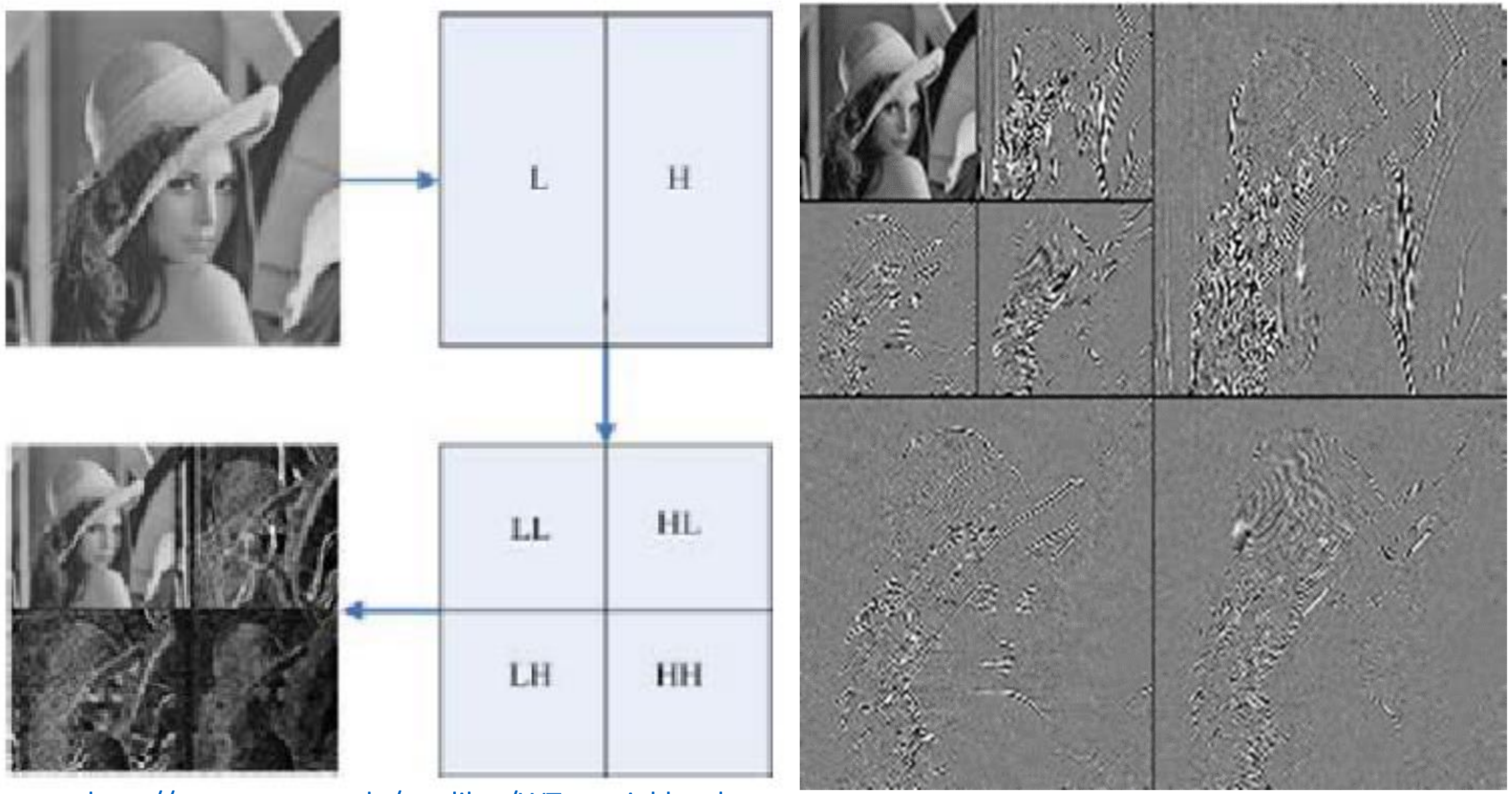
Two Sine Waves

Two Sine Waves + Noise

Frequency

What Is Wavelet Transform?

- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression

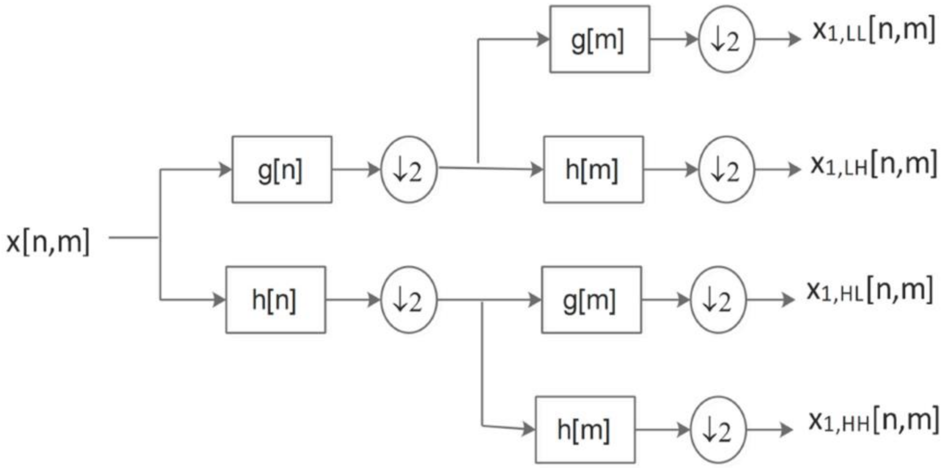


http://users.rowan.edu/~polikar/WTtutorial.html

Wavelet Transformation

Method:

- Length, L, must be an integer power of 2 (padding with 0's, when necessary)
- Each transform has 2 functions: smoothing (g), difference (h)
- Applies to pairs of data, resulting in two set of data of length L/2
- Applies the two functions recursively, until the desired level of decomposition is reached



$x(n,1)$	**56**	**40**	**8**	**24**	**48**	**48**	**40**	**16**
$g(n)=\frac{1}{2}[1,1]$	48	24	16	36	48	44	28	16
$h(n)=[1,-1]$	16	32	16	24	0	8	24	0
$g(n)\downarrow 2$	48		16		48		28	
$h(n)\downarrow 2$		32		24		8		0

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

https://www.math.aau.dk/digitalAssets/120/120646_r-2003-24.pdf

Why Wavelet Transform?

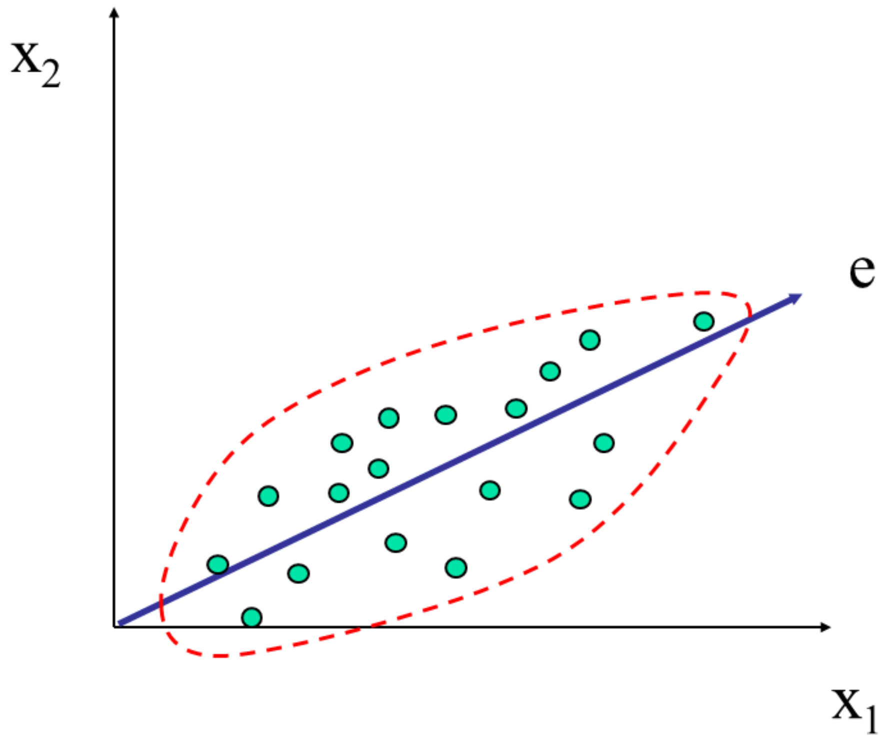
- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity O(N)
- Only applicable to low dimensional data

Principal component analysis

- Simplify data
- Understand relationship between variables
- Get an insight to patterns

Principal Component Analysis (PCA)

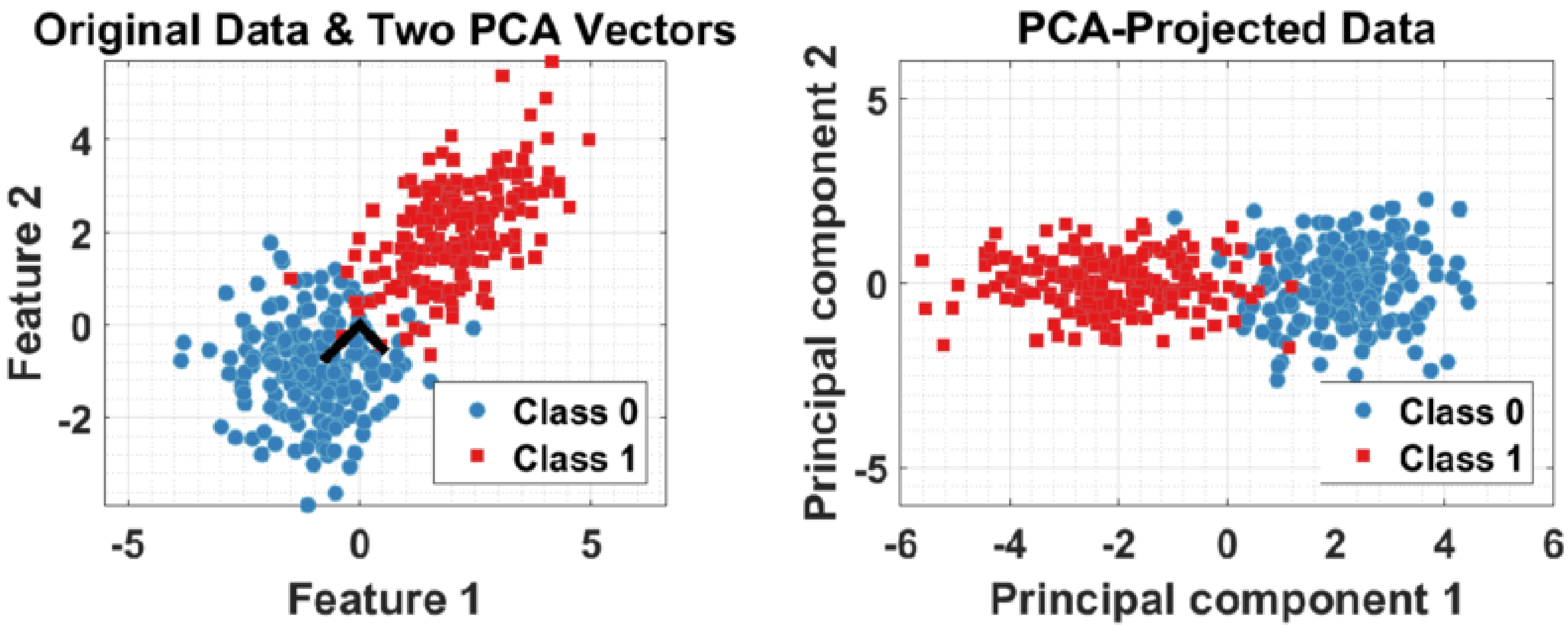
- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Principal Component Analysis (Steps)

- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
  - Works for numeric data only

PCA: Data in the Eigen Space – A different representation



https://www.researchgate.net/publication/320410861_Physically_Motivated_Feature_Development_for_Machine_Learning_Applications/figures?lo=1

Principal component analysis

- Get data
- Subtract mean
  (or bring it to zero mean, unit standard deviation form)
- Compute the covariance matrix
- Find Eigen values and Eigen vectors
- Select principal Eigen vectors (PCA)
  - Use proportion of variance retained by an eigen vector (using eigen values)
- Project data onto selected Eigen vectors
- Plot data

PCA example

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Covariance, Eigen analysis

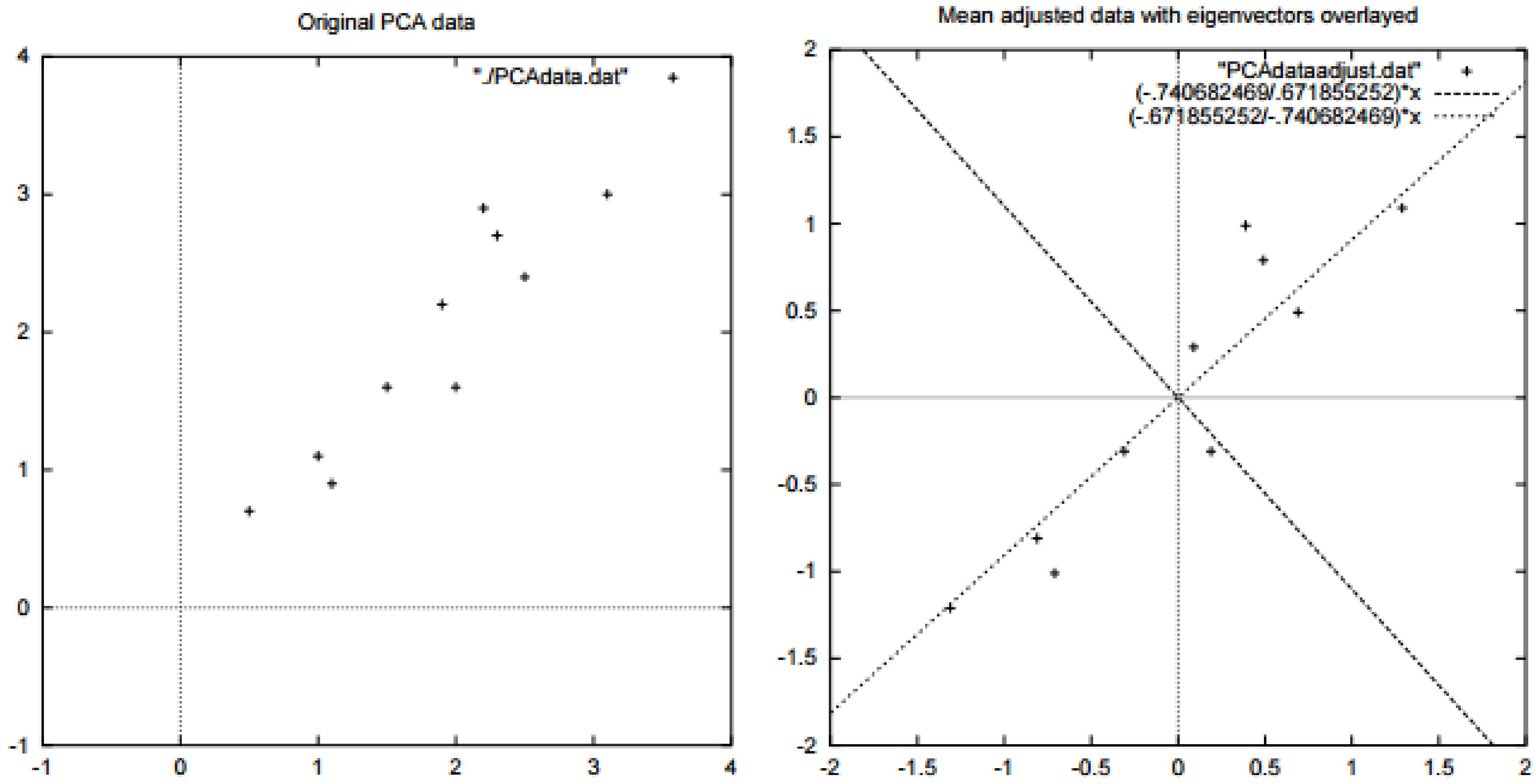
$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

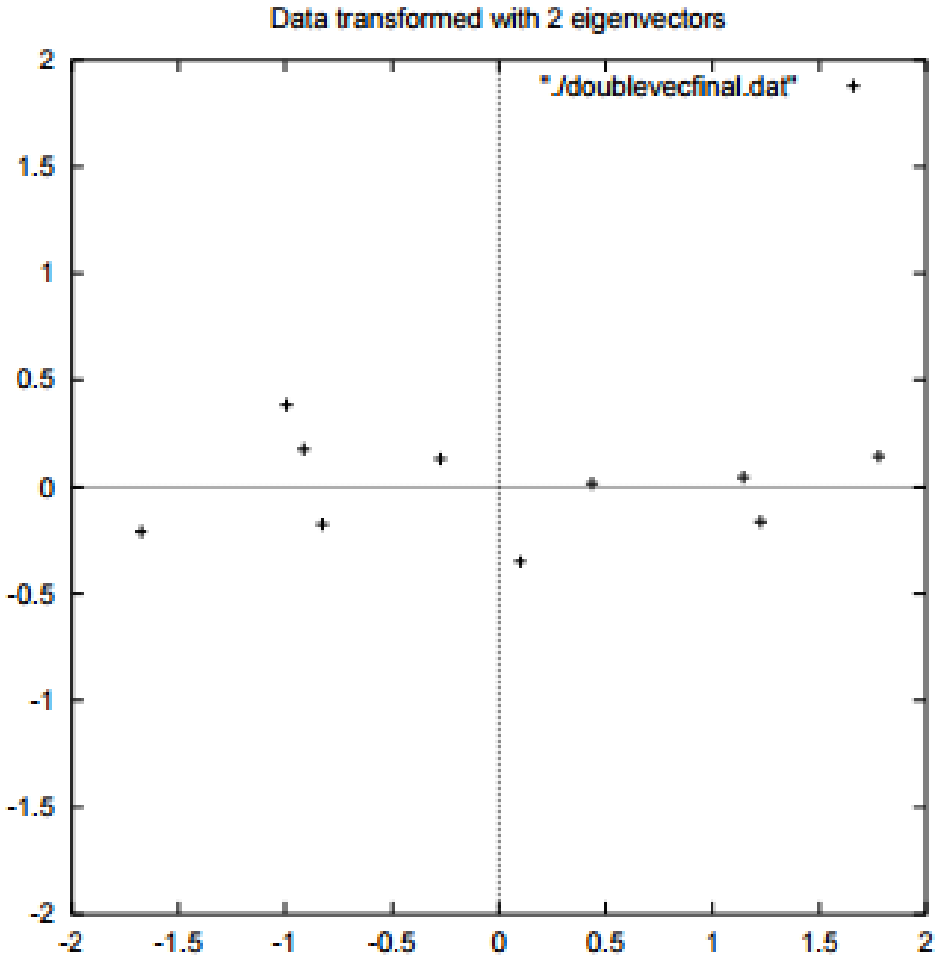
http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

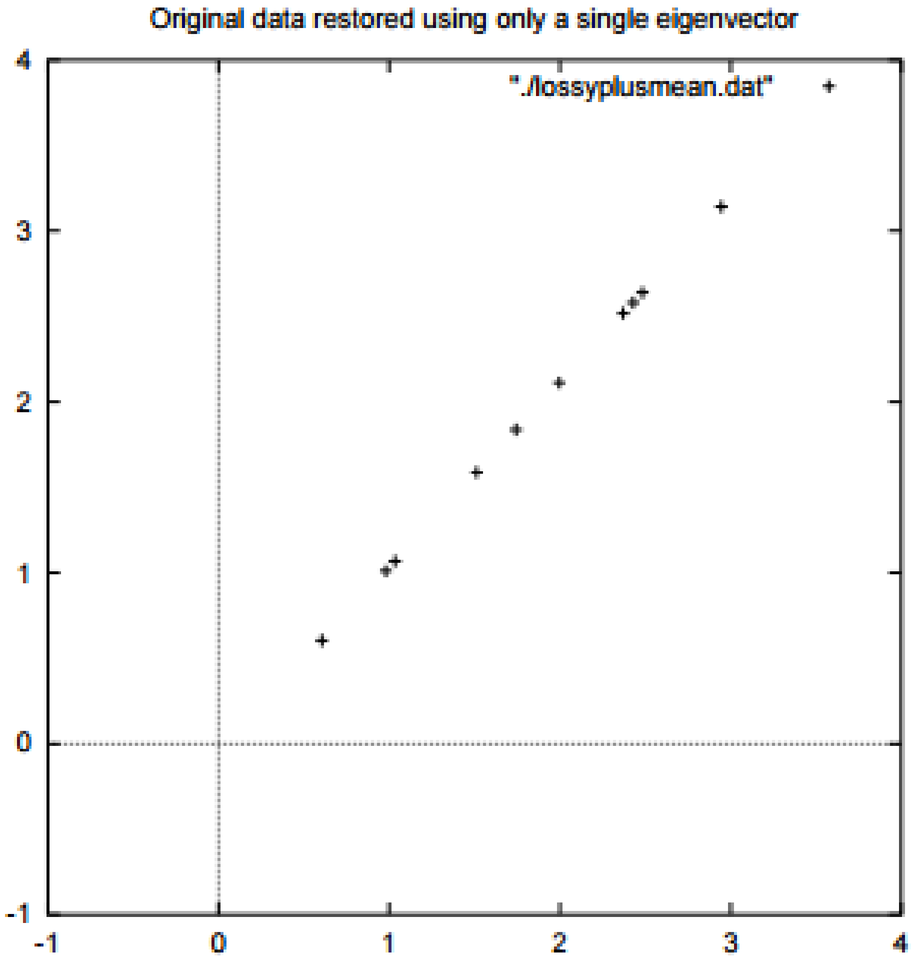
Choosing an appropriate 'axis'



http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

A new representation





http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

THANK YOU

Dr.Mamatha H R
Professor,Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834