DIMENSIONALITY REDUCTION

3rd Sem, MCA

Contents

- Dimensionality Reduction
  - Subset Selection,
  - PCA,
  - Factor Analysis,
  - Multidimensional Scaling,
  - Linear Discriminant Analysis.

Dimensionality

- **Dimensionality**: number of input variables or features for a dataset.
- **Dimensionality reduction**: techniques that reduce number of input variables in a dataset.

Benefits of applying Dimensionality Reduction

- Space required to store the dataset also gets reduced.
- Less Computation training time is required.
- Help in visualizing the data quickly.
- Removes the redundant features (if present).

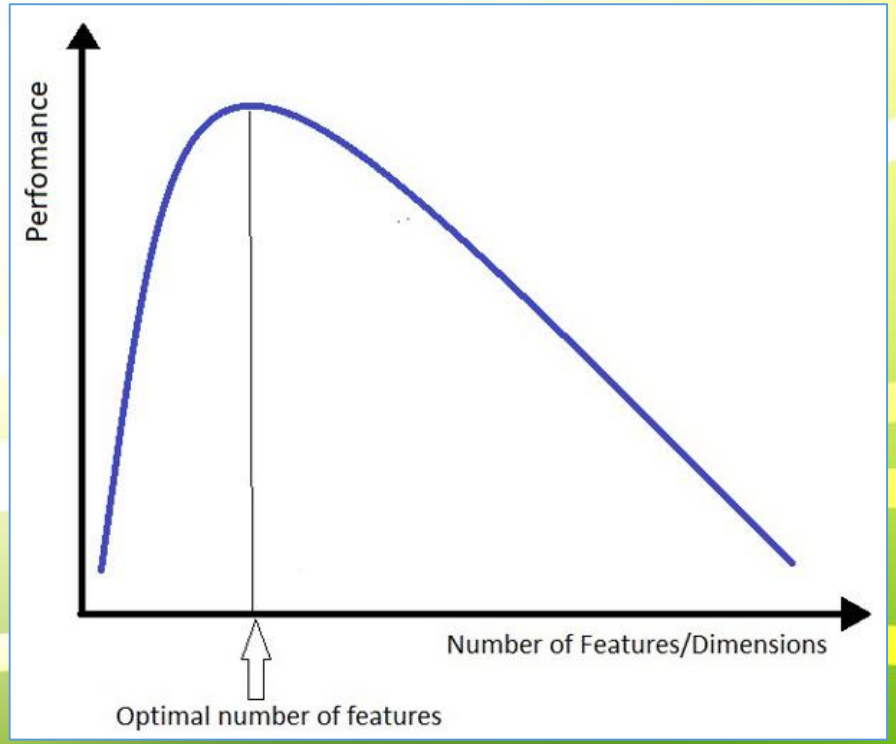
Disadvantages of dimensionality Reduction

- Some data may be lost due to dimensionality reduction.
- Sometimes the principal components required to consider are unknown.

Curse of dimensionality (CoD)

- Handling the high-dimensional data is very difficult in practice.
- **CoD** - difficulties related to training machine learning models due to high dimensional data
- If dimensionality of input dataset increases;
- number of samples also gets increased proportionally,
- chance of overfitting also increases.
- any machine learning model becomes more complex.
- If model is trained on high-dimensional data, it becomes overfitted and results in poor performance.

- Dimensionality reduction is very important.



Dimensionality Reduction

- some features can be quite redundant, adding noise to dataset and it makes no sense to have them.
- dimensionality reduction essentially transforms data from high-dimensional feature space to a low-dimensional feature space.
- also important that meaningful properties present in data are not lost during transformation.
- Dimensionality reduction is commonly used in data visualization to understand and interpret data.

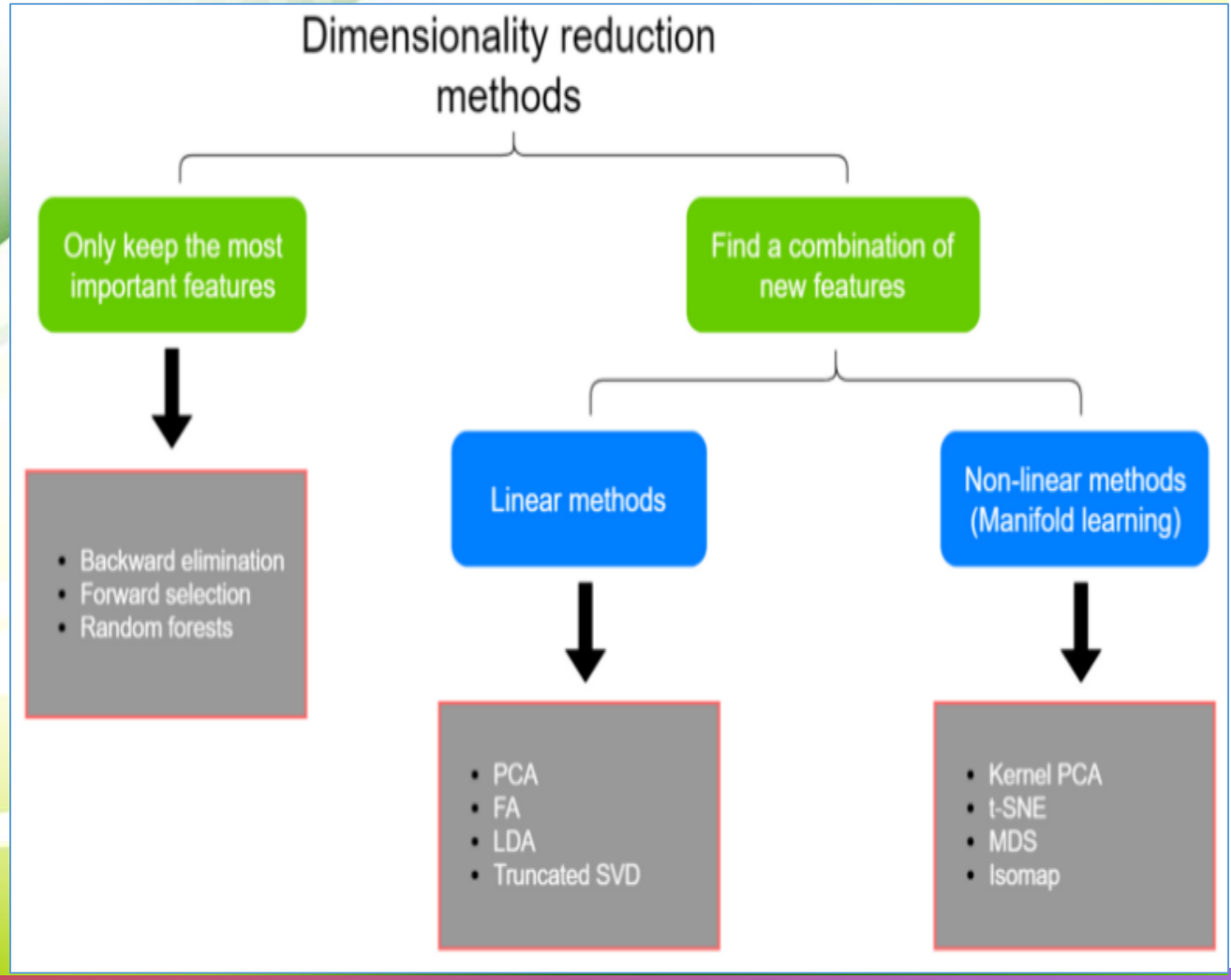
Two components of dimensionality reduction:

- **Feature selection**: find a subset of original set of variables. It involves three ways:
  - Filter
  - Wrapper
  - Embedded
- **Feature Extraction:** reduces data in a high dimensional space to a lower dimension space (lesser no. of dimensions).

Dimensionality Reduction

Two types of dimensionality reduction methods.

1. only keep most important features in dataset and removes redundant features.
   - no transformation applied to set of features.
2. find combination of new features.
   - appropriate transformation is applied to set of features.
   - new set of features contains different values instead of original values.
   - Linear methods
   - Non-linear methods (Manifold learning).



Dimensionality Reduction

- **Feature selection**: selecting optimal features from input dataset.
  - process of selecting subset of relevant features and leaving out irrelevant features present in dataset to build a model of high accuracy.
- Three methods are used for feature selection:
  - *Filter, Wrapper, & Embedded methods.*

Filter method: dataset is filtered, and a subset that contains only relevant features is taken.

- Common techniques of filters method:
  - **Correlation test**
  - **Chi-Square Test**
  - **ANOVA test**
  - **Information Gain, etc.**

Dimensionality Reduction

Wrappers Methods

- some features are fed to ML model, and evaluate performance.
- performance decides whether to add those features or remove to increase accuracy of model.
- more accurate than filtering method but complex to work.
- Common techniques of wrapper methods :
  - **Forward Selection, Backward Elimination, Bi-directional Elimination**
  - Best subset
  - Model selection based on validation dataset

Embedded Methods:

- check different training iterations of model and evaluate importance of each feature.
- Common techniques of Embedded methods are:
  - **LASSO, Elastic Net, Ridge Regression, etc.**

Dimensionality Reduction (DR)

Feature extraction

- process of transforming space containing many dimensions into space with fewer dimensions.
- useful to keep whole information but use fewer resources while processing the information.
- Common feature extraction techniques:
  - **Principal Component Analysis (PCA)**
  - **Linear Discriminant Analysis**
  - **Kernel PCA**
  - **Quadratic Discriminant Analysis**

DR - Feature Selection

Backward feature elimination:

- mainly used while developing Linear Regression or Logistic Regression model.

Steps :

1. firstly, all n variables of given dataset are taken to train the model.
2. performance of the model is checked.
3. remove one feature each time and train the model on n-1 features (repeat for n times),
   - compute the performance of the model each time.
4. check for the variable that has made smallest or no change in performance of model,
   - drop that variable or features; (remaining with 'n-1' features).
5. Repeat step 3-4 until no more feature can be dropped.

- Select optimum performance of model and maximum tolerable error rate to find optimal number of features require for the machine learning algorithms.

DR - Feature Selection

Backward feature elimination:

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy using all the variables = 92%

Variable_dropped	Accuracy
Calories_burnt	90%
Gender	91.60%
Plays_Sport?	88%

gender does not have a high impact on Fitness_Level variable → **dropped**.

DR - Feature Selection

Forward Feature Selection

- inverse process of backward elimination process.
- don't eliminate feature; instead, find/add the best features that can produce highest increase in performance of the model.

steps:

1. start with a single feature only,
2. progressively add all features, one at a time.
3. train the model on each feature separately.
4. feature with best performance is selected to be added.
5. process will be repeated until there is NO significant increase in performance of the model.

DR - Feature Selection

Forward Feature Selection:

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Variable used	Accuracy
Calories_burnt	87.00%
Gender	80.00%
Plays_Sport?	85.00%

- Calories_Burnt produced best result → **selected**.
- Next, repeat this process and add one variable to **Calories_Burnt** variable.

DR - Feature Selection

ID	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
AB101	1.0	0.0	0.0	1.0	9.84	14.395	81.0	NaN	16
AB102	1.0	NaN	0.0	NaN	9.02	13.635	80.0	NaN	40
AB103	1.0	0.0	NaN	1.0	9.02	13.635	80.0	NaN	32
AB104	NaN	0.0	NaN	1.0	9.84	14.395	75.0	NaN	13
AB105	1.0	NaN	0.0	NaN	9.84	14.395	NaN	16.9979	1
AB106	1.0	0.0	NaN	2.0	9.84	12.880	75.0	NaN	1
AB107	1.0	0.0	0.0	1.0	9.02	13.635	80.0	NaN	2
AB108	1.0	NaN	0.0	1.0	8.20	12.880	86.0	NaN	3
AB109	NaN	0.0	0.0	NaN	9.84	14.395	NaN	NaN	8
AB110	1.0	0.0	0.0	1.0	13.12	17.425	76.0	NaN	14

Missing Value Ratio

- If dataset has too many missing values, then drop those variables
- These variable do not carry much useful information.

Steps:

1. set a threshold level,
2. if a variable has missing values more than that threshold, drop that variable.

- The higher the threshold value, the more efficient the reduction.
  - Dropping a feature may result in information loss.
- Replacing missing data with some substitute value to retain most of the data/information of the dataset:
  - **Mean Imputation**
  - **KNN Imputation**
  - **Hot Deck Imputation**
  - **Cold Deck Imputation**
  - **Stocastic Regression Imputation**

DR - Feature Selection

Missing Value Ratio

$$\text{Ratio of missing values} = \frac{\text{Number of missing values}}{\text{Total number of observations}} * 100$$

ID	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
AB101	1.0	0.0	0.0	1.0	9.84	14.395	81.0	NaN	16
AB102	1.0	NaN	0.0	NaN	9.02	13.635	80.0	NaN	40
AB103	1.0	0.0	NaN	1.0	9.02	13.635	80.0	NaN	32
AB104	NaN	0.0	NaN	1.0	9.84	14.395	75.0	NaN	13
AB105	1.0	NaN	0.0	NaN	9.84	14.395	NaN	16.9979	1
AB106	1.0	0.0	NaN	2.0	9.84	12.880	75.0	NaN	1
AB107	1.0	0.0	0.0	1.0	9.02	13.635	80.0	NaN	2
AB108	1.0	NaN	0.0	1.0	8.20	12.880	86.0	NaN	3
AB109	NaN	0.0	0.0	NaN	9.84	14.395	NaN	NaN	8
AB110	1.0	0.0	0.0	1.0	13.12	17.425	76.0	NaN	14

Variable	Missing value ratio
ID	0%
season	20%
holiday	30%
workingday	30%
weather	30%
temp	0%
atemp	0%
humidity	20%
windspeed	90%
count	0%

- "windspeed" variable has missing value ratio of 90% > threshold of 70% → **dropped**.

DR - Feature Selection

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	1	7	8.20	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

Low Variance Filter

- Variance is measure of variability/dispersion.
- This method acts similar like missing value ratio technique.
- Variables with little changes in data have less information
- low variance features will not significantly affect target variable.

Steps:

1. set a threshold level,
2. calculate variance of each variable,
3. all features with variance lower than threshold are dropped.

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

DR - Feature Selection

Low Variance Filter

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	1	7	8.20	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

- variance of ‘f5’ variable is zero → less impact on target variable → **drop.**
- set a threshold value of variance.
- if variance of a variable < threshold → drop.
- **Variance is range-dependent** → apply **normalization** before applying this technique.

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

DR - Feature Selection

High Correlation filter

- Correlation describes the size and direction of a relationship between two variables.
- High Correlation refers that two variables carry approximately **similar/related/dependent** information.
- This may lead the performance of model to degrade.

Steps:

1. Set a threshold value.
2. calculated the correlation coefficient between the pair of variables.
3. If this value is higher than threshold value, then remove one of variables from dataset.

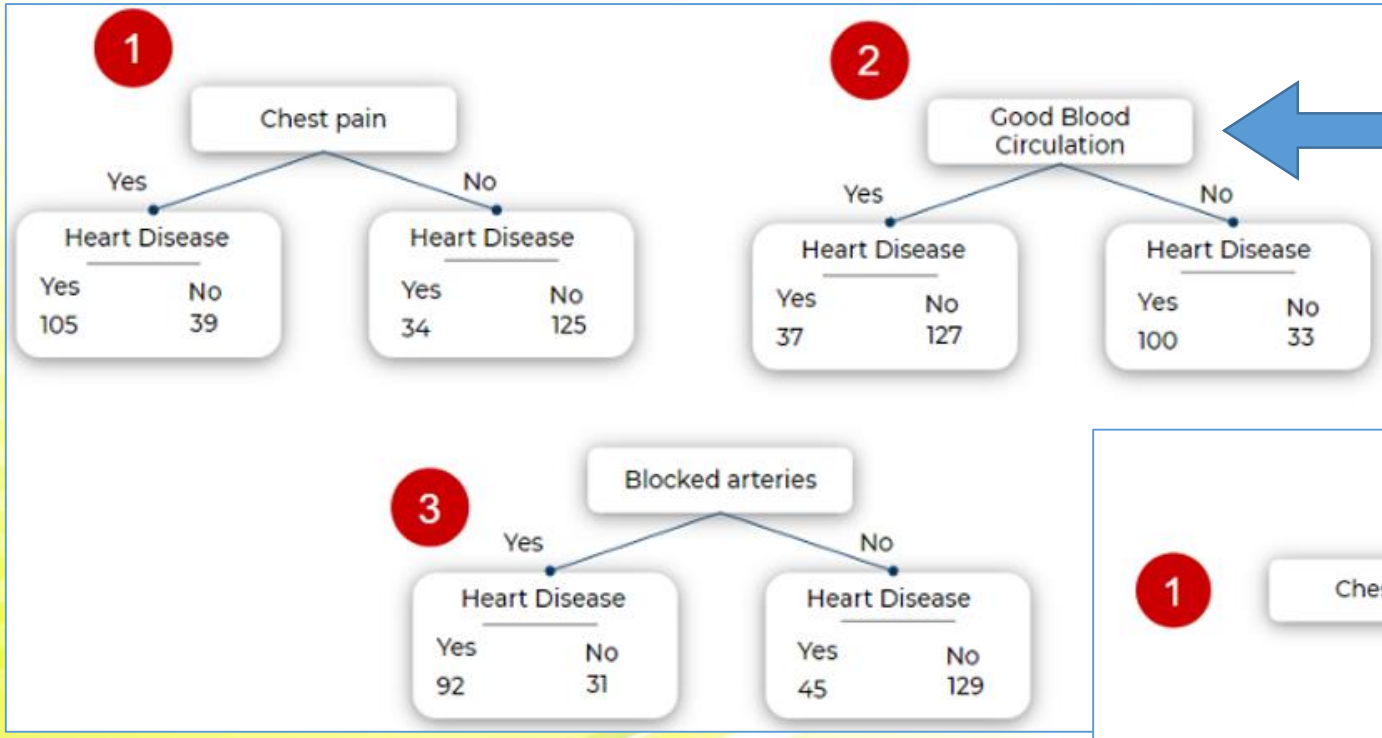
DR - Feature Selection

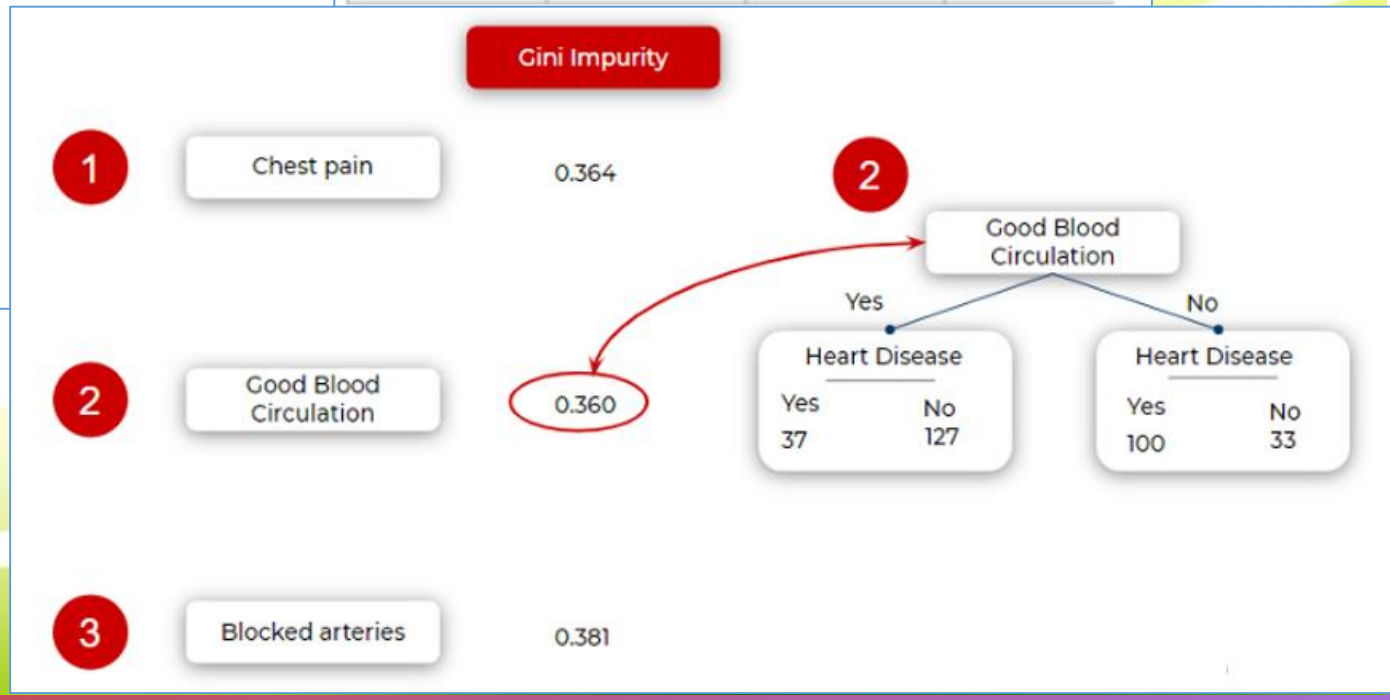
Random Forest

- tree-based model; widely used for regression and classification tasks on non-linear data.
- Each tree of random forest can calculate importance of a feature according to its ability to increase pureness of leaves.
  - Uses ‘gini’ coefficient to calculate feature_importance.
  - ***Gini****: a measure of quality of a split of internal nodes in the tree.*
  - higher the increment in leaves purity, higher the importance of feature.

DR - Feature Selection

Chest pain	Good blood circulation	Blocked arteries	Heart disease
yes	no	yes	no
yes	yes	yes	yes
no	no	yes	no
yes	yes	no	yes
...	...	...	...





DR - Feature Selection

LASSO (Least Absolute Shrinkage and Selection Operator)

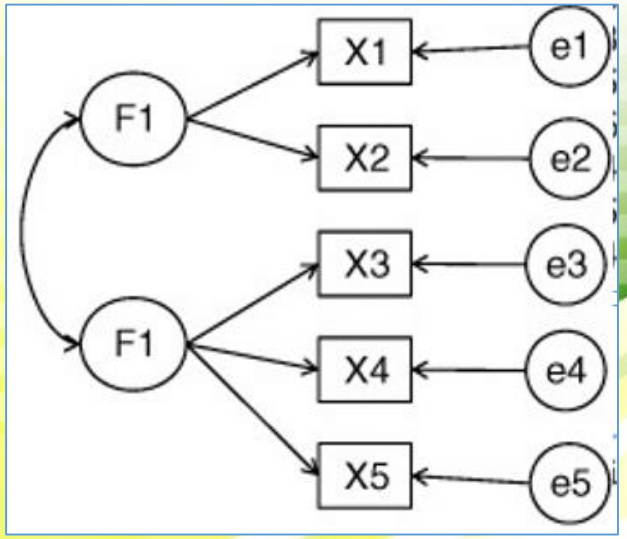
- Used for feature selection and regularization of data models.
- Regularization is one approach to tackle the problem of overfitting.
- LASSO introduces parameters to the sum of model, giving it an upper bound that acts as a constraint for the sum to include absolute parameters within an allowable range.
- LASSO method regularizes model parameters by shrinking regression coefficients, reducing some of them to zero.
- feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model.
- significant in minimization of prediction errors.
- Leads to high prediction accuracy.
- shrinkage of coefficients → reduces variance and minimizes bias.
- It performs best when number of observations is low and number of features is high.

DR - Feature Extraction

Factor Analysis

- each variable is kept within a group (**factor)** according to correlation with other variables,
  - variables within a group can have a high correlation between themselves,
  - but they have a low correlation with variables of other groups.
    - **example**, two variables Income and spend, have a high correlation → people with high income spends more, and vice versa.
  - number of these factors will be reduced as compared to original dimension of dataset.

DR - Factor Analysis



rating	complaints	privileges	learning	raises	critical	advance
43	51	30	39	61	92	45
63	64	51	54	63	73	47
71	70	68	69	76	86	48
61	63	45	47	54	84	35
81	78	56	66	71	83	47
43	55	49	44	54	49	34

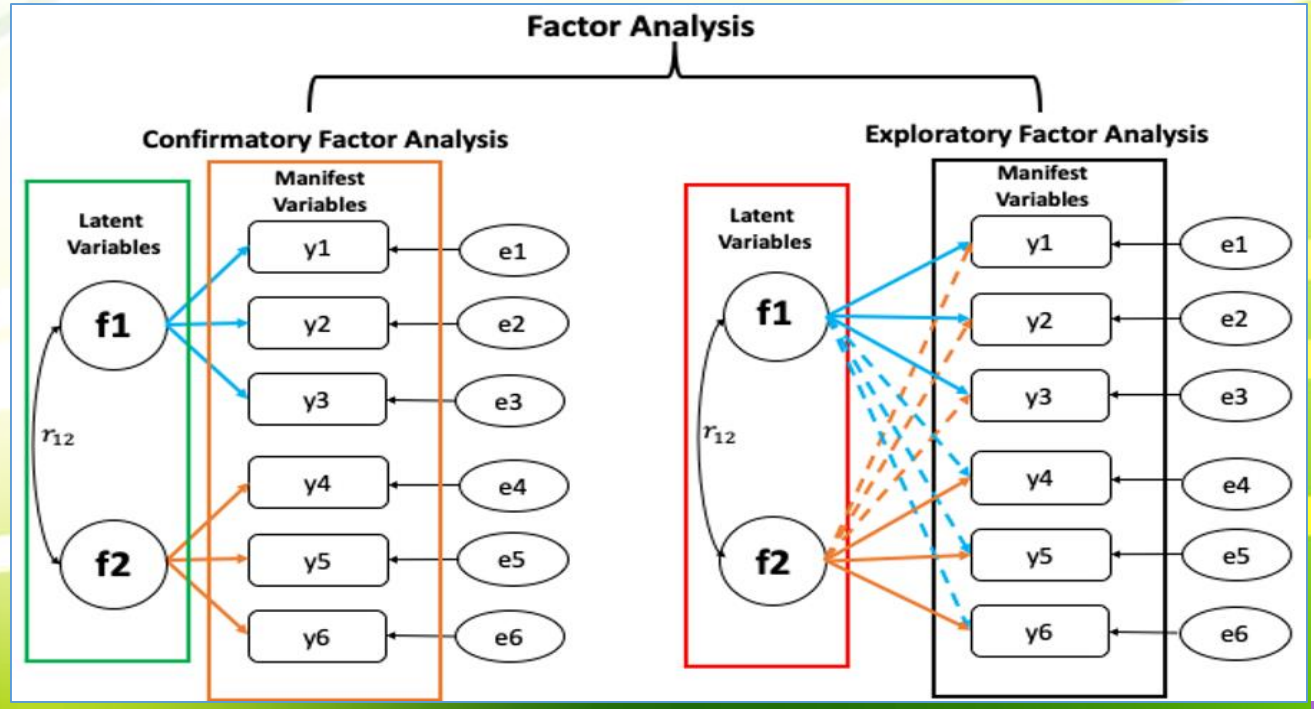
Group-1 Work Culture | Group-2 Learning Opportunities | Group-3 Promotions

rating	complaints	privileges	learning	raises	critical	advance
43	51	30	39	61	92	45
63	64	51	54	63	73	47
71	70	68	69	76	86	48
61	63	45	47	54	84	35
81	78	56	66	71	83	47
43	55	49	44	54	49	34

Average Ratings

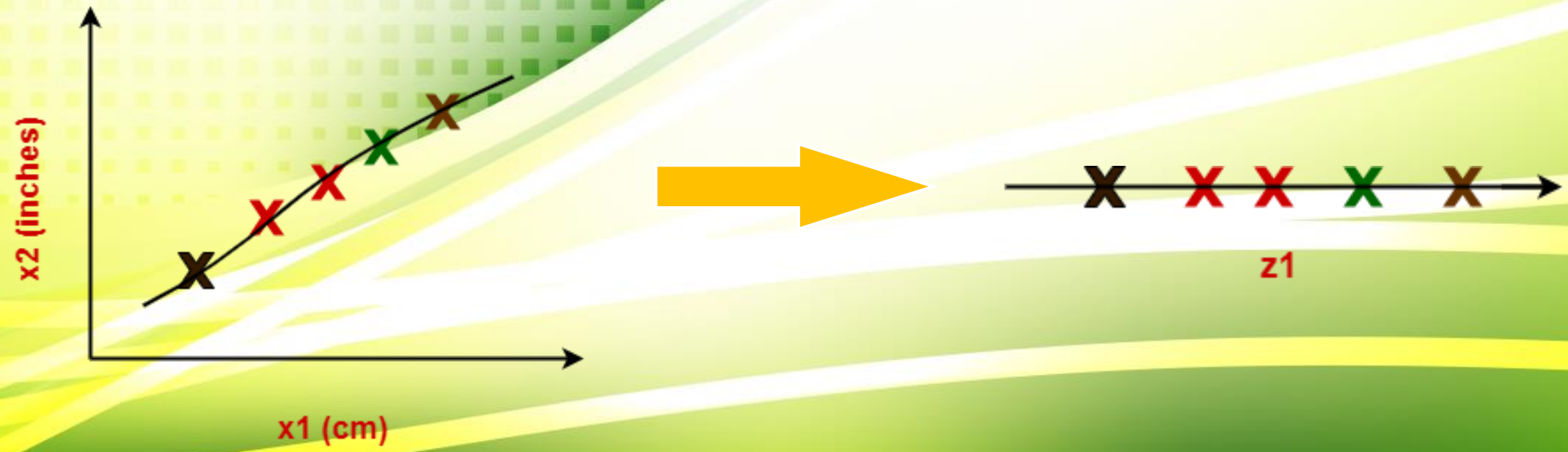
DR - Factor Analysis

- In exploratory factor analysis (EFA), all measured variables are related to every latent variable.
- In confirmatory factor analysis (CFA), researchers can specify the number of factors required in the data and which measured variable is related to which latent variable.
- EFA is used when it is not known how many factors there are between the items & which factors are determined by which items.
- CFA is used if there is a strong theory about the structure.



DR - Principal Component Analysis (PCA)

- PCA is a linear dimensionality reduction technique that transforms a set of correlated variables (p) into a smaller k (k<p) number of uncorrelated variables called ***principal components (PC)*** while retaining as much of the variation in the original dataset as possible.



DR - Principal Component Analysis (PCA)

Steps:

1. Original Data
2. Normalize the original data (mean =0); Center the data.
3. Calculate covariance matrix
4. Calculate Eigen values, Eigen vectors of covariance matrix
5. Normalize Eigenvectors
6. Calculate Principal Component (PC)
7. Plot the graph for orthogonality between PCs

DR - Principal Component Analysis (PCA)

- **Variance** – for calculating the variation of data distributed across dimensionality of graph.
- **Covariance** – calculating dependencies and relationship between features.

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- **Standardizing data** – Scaling dataset within a specific range for unbiased output.
- **Covariance matrix** – Used for calculating interdependencies between features & helps to reduce it to improve performance.
- **EigenValues** - Magnitude of Eigenvector. Eigenvalue indicates variance in a particular direction.
- **EigenVectors** – special set of vectors that help to understand structure & property of data that would be PCs.
  - Helps to find largest variance that exists in dataset to calculate Principal Component.
  - eigenvector indicates about expanding or contracting X-Y (2D) graph without altering the direction.
- The highest eigenvalues and their corresponding eigenvectors make the most important principal components.

DR - Principal Component Analysis (PCA)

- Covariance Matrix (dispersion matrix and variance-covariance matrix) is a measure of how much two random variables gets change together.

$$\begin{bmatrix} Var(x_1) & \cdots & Cov(x_n, x_1) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix}$$

$$var(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$cov(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix}$$

Covariance Matrix 2 × 2

$$\begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix}$$

Covariance Matrix 3 × 3

DR - Principal Component Analysis (PCA)

Find the Covarience matrix for following data.

Student	Score	Age
1	68	29
2	60	26
3	58	30
4	40	35

$\mu_x = 56.5, n = 4$

$var(x) = [(68 - 56.5)^2 + (60 - 56.5)^2 + (58 - 56.5)^2 + (40 - 56.5)^2] / 4 = 104.75$

$\mu_y = 30, n = 4$

$var(y) = [(29 - 30)^2 + (26 - 30)^2 + (30 - 30)^2 + (35 - 30)^2] / 4 = 10.5$

$$var(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$cov(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix}$$

$$cov(x, y) = \frac{\sum_1^4 (x_i - \mu_x)(y_i - \mu_y)}{4}$$

$cov(x, y) = -27$

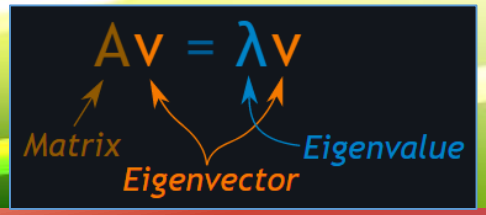
$$\begin{bmatrix} 104.7 & -27 \\ -27 & 10.5 \end{bmatrix}.$$

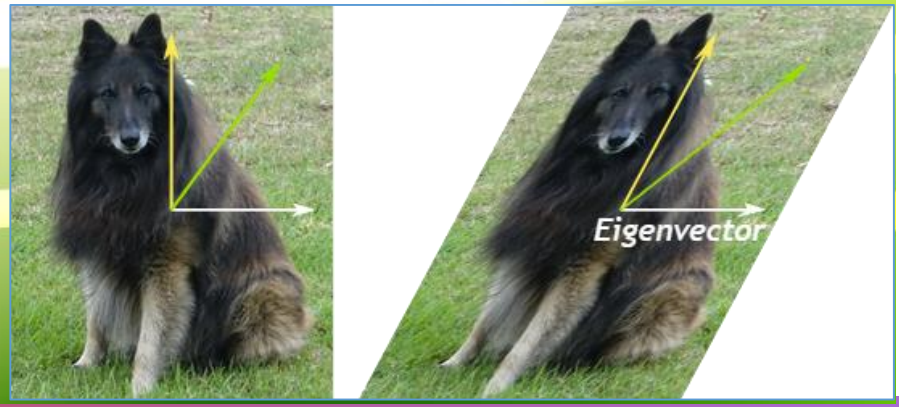
DR - Principal Component Analysis (PCA)

- Eigenvalues are associated with eigenvectors in Linear algebra., combinedly used in analysis of linear transformations.
  - special set of scalar values associated with the set of linear equations (most probably in matrix equations).
- Eigenvectors is a non-zero vector that do not change the direction when any linear transformation is applied. It can be changed at most by its scalar factor by linear transformations. And the corresponding factor which scales the eigenvectors is called an eigenvalue.
- If A is a linear transformation from a vector space V and **x** is a vector in V, which is not a zero vector, then v is an eigenvector of A if A(X) is a scalar multiple of **x**.
- If A is an "n × n" matrix and λ is an eigenvalue of matrix A, then **x**, a non-zero vector, is called as eigenvector if it satisfies the given below expression;

Ax = λx **(**scalar value "**λ"** is an eigenvalue of A.)

(x is an eigenvector of A corresponding to eigenvalue, λ.)





DR - Principal Component Analysis (PCA)

Find the eigenvalues of the 2 x 2 matrix

$$A = \begin{bmatrix} 0 & -2 \\ 3 & 4 \end{bmatrix}$$

$$A \cdot v = \lambda \cdot v$$

$$|A - \lambda I| = 0$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\left| \begin{bmatrix} 0 & -2 \\ 3 & 4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} 0 & -2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} -\lambda & -2 \\ 0 & 4-\lambda \end{vmatrix} = 0$$

$-\lambda(4 - \lambda) - (-2)(0) = 0$

$-4\lambda + \lambda^2 = 0$

$\lambda(\lambda - 4) = 0$

$\lambda = 0$ or $\lambda - 4 = 0$

two eigenvalues of the given matrix are λ = 0 and λ = 4.

DR - Principal Component Analysis (PCA)

Find the eigenvalues of the 3 x 3 matrix

$$A = \begin{bmatrix} 4 & 6 & 10 \\ 3 & 10 & 13 \\ -2 & -6 & -8 \end{bmatrix}$$

$$A \cdot v = \lambda \cdot v$$

$$|A - \lambda I| = 0$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$det(A-\lambda I) = \left| \begin{bmatrix} 4 & 6 & 10 \\ 3 & 10 & 13 \\ -2 & -6 & -8 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right|$$

$$= \left| \begin{bmatrix} 4 & 6 & 10 \\ 3 & 10 & 13 \\ -2 & -6 & -8 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right|$$

$$det(A-\lambda I) = \begin{vmatrix} 4-\lambda & 6 & 10 \\ 3 & 10-\lambda & 13 \\ -2 & -6 & -8-\lambda \end{vmatrix}$$

$$= (4-\lambda) \begin{vmatrix} 10-\lambda & 13 \\ -6 & (-8-\lambda) \end{vmatrix} - (6) \begin{vmatrix} 3 & 13 \\ -2 & -8-\lambda \end{vmatrix} + 10 \begin{vmatrix} 3 & 10-\lambda \\ -2 & -6 \end{vmatrix}$$

$\det (A-\lambda I) = -\lambda^3 + 6\lambda^2 - 6\lambda - 8 - 12 + 18\lambda + 20 - 20\lambda$

$-\lambda^3 + 6\lambda^2 - 8\lambda = 0$

Three eigenvalues of the given matrix are λ = 0, 2 & 4.

DR - Principal Component Analysis (PCA)

- Normalized eigenvector is an eigenvector having unit length.
- It can be found by simply dividing each component of the vector by the length of the vector.

Normalize the eigenvector to Unit length.

$$\begin{bmatrix} 1 \\ -5 \\ -1 \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix}$$

$$L = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

$$L = \sqrt{1^2 + (-5)^2 + 1^2}$$

$$L = 3\sqrt{3}$$

Normalized the eigenvector to Unit length:

$$\begin{bmatrix} \frac{1}{3\sqrt{3}} \\ \frac{-5}{3\sqrt{3}} \\ \frac{-1}{3\sqrt{3}} \end{bmatrix}$$

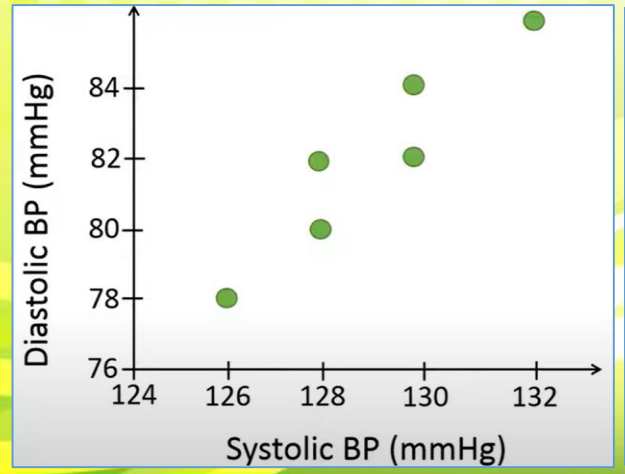
DR - Principal Component Analysis (PCA)

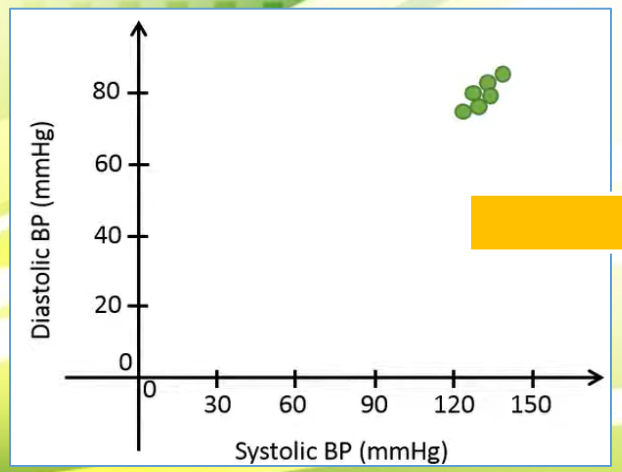
Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

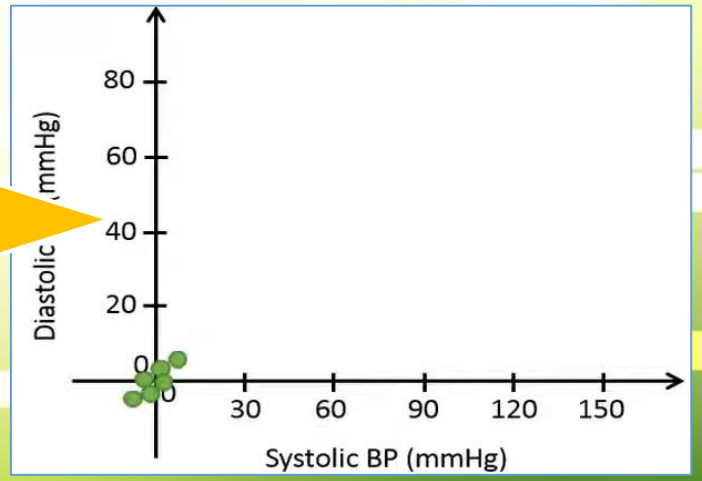
Center the Data

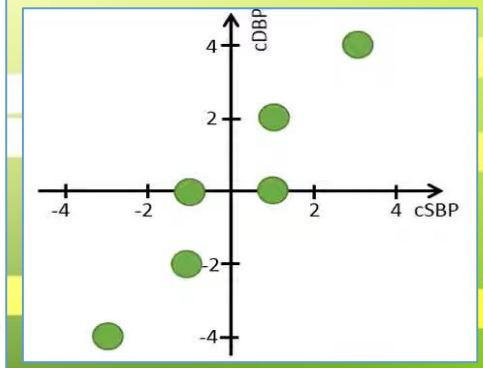
Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

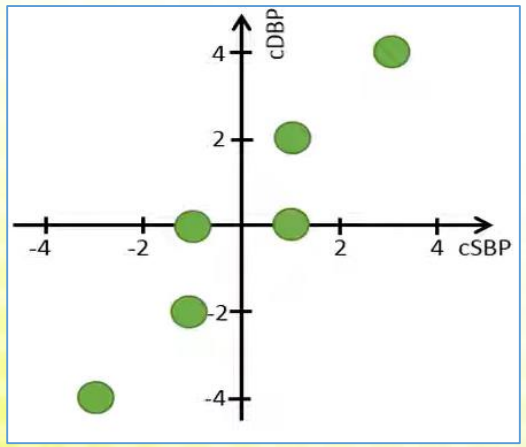








DR - Principal Component Analysis (PCA)



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

Covariance Matrix

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$\text{cov}(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

Var, Cov calculation is simpler with Mean=0 (centered data)

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = \big((-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4\big)/(6-1) = 28/5 = 5.6$$

DR - Principal Component Analysis (PCA)

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Eigenvalue

$|A - \lambda I| = 0$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

λ = 0.32 or 12.08

$$\det\left(\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$$(4.4-\lambda)(8.0-\lambda) - 5.6 \cdot 5.6 = 0$$

$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\det\left(\begin{bmatrix} (4.4-\lambda) & 5.6 \\ 5.6 & (8.0-\lambda) \end{bmatrix}\right) = 0$$

DR - Principal Component Analysis (PCA)

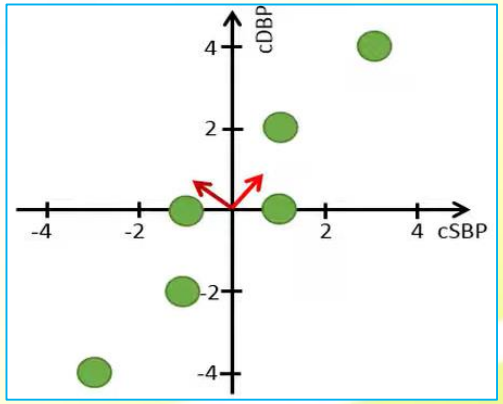
$$A \cdot v = \lambda \cdot v$$

Eigenvalue λ = 0.32 or 12.08

Eigenvector

$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$



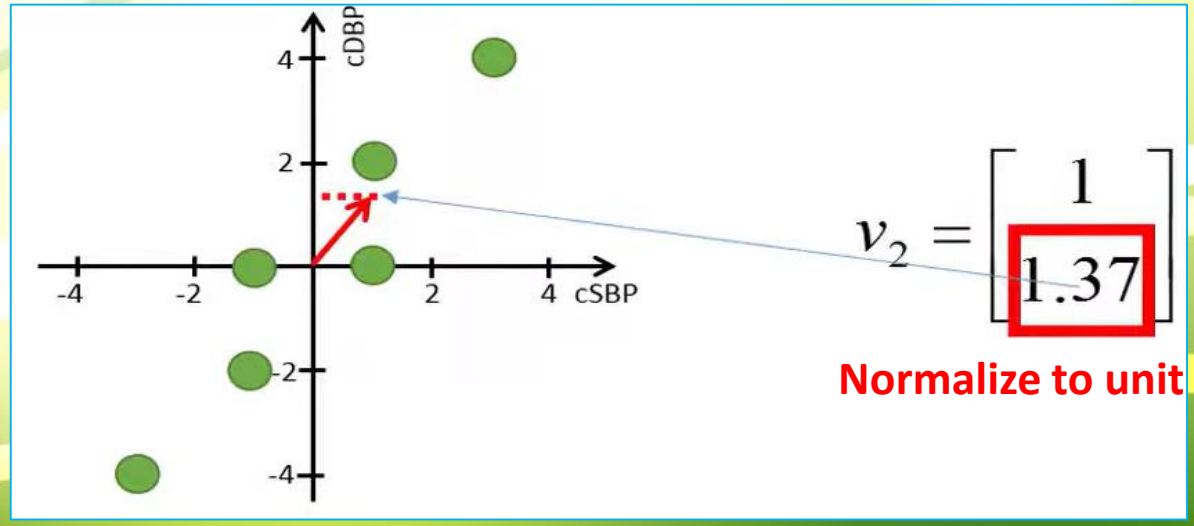
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$

$$5.6y = 7.68x$$
$$5.6x = 4.08y$$

$$y = 1.37x$$
$$1.37x = y$$

$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$



$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$

Normalize to unit length

$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix}$$

DR - Principal Component Analysis (PCA)

$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

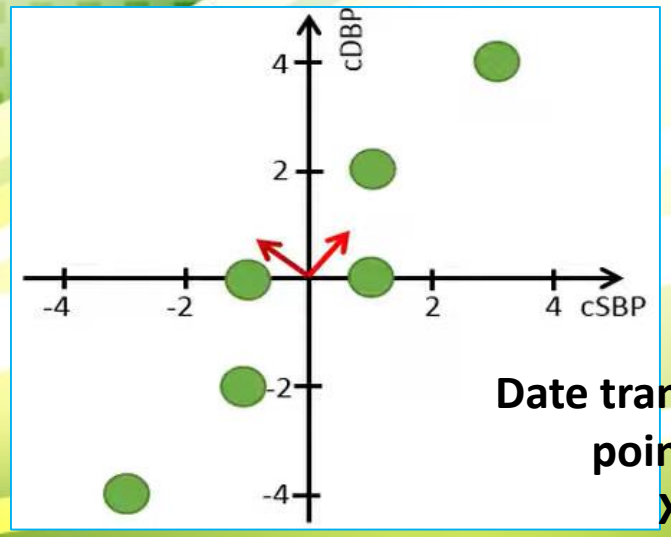
$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

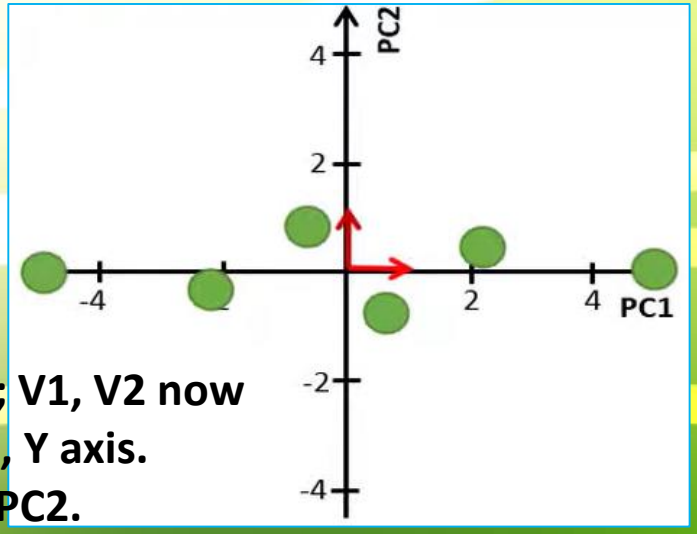
Eigenvector with higher Eigenvalue is primary/first Eigenvector.

$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

$$D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$

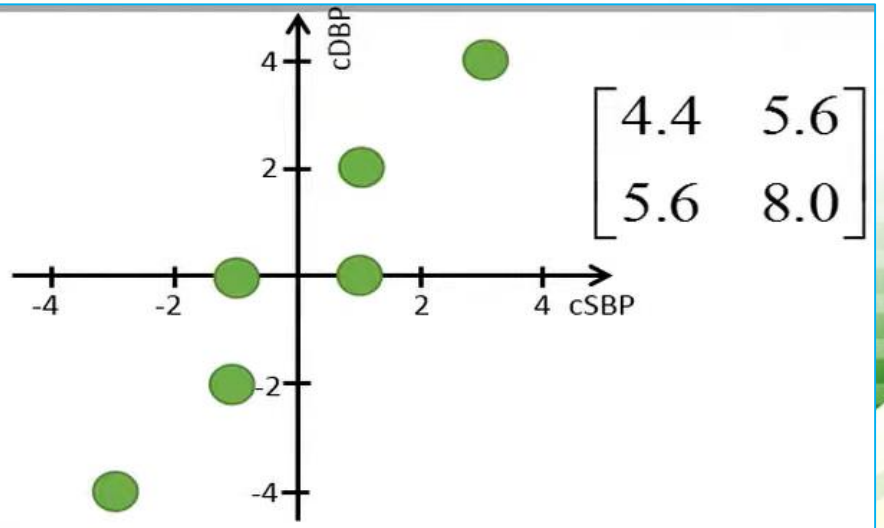
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4





Date transformation; V1, V2 now point towards X, Y axis.
X, Y → PC1, PC2.

DR - Principal Component Analysis (PCA)



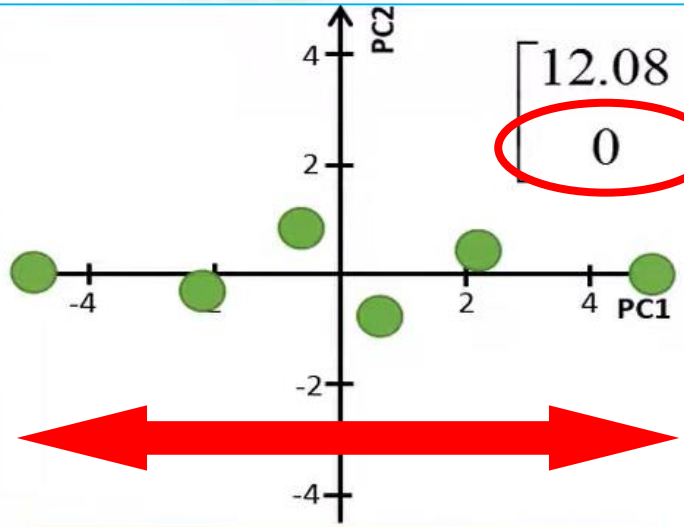
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$$

Centered data

vs transformed data



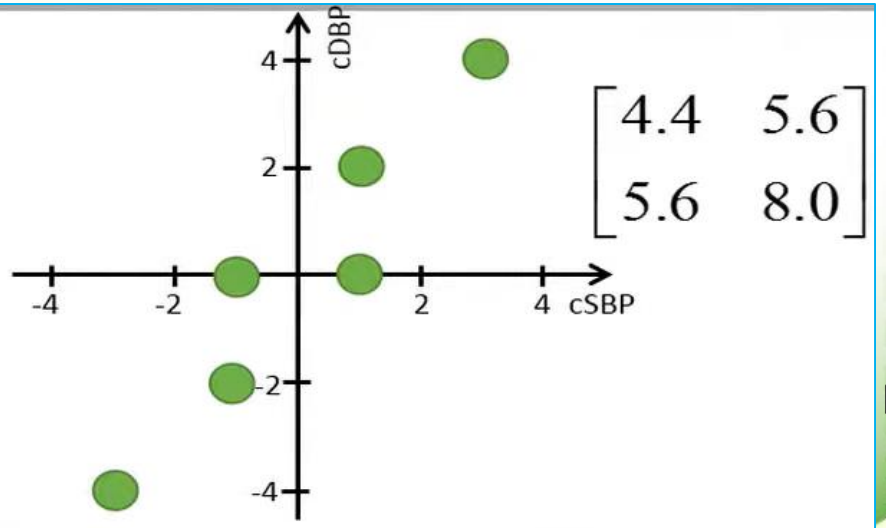
$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

PC1 contains significant variance; PC2 has min.

$$\%\,var = \frac{12.08}{12.08+0.32} = 97.4\%$$

DR - Principal Component Analysis (PCA)



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

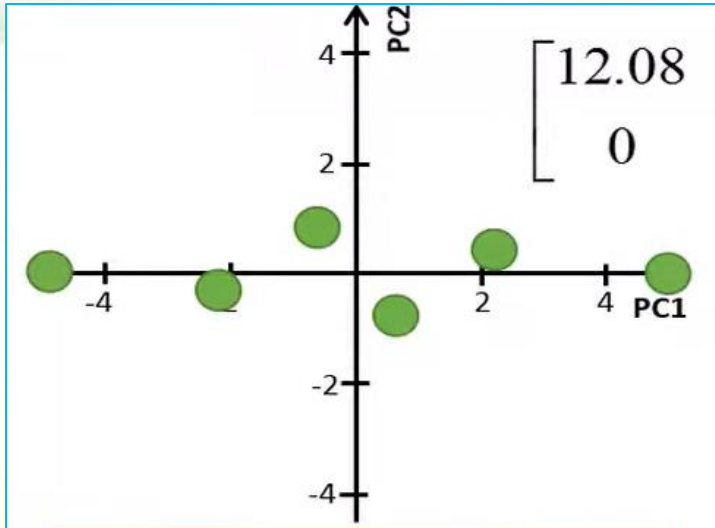
$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

PC1 & PC2 Variances = Eigenvalues

Total Variance remains same.

4.4 + 8.0 = 12.4 = 12.08 + 0.32

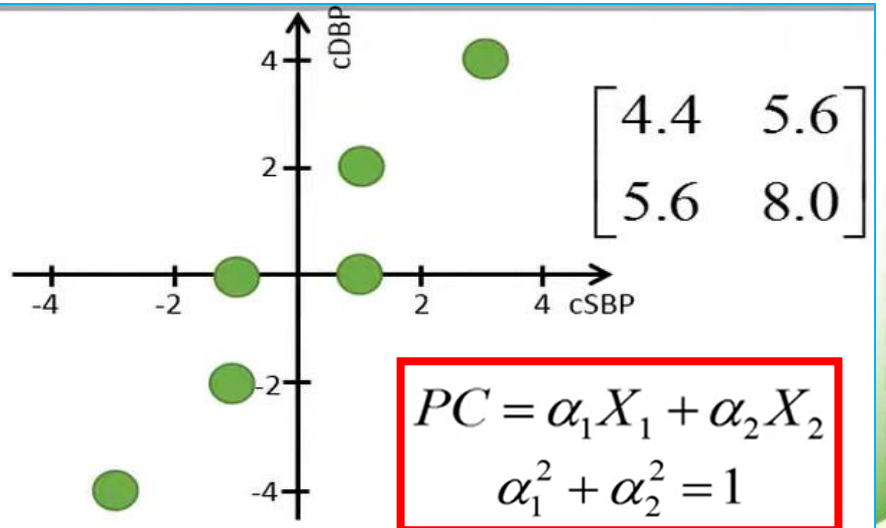
→ No loss of data → shift/transformation of variance of X-Y to PC1-PC2



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	
0.6	
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

DR - Principal Component Analysis (PCA)



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$
$$\alpha_1^2 + \alpha_2^2 = 1$$

$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

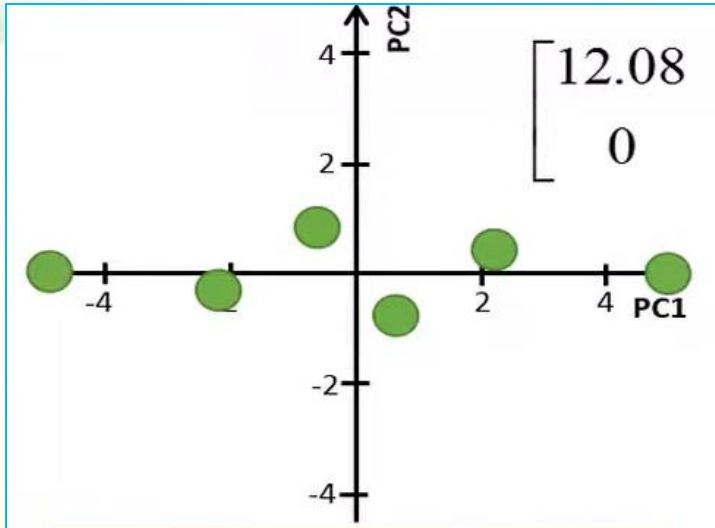
Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

$$(SBP - \overline{SBP})$$

$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

DR - Principal Component Analysis (PCA)

Apply PCA on following data.

Student	Score	Age
1	68	29
2	60	26
3	58	30
4	40	35

Steps:

1. ~~Original Data~~
2. Normalize original data (mean = 0); Center the data.
3. Calculate covariance matrix
4. Calculate Eigen values, Eigen vectors of covariance matrix
5. Normalize Eigenvectors
6. Calculate Principal Component (PC)
7. ~~Plot the graph for orthogonality between PCs~~

$$\text{var}(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$\text{cov}(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

$$|A - \lambda I| = 0$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} \qquad \begin{bmatrix} \frac{1}{3\sqrt{3}} \\ \frac{-5}{3\sqrt{3}} \\ \frac{-1}{3\sqrt{3}} \end{bmatrix}$$

$$L = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

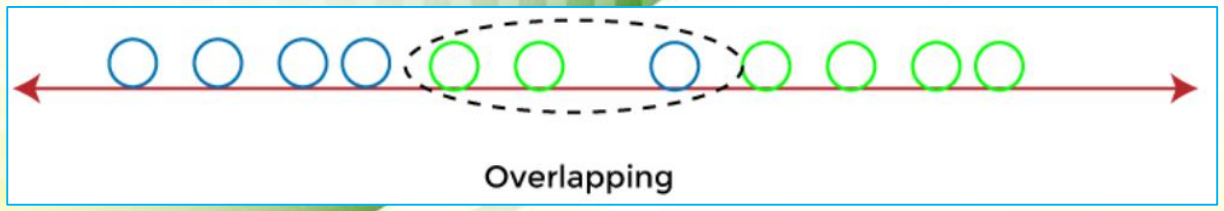
$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

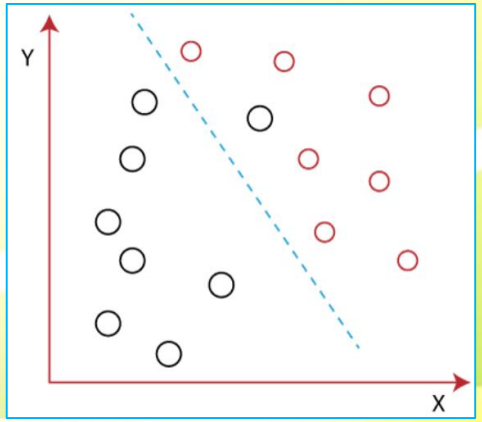
DR - Linear Discriminant Analysis (LDA)

- Also known as Normal Discriminant Analysis (NDA) or Discriminant Function Analysis (DFA).
- Linear Discriminant Analysis (LDA) is commonly used DR technique to solve multi-class classification problems.

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

$$LD = \alpha_1 X_1 + \alpha_2 X_2$$



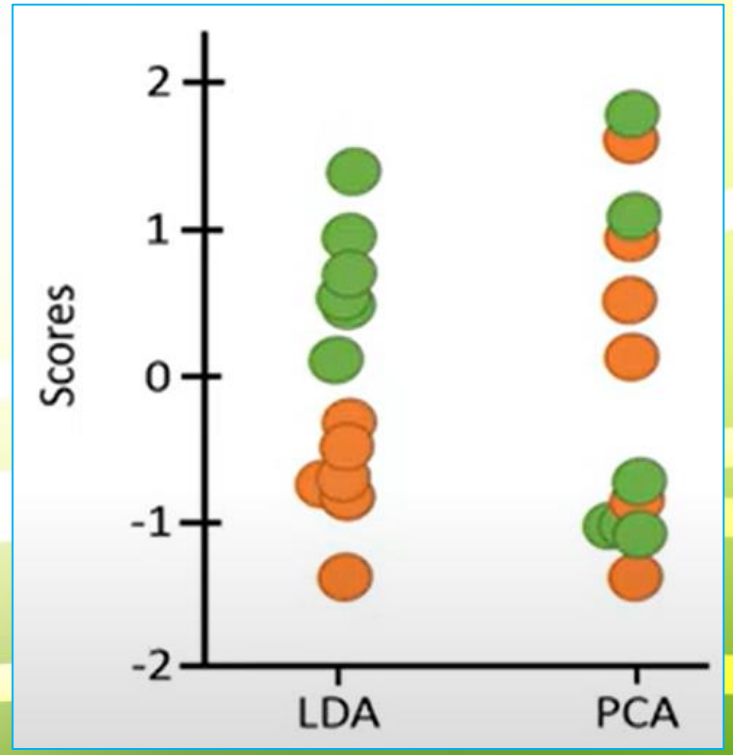


- PCA (**unsupervised** algorithm) does not care about classes/labels and only aims to find the principal components to **maximize the variance** in the given dataset.
- LDA (**supervised** algorithm) aims to find the linear discriminants to represent the axes that **maximize separation** between different classes of data.
- LDA is more suitable for multi-class classification tasks compared to PCA.
- PCA is good performer for a comparatively small sample size.

DR - Linear Discriminant Analysis (LDA)

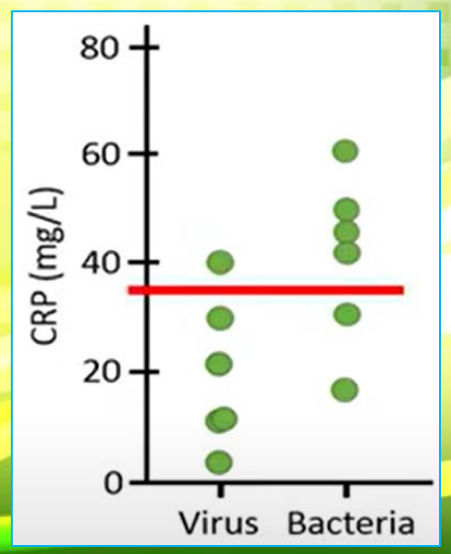
Infection	CRP (mg/L)	Temp (C)	Scores LDA	Scores PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.1	-0.9	-1.4	0.1
Viral	0.0	-1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	-0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	**1**	**1**	**0.71**	**1.29**

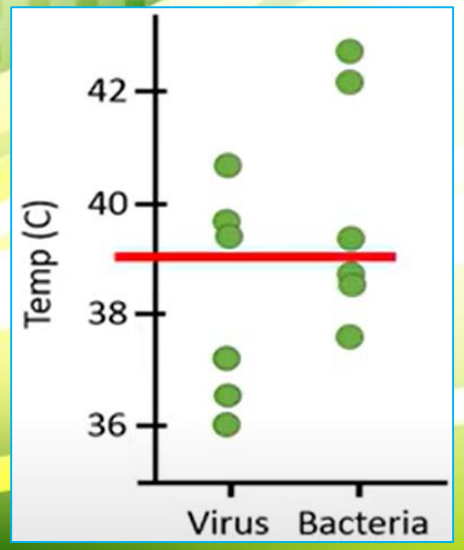
$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$
$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

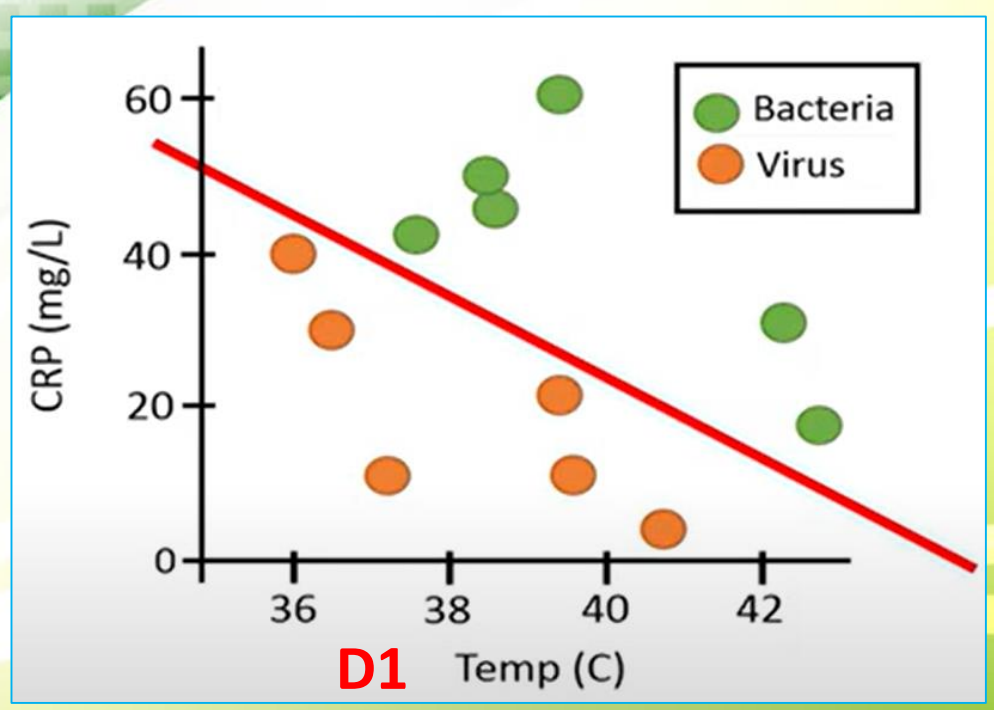


DR - Linear Discriminant Analysis (LDA)

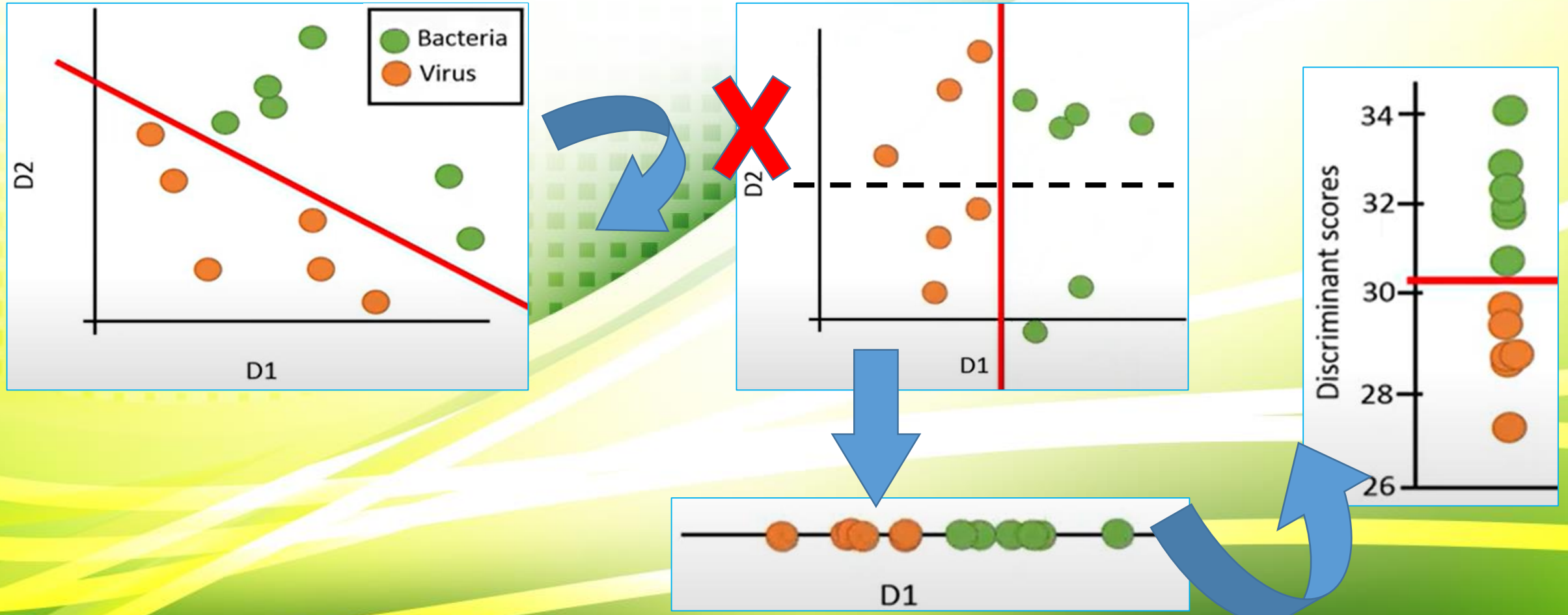
Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7







DR - Linear Discriminant Analysis (LDA)



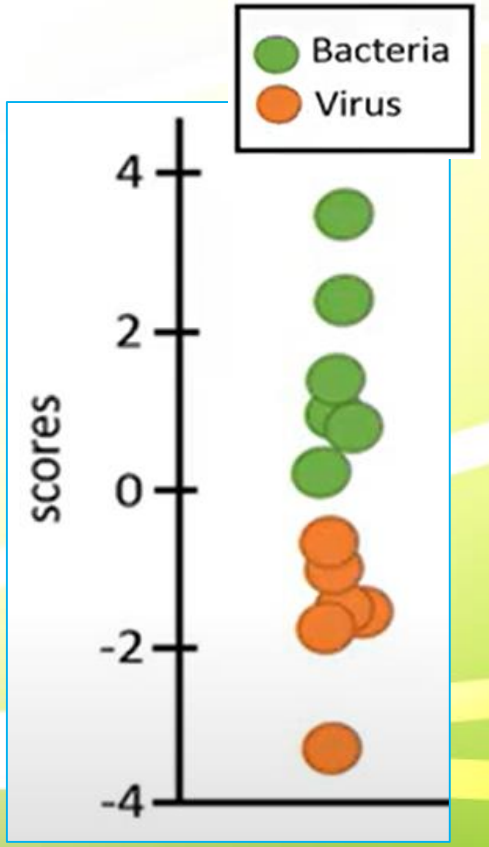
DR - Linear Discriminant Analysis (LDA)

Steps:

1. Original Data
2. Normalize the original data (mean =0); Center the data.
3. Calculate Separation matrix by finding within-group & between-group variances.
4. Calculate Eigen values, Eigen vectors of Separation matrix
5. Normalize Eigenvectors
6. Calculate Linear Discriminant (LD)
7. Plot the graph between LDs

DR - Linear Discriminant Analysis (LDA)

Infection	CRP (mg/L)	Temp (C)	Scores	Cent. scores
Viral	40.0	36.0	29.5	-1.1
Viral	11.1	37.2	27.3	-3.3
Viral	30.0	36.5	28.8	-1.8
Viral	21.4	39.4	29.9	-0.7
Viral	10.7	39.6	28.9	-1.7
Viral	3.4	40.7	28.9	-1.7
Bacterial	42.0	37.6	30.8	0.2
Bacterial	31.1	42.2	32.9	2.3
Bacterial	50.0	38.5	32.3	1.7
Bacterial	60.4	39.4	34.0	3.5
Bacterial	45.7	38.6	31.9	1.3
Bacterial	17.3	42.7	31.8	1.2



$$LD = \alpha_1 X_1 + \alpha_2 X_2$$

$$LD = 0.11 \cdot \mathrm{CRP} + 0.70 \cdot \mathrm{Temp}$$

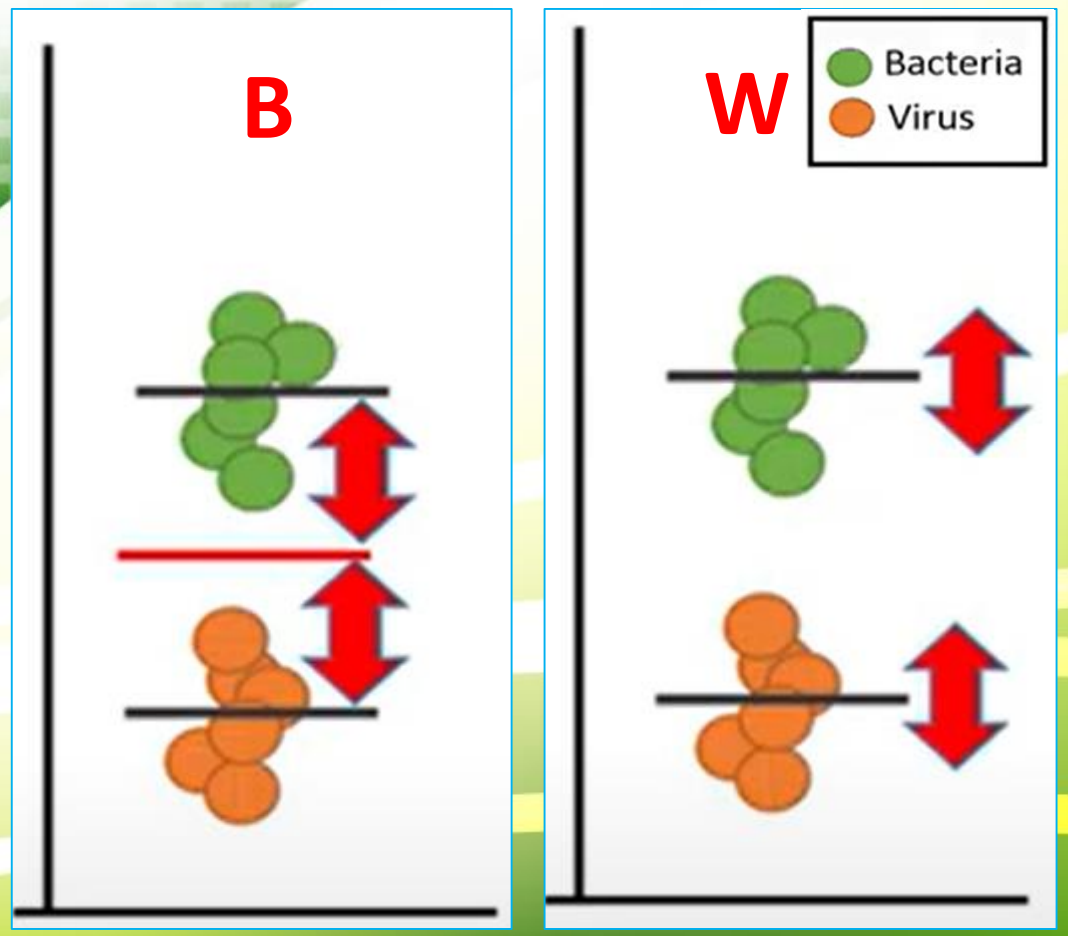
$$LD = 0.11 \cdot \left(\mathrm{CRP} - \overline{\mathrm{CRP}}\right) + 0.70 \cdot \left(\mathrm{Temp} - \overline{\mathrm{Temp}}\right)$$

DR - Linear Discriminant Analysis (LDA)

- PCA: EigenVector of covariance matrix
- LDA: EigenVector of Separation matrix

$$\text{Separation} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

$$S = W^{-1}B$$



DR - Linear Discriminant Analysis (LDA)

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

$$W = \frac{(n_1 - 1)\text{cov}(A) + (n_2 - 1)\text{cov}(B)}{n_1 + n_2 - 2}$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$\text{cov}_{\text{virus}} = \begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix}$$

$$\text{cov}_{\text{Bacteria}} = \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}$$

$$W = \frac{\begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix} + \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}}{2} = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

DR - Linear Discriminant Analysis (LDA)

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$T = B + W$$

$$B = T - W$$

$$B = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix} - \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix} = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

$$S = W^{-1}B$$

$$S = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix} \cdot \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix} = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$W^{-1} = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

DR - Linear Discriminant Analysis (LDA)

Infection	CRP (mg/L)	Temp (C)	Scores	Scores
Viral	40.0	36.0	41.6	29.5
Viral	11.1	37.2	38.4	27.3
Viral	30.0	36.5	40.6	28.8
Viral	21.4	39.4	42.2	29.9
Viral	10.7	39.6	40.8	28.9
Viral	3.4	40.7	40.8	28.9
Bacterial	42.0	37.6	43.5	30.8
Bacterial	31.1	42.2	46.4	32.9
Bacterial	50.0	38.5	45.5	32.3
Bacterial	60.4	39.4	48.0	34.0
Bacterial	45.7	38.6	45.0	31.9
Bacterial	17.3	42.7	44.8	31.8

$$LD1 = 0.11 \cdot CRP + 0.70 \cdot Temp$$

$$LD1 = \frac{0.150 \cdot CRP + 0.989 \cdot Temp}{\sqrt{1.989}}$$

$$var(scores)_{Viral} = 1.60$$

$$var(scores)_{pooled} = 1.989$$

$$var(scores)_{Bacterial} = 2.37$$

Scores
29.5
27.3
28.8
29.9
28.9
28.9
30.8
32.9
32.3
34.0
31.9
31.8

$var(scores)_{Viral} = 0.81$

$var(scores)_{pooled} = 1$

$var(scores)_{Bacterial} = 1.19$

DR - Linear Discriminant Analysis (LDA)

$$LD = 0.11 \cdot \left(\text{CRP} - \overline{\text{CRP}}\right) + 0.70 \cdot \left(\text{Temp} - \overline{\text{Temp}}\right)$$

Infection	CRP (mg/L)	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.1	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	28.9	-1.7
Viral	3.4	40.7	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	0.2
Bacterial	31.1	42.2	46.4	32.9	2.3
Bacterial	50.0	38.5	45.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.9	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

DR - Linear Discriminant Analysis (LDA)

Apply LDA on following data.

Score A	Score B	Status
2.95	6.63	Passed
2.53	7.79	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	4.46	Not Passed
2.16	6.22	Not Passed
3.27	3.52	Not Passed
2.87	4.85	Not Passed

$$\text{var}(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$\text{cov}(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\text{Separation} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

$$S = W^{-1}B$$

$$|A - \lambda I| = 0$$

$$B = T - W$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Steps:

1. ~~Original Data~~
2. Normalize original data (mean = 0); Center the data.
3. Calculate Separation matrix by finding within-group & between-group variances.
4. Calculate Eigen values, Eigen vectors of Separation matrix
5. Normalize Eigenvectors
6. Calculate Linear Discriminant (LD)
7. ~~Plot the graph between LDs~~

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} \quad \begin{bmatrix} \frac{1}{3\sqrt{3}} \\ \frac{-5}{3\sqrt{3}} \\ \frac{-1}{3\sqrt{3}} \end{bmatrix}$$

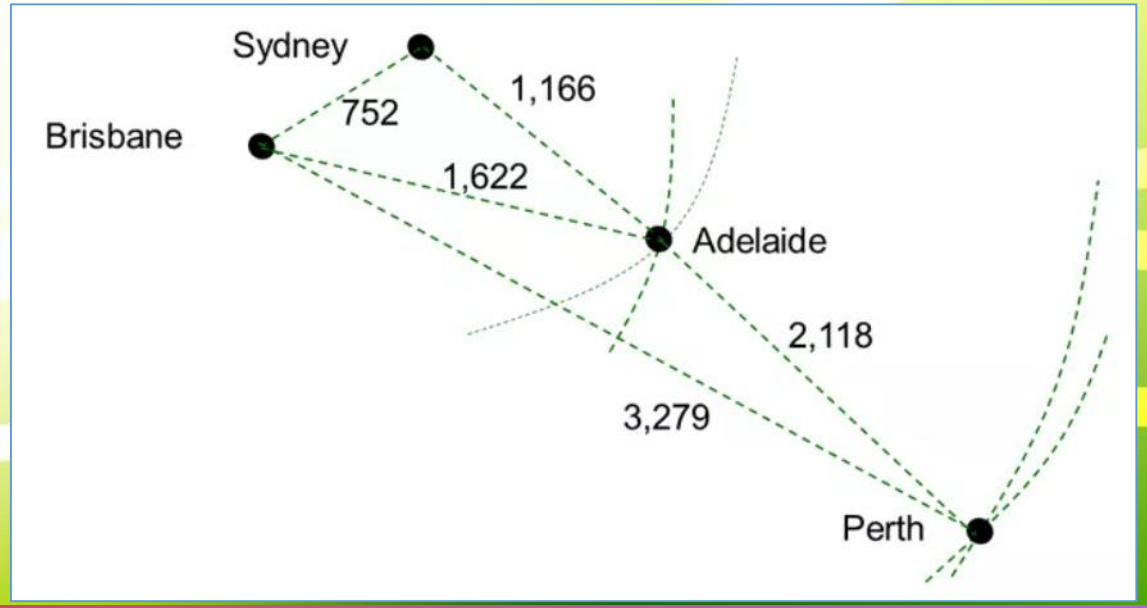
$$L = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

$$\text{LD1} = 0.11 \cdot \text{CRP} + 0.70 \cdot \text{Temp}$$

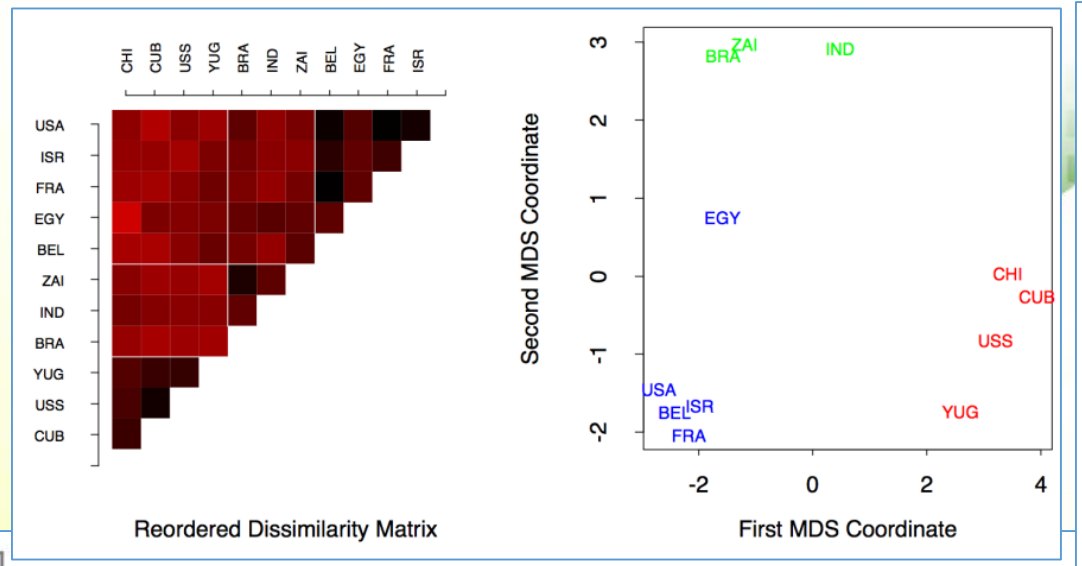
DR - Multidimensional Scaling (MDS)

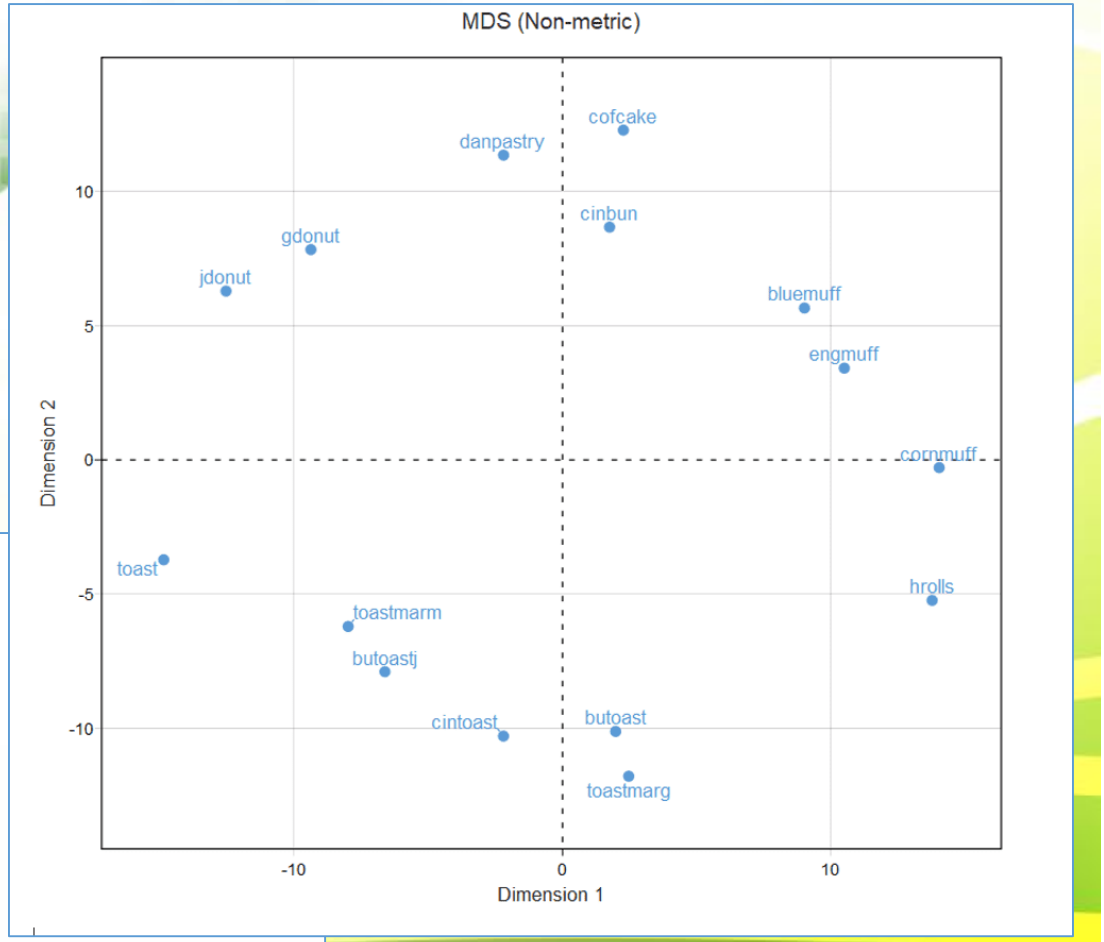
- Given pairwise dissimilarities, MDS reconstruct a map that preserves distances in lower dimension.
- MDS helps in visualizing distances between objects, where distance is known between pairs of the objects.
- Input to MDS is a *distance matrix.*
- *O*utput is typically a 2-D scatterplot, where each of the objects is represented as a point.

Adelaide	1,166		
Brisbane	752	1,622	
Perth	3,279	2,118	3,606
	Sydney	Adelaide	Brisbane



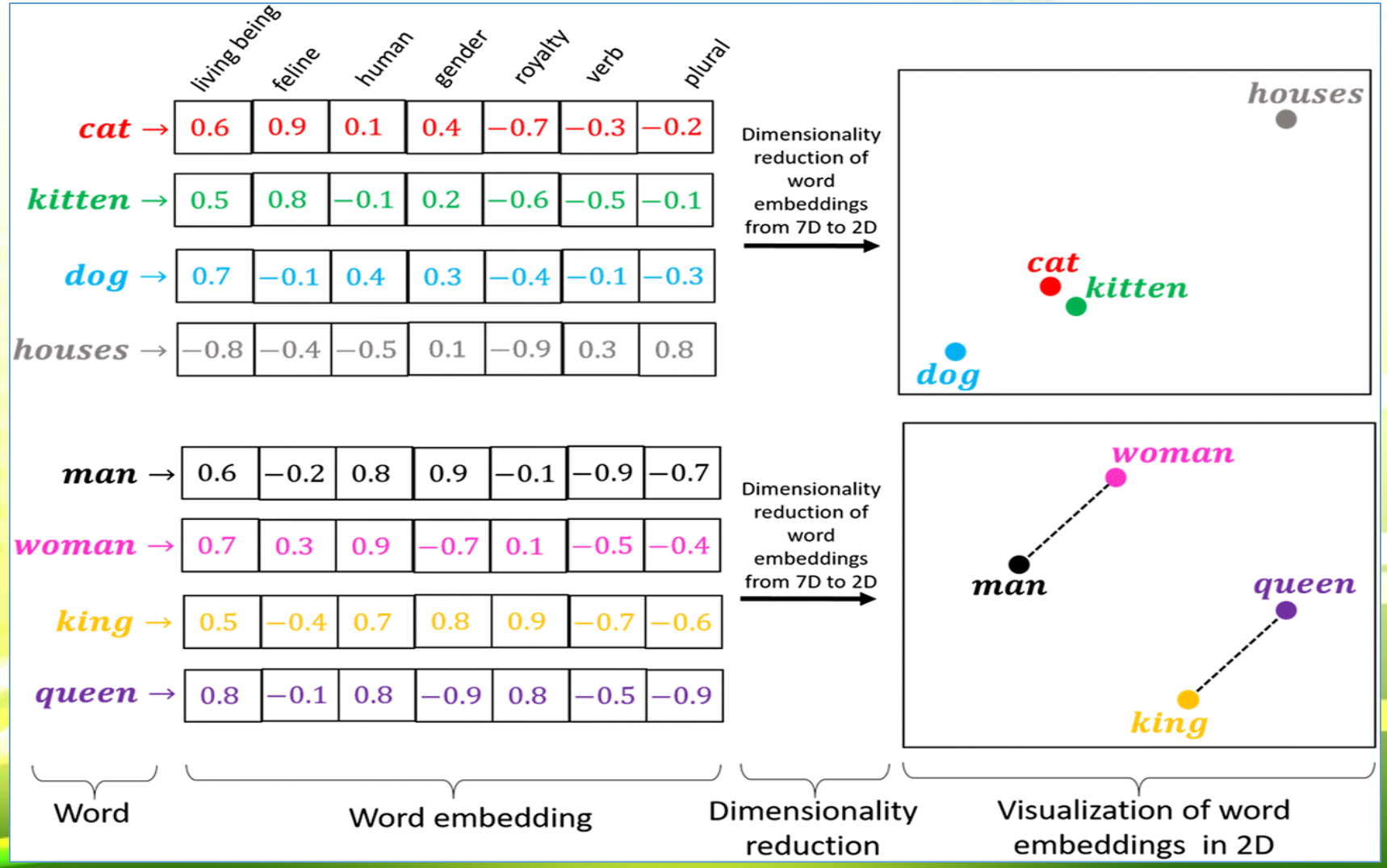
DR - Multidimensional Scaling (MDS)





	toast	butoast	engmuff	jdonut	cintoast	bluemuff	hrolls	toastmarm	butoastj	toastmarg	cinbun	danpastry	gdonut	cofcake
butoast	15													
engmuff	25	15												
jdonut	3	24	22											
cintoast	14	3	17	22										
bluemuff	24	17	2	21	19									
hrolls	28	8	4	27	18	8								
toastmarm	7	7	20	11	6	18	23							
butoastj	8	6	21	12	5	19	22	2						
toastmarg	16	2	16	25	4	18	9	8	7					
cinbun	26	17	10	17	12	7	18	20	19	18				
danpastry	21	25	11	5	19	10	22	17	16	26	2			
gdonut	20	18	24	2	23	22	25	11	12	17	4	11		
cofcake	16	22	11	13	21	7	21	21	20	23	6	7	11	
cornmuff	27	11	3	26	16	4	5	25	24	12	12	16	24	16

DR - Multidimensional Scaling (MDS)



DR : t-SNE

- PCA is simple, fast, easy to use, and retains the overall variance of the dataset.
- One major drawbacks of PCA is that it does not retain non-linear variance.
- PCA works on retaining only global (linear) variance, and t-SNE retains local (non-linear) variance.
- t-SNE is a **nonlinear dimensionality reduction** technique that is well suited for embedding high dimension data into lower dimensional data (2D or 3D) for data visualization.
- t-SNE stands for **t-distributed Stochastic Neighbor Embedding**, which tells the following :

  Stochastic → not definite but random probability

  Neighbor →concerned only about retaining the variance of neighbor points

  Embedding → plotting data into lower dimensions

- t-SNE is a machine learning algorithm that generates slightly different results each time on the same data set, focusing on retaining structure of neighbor points.

DR : t-SNE

- t-SNE works by converting high dimensional data points to joint probabilities and uses these probabilities to minimize the KL divergence so that low dimensional embeddings can be obtained.

t-SNE steps:

  - t-SNE constructs a probability distribution on pairs in higher dimensions such that similar objects are assigned a higher probability and dissimilar objects are assigned lower probability.
  - Then, t-SNE replicates the same probability distribution on lower dimensions iteratively till the Kullback-Leibler divergence is minimized.

- KL divergence is a measure of difference between probability distributions.
- KL divergence is mathematically given as the expected value of the logarithm of the difference of these probability distributions.

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot log \frac{p(x_i)}{q(x_i)}$$

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

DR : t-SNE effectiveness

- t-SNE plots are highly influenced by parameters.
- Two important parameters are;
  - **n_iter:** number of iterations that the algorithm runs
  - **perplexity**: number of neighboring points t-SNE must consider
- It is necessary to perform t-SNE using different parameter values before analyzing results.
- Since t-SNE is stochastic, each run may lead to slightly different output. This can be solved by fixing the value of random_state parameter for all the runs.
- t-SNE doesn't retain distance between clusters from raw data. Distance between clusters might vary post dimensionality reduction in t-SNE. One shouldn't obtain any conclusion solely from distance between clusters.
- t-SNE shrinks widespread data and expands densely packed data. It is hence suggested not to decide the size and density/spread/variance of the clusters based on the output.
- Lower perplexity values might result in fewer clusters. It is hence recommended to try various perplexity values ranging from 2 to the number of data points to obtain better results.

DR

- Linear DR transforms data to a reduced dimension space using a linear combination of the original variables.
- The aim is to replace original variables by a smaller set of underlying variables.
- Unsupervised linear dimensionality reduction techniques include PCA and Multidimensional Scaling (MDS).
- Linear DR techniques are unsuitable if dataset contains nonlinear relationships among variables.
- If original high-dimensional data set contains nonlinear relationships, then nonlinear dimensionality reduction techniques may be more appropriate.
- Some methods such as LLE (Locally Linear Embedding) and Isomap rely on applying linear techniques on a set of local neighborhoods, which are assumed to be locally linear in nature. As such, they fall into the category of local linear dimensionality reduction techniques.

DR : isomap

- Isomap (Isometric Feature Mapping), unlike Principle Component Analysis (PCA), is a non-linear feature reduction method.
- Isomap should be used when there is a non-linear mapping between higher-dimensional data and lower-dimensional manifold.
- Isomap is better than linear methods when dealing with almost all types of real image and motion tracking.
- Isomap uses MDS techniques with geodesic interpoint distances instead of Euclidean distances.
- Geodesic distances represent the shortest paths along the curved surface.
- Unlike the linear techniques, Isomap can discover the nonlinear degrees of freedom that underlie complex natural observations.

DR – Subset Selection

- Subset selection is a classic topic of model selection in statistical learning and is encountered whenever we are interested in understanding the relationship between a response and a set of explanatory variables.
- Feature Selection intends to select a subset of attributes or features that makes the most meaningful contribution to a machine learning activity.
- Objectives of subset selection are:
  - Having a faster and more cost-effective (less need for computational resources) learning model
  - Having a better understanding of the underlying model that generates the data.
  - Improving the efficacy of the learning model.
- Main Factors Affecting Subset Selection:
  - Feature Relevance (information theory; mutual information, etc.)
  - Feature Redundancy (distance; correlation, etc.)

End !!