DATA REDUCTION

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Generally, large data sets will make data mining analysis less efficient.
- Data scientists may also be easily confused in conducting analysis.

- To overcome this problem, data reduction techniques can be used to reduce the dimensions/numerosity of a data set.
- However, it still provides information that is almost the same as the original data.

- Two approaches in data reduction:

i) Dimensional Data Reduction

ii) Numerosity Data Reduction

DIMENSIONAL DATA REDUCTION:

- The large dimension in the data makes the efficiency of algorithm in mining methods less efficient (curse of dimensionality).

- In fact, the size of the data storage may not be sufficient to store an excessive amount of data.

- Data with large dimensions can be reduced through the following methods:

i) Removing Attributes

ii) Principal Component Analysis

iii) Factor Analysis

REMOVING ATTRIBUTES:

- Removing certain attributes is the simplest method of reducing data dimensions.
- This can be done by removing attributes that have the following characteristics:

i) Attributes that provide almost the similar information:

- If we found that there are attributes that are duplicates of other attributes, then the same information can be obtained among those attributes.
- Example: the price of a product and the amount of sales tax.

ii) Irrelevant attributes:

- An attribute only provides useful information if it is needed to achieve the objectives of the analysis.
- Example: the student ID attribute is irrelevant for analyzing student performance.

REMOVING ATTRIBUTES:

iii) Insignificant attributes:

- An attribute that is found to be insignificant can be removed from the data

Example:

- Insignificant attributes detected through regression model analysis.
- That is, when the relationship of the response variable *Y* and the regressor variable X_j can be described linearly as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

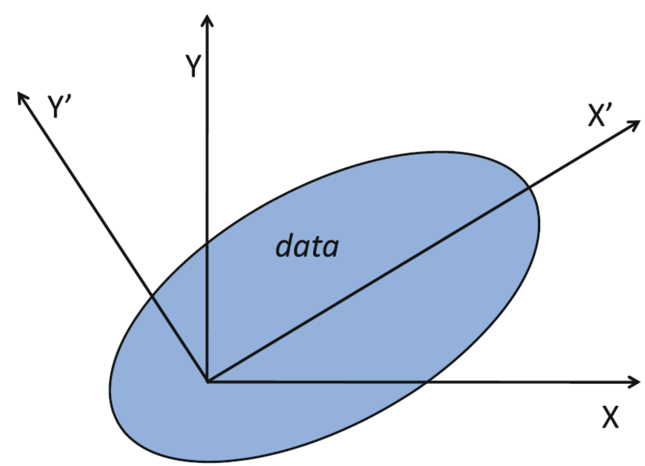
- We are interested in testing either $H_0:\beta_j=0$ vs $H_a:\beta_j \neq 0$.
- The attribute X_j for which the parameter $\beta_j=0$ is found to be significant, can be removed from the data.

PRINCIPAL COMPONENT ANALYSIS (PCA):

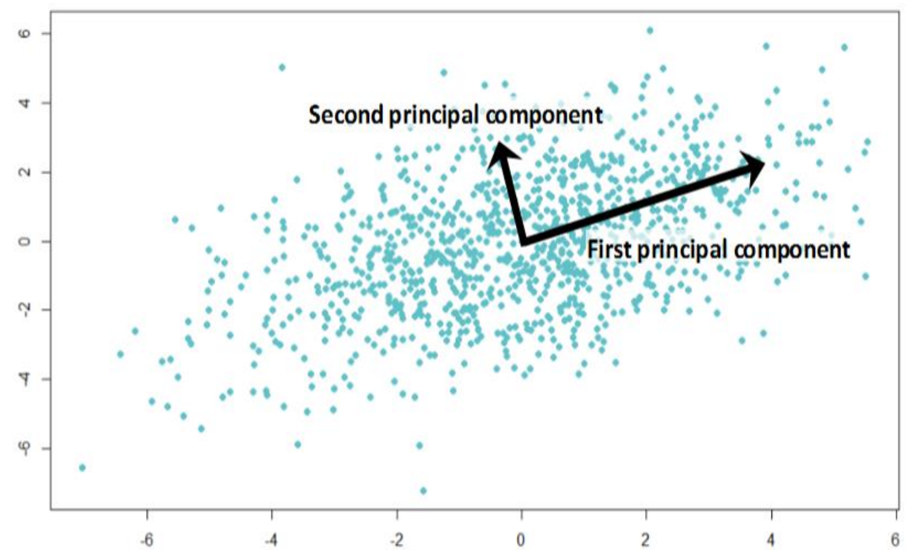
- The basic idea of PCA is to obtain a set of linear transformations with a smaller number of variables than the original set of variables that can represent most of the variance of the original data.
- That is, a set of orthogonal vectors *k* that can represent original data, with $k \leq p$ (p is the original data dimension).
- This new set of attributes is represented in a smaller order of variation contributions.
- The first variable of the PCA, is called the first principal component. Its contains the largest variance against the original data set.
- The second variable of PCA, is called the second principal component contains the second largest variance against the original data set.
- The third variable, and so on.

PRINCIPAL COMPONENT ANALYSIS (PCA):

- A general procedure is to determine several principal components that able to retain 80% or more of the variance of the original data set.
- PCA is useful when the original data contain too many attributes and the correlation between some attributes is quite high (correlated/related).
- Information related to the principal components are represented through the eigenvalues and eigenvectors which can obtained from the correlation matrix.



.1 PCA. X' and Y' are the first two principal components obtained



PCA PROSEDURES:

i) Scale the input data by standardizing the range for each attribute involved (*z*-score).

ii) Determine *k*-set of orthonormal vectors based on the standardized data.

iii) The main principal components are arranged in descending contribution based on eigenvalue information. The main component serves as a new set of axes according to the largest percentage of variance.

iv) Data dimension reduction was carried out by removing components that contributed a low variance.

- However, PCA analysis can only be performed if all attributes are numerical.

PCA PROSEDURES:

- Given a random vector **X= (X$_1$,X$_2$, …, X$_p$)**.
- Compute correlation/covariance matrix:

$$\mathrm{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

- PCA can be obtained through the following linear relationship:

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

- Where **Y= (Y$_1$,Y$_2$, …, Y$_p$)** is a new set of PCA variables.
- **e$_i$** is the set of eigenvectors for the covariance matrix (or correlation matrix).
- The first principal component (**Y$_1$**) stores the largest variance of the original data, followed by **Y$_2$** and so on. This represent by eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

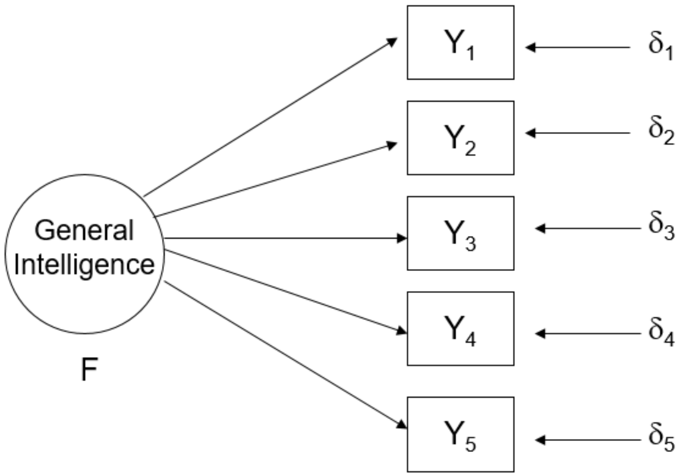
FACTOR ANALYSIS (FA):

- FA is a data reduction technique used to explain the covariance between observed variables in the form of unobserved variables (latent) with smaller dimensions.

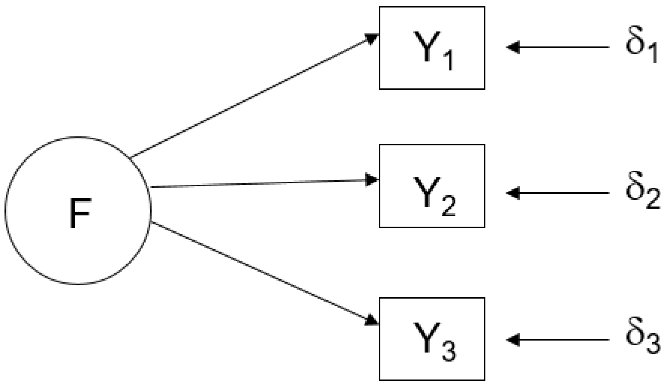
- FA objective is to find hidden factors in the original data change.

- In FA, we assume that there is a set of latent (unobserved) factors F_j , $j = 1, \ldots, k$; that can be derived from the original data.

- FA characterizes the dependency between the attributes of the original data through smaller-dimensional factors.



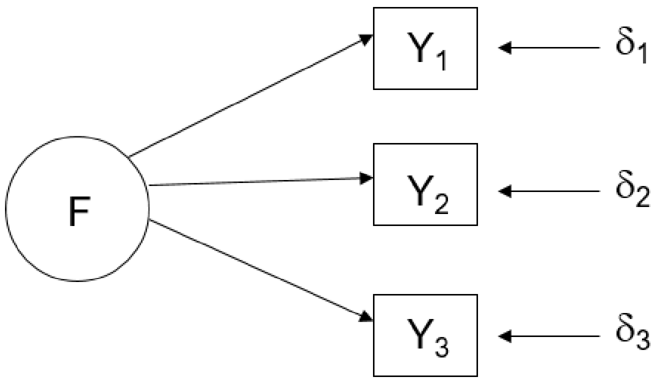
EXAMPLE: ONE-FACTOR MODEL



$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

- Y_1, Y_2, and Y_3 are the variables for the original data.
- F is an unobserved latent factor.
- δ_i is an error representing a variation in Y_i that cannot be explained by a factor of F.
- Y_i *can be* described through a linear relationship of factor F and error δ_i.

ASSUMPTIONS IN ONE-FACTOR MODEL:



$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

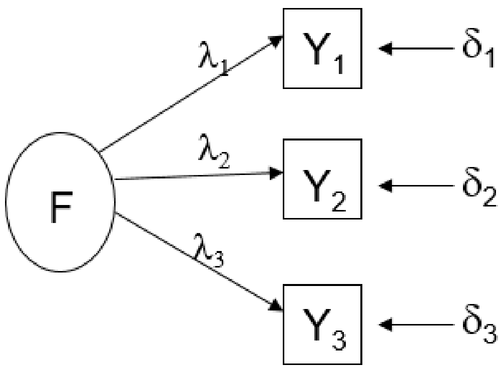
- F is the latent factor for Y_1, Y_2, Y_3.
- F is independent of δ_j, i.e $cov(F, \delta_j)=0$
- δ_i and δ_j is independent for $i \neq j$, i.e $cov(\delta_i, \delta_j)=0$
- Conditional independent: The variables Y_i and Y_j are independent of each other, given the factor F, i.e $cov(Y_i, Y_j \mid F) = 0$.

ASSUMPTIONS IN ONE-FACTOR MODEL:

- Given Y_1, Y_2, Y_3 which has been standardized: $var(Y_i) = var(F) = 1$

- Factor loadings:

  $\lambda_i = corr(Y_i, F)$



- Communality for variable Y_i:

  $h_i^2 = \lambda_i^2 = [corr(Y_i, F)]^2$

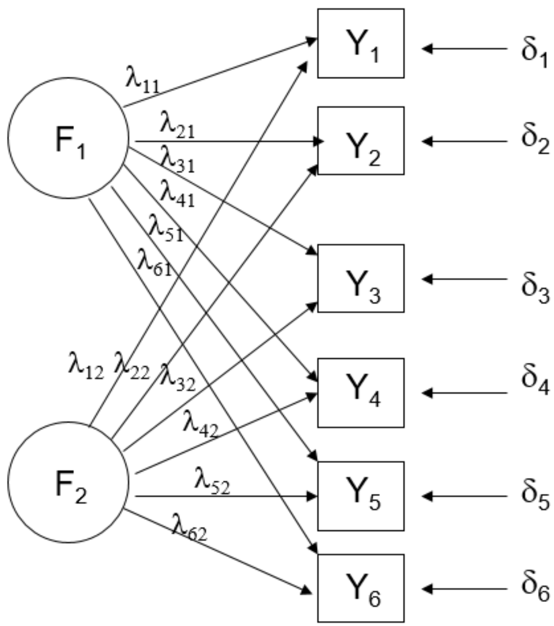
  =% of variance for Y_i which explained by the factor F.

$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

- Uniqueness for Y_i:

  $1 - h_i^2$ = residual variance for Y_i.

EXAMPLE: TWO-FACTOR MODEL



$$Y_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \delta_1$$
$$Y_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \delta_2$$
$$Y_3 = \lambda_{31}F_1 + \lambda_{32}F_2 + \delta_3$$
$$Y_4 = \lambda_{41}F_1 + \lambda_{42}F_2 + \delta_4$$
$$Y_5 = \lambda_{51}F_1 + \lambda_{52}F_2 + \delta_5$$
$$Y_6 = \lambda_{61}F_1 + \lambda_{62}F_2 + \delta_6$$

- Factors F_1 and F_2 are common factors because these two factors share more than two variables from the same Y_1, Y_2, Y_3 , Y_4, Y_5, Y_6 in each factor.
- Models with *m*-Factors and n-variables will lead to a more complex model specifications.

FACTOR ROTATION:

- Factor rotation aims to obtain a simpler structure and make the factors easier to be interpreted.
- Factor rotation does not affect the fitting of the factor model.
- The number of factors and communalities for Y obtained from model fitting were unchanged after the factor rotation was carried out.

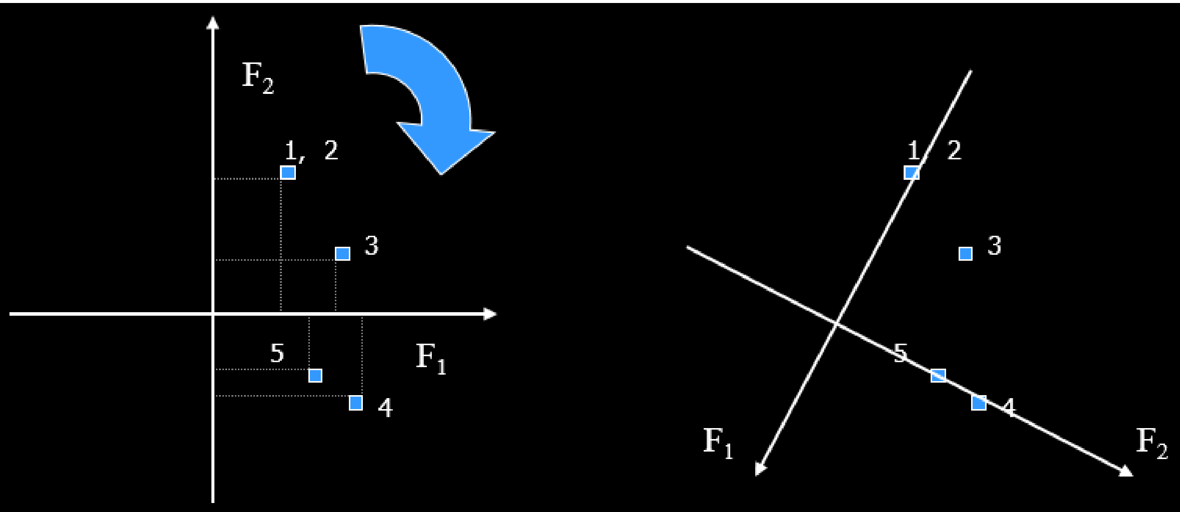
- Factor rotation is made by redefining the factors obtained in such a way such that:

i) The values matrix loadings for some factors tend to increase approaching the value of -1 or 1.

ii) The value of the weighting matrix for some of the other factors tends to decrease to a value close to 0.

- This process makes the influence of each factor on the original variable more significant and easier to be interpreted.

EXAMPLE OF FACTOR ROTATION:



	Faktor 1	Faktor 2
x1	0.4	0.69
x2	0.4	0.69
x3	0.65	0.32
x4	0.69	-0.4
x5	0.61	-0.35

	Faktor 1	Faktor 2
x1	-0.8	0
x2	-0.8	0
x3	-0.6	0.4
x4	0	0.8
x5	0	0.7

NUMEROSITY DATA REDUCTION:

- Numerosity Data Reduction can be made by substituting the original data into another alternative form:

i. Parametric Model:

Example:

- Regresion model
- Log-linear model
- Probability distribution, and etc.

ii. Non-Parametric Model:

Example:

- Histogram
- Resampling techniques (bootstraps, jackknife method).
- Clustering, and etc.

PARAMETRIC MODEL:

- The best statistical model obtained from the data fit will be used as a representation of the data.
- Only the parameters of a model will be used to represent the data.
- The simulation data generated from the model will resemble the actual data.
- Example:

i) Linear Regression Model:

- The variable Y is numerical and must obey the assumption of Normal distribution

ii)Log-linear Model:

- Y is discrete and multi-dimension.

iii) Probability Distribution Model:

- Univariate model: Normal, Poisson, Weibull, Gamma, Pareto, and etc.
- Multivariat Model: Joint distribution and Copula.

ii) And various other Statistical models.

NON-PARAMETRIC MODEL:

i) Histogram/Discretization:

- Data will be allocated to a specific intervals.
- Data will be stored in the form of interval data measurement, i.e.; average, median, mod and etc.

ii) Clustering:

- Data will be partitioned into several sets of clusters based on similarity features that exist between the data.
- Data in the same cluster had similar characteristics with small variations.
- Data between different clusters did not have the same characteristics and having a large variation..

iii) Resampling techniques (bootstraps, jackknife method):

- Some portion samples will be taken randomly from the full set of original data.
- These samples will be used to represent the original data set

TYPES OF SAMPLING:

i)Simple Random Sampling: Each item in the data set has the same probability of being selected.

ii) Sampling without replacement: Once a data item is selected, it will be removed from the original data set.

iii) Sampling with replacement: Selected data items, re -entered in the original data set. It is possible to be re-elected.

iv) Stratified sampling: Data is partitioned into specific groups (strata) based on the nature of the data. Then, a random sample will be taken from each strata.

- Other sampling includes Clustered Sampling, Systematic Sampling, Multistage Sampling and etc.

NEXT TOPIC:

Mining Association Rules