# Data Reduction

## Week 5

Nurul Afiqah Burhanuddin
nurul.afiqah@ukm.edu.my

Room 2119
Department of Mathematical Science

# Outline

# Introduction

- Data mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.
- Mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.
- Two approaches in data reduction:
  - Numerosity reduction: replace the original data volume by alternative, smaller forms of data representation.
    - Non-parametric reduction
    - Parametric reduction
  - Dimensionality reduction: the process of reducing the number of variables/attributes under consideration.
    - Variable selection
    - Principle component analysis
    - Factor analysis

# Numerosity Reduction

## Non-parametric reduction

Do not hold any assumption of the data fit into a model. Unlike the parametric method, these methods may not give a huge reduction in data, though they generate reduced data in a systematic way.

- sampling
- data aggregation

# Numerosity Reduction

## Non-parametric reduction

**Sampling**

Allows a large data set to be represented by a much smaller random data sample (or subset).

- Simple random sample with replacement:
  - This is created by drawing $n'$ of the $n$ observation from the data, where the probability of drawing any observation is $1/n$, that is, all observations are equally likely to be sampled.
  - Each time an observation is drawn, it may be drawn again.
  - The items in the sample are independent because the outcome of one random draw is not affected by the previous draw.
  - Eg: Mark and recapture method to estimate an animal population. The captured animals will be marked and released.

# Numerosity Reduction

Non-parametric reduction

**Sampling**

- Simple random sample without replacement:
  - This is similar to simple random sample with replacement, except that each time an observation is drawn from it could not be selected again.
  - The items in the sample are dependent because the outcome of one random draw is affected by the previous draw.
  - Eg: we want to estimate the household income in a population, we might want to collect a random sample of 10,000 households but we don't want any duplication on the sampled households.

# Numerosity Reduction

Non-parametric reduction

## Sampling

- Stratified sampling:
  - Strata is constructed such that every member of the population fits into one, and only one, stratum. These strata must collectively contain all members of the population.
  - It is usually used when we want to examine subgroups (strata) within a population.
  - By explicitly incorporating the strata into the sampling, ensuring that the sample represents all groups. When you have smaller groups in your study, simple random sampling can miss some of them by chance. Stratified sampling helps retain the complete variety of the population in the sample.
  - Eg: We divide a sample of adults into subgroups by age, like 18–29, 30–39, 40–49, 50–59, and 60 and above. To stratify this sample, we would then randomly select proportional amounts of people from each age group.

# Numerosity Reduction

## Non-parametric reduction

**Data aggregation**

Allows large datasets be represented by fewer data points, typically in the form of summary statistics.

- Enable us to access and examine large amounts of data in a reasonable time frame. A row of aggregate data can represent hundreds, thousands or even more data points. When the data is aggregated, it can be queried quickly instead of requiring all of the processing cycles to access each underlying data row and aggregate it in real time when it is queried.
- Before aggregating, it is crucial that the data is analyzed for accuracy and that there is enough data for the aggregation to be useful.
- Aggregate data does not need to be numeric. You can, for e.g., count the number of any non-numeric data element.

# Numerosity Reduction

Non-parametric reduction

**Data aggregation**

- Examples of data aggregation:
  - Voter turnout by state or county. Individual voter records are not presented, just the vote totals by candidates for the specific region.
  - Average age of customer by product. Each individual customer is not identified, but for each product, the average age of the customer is saved.
  - Number of customers by country. Instead of examining each customer, a count of the customers in each country is presented.

# Numerosity Reduction

## Parametric reduction

Hold an assumption that the data fit into a model. Hence, it estimates the model parameters and stores only these estimated parameters and not the original or the actual data.

- linear model
- probability distribution

# Dimensionality Reduction

## Variable selection

- Reduces the data set size by removing irrelevant or redundant variables (or dimensions).
- The goal of variable selection is to find a minimum set of variables so that the analytical results of the reduced set of variable are as close as possible to the analytical results obtained using all variables.
- Mining on a reduced set of variables has an additional benefit: It reduces the number of variables appearing in the discovered patterns, helping to make the patterns easier to understand.

# Dimensionality Reduction

## Variable selection

- Three strategies for variable selection:
  - **Forward selection:** The procedure starts with an empty set of variables as the reduced set. The best of the original variables is determined and added to the reduced set. At each subsequent step, the best original variables are added to the set.
  - **Backward elimination:** The procedure starts with the full set of variables. At each step, it removes the worst from among the remaining variables.
  - **Combination of forward selection and backward elimination:** The procedure starts with an empty set of variables. At each step, the procedure selects the best variable and removes the worst from among the remaining variables.

    The strategies may employ a threshold on some measures to determine when to stop the selection process. The stopping criteria may vary from method to method. The most commonly used measures are AIC, BIC, and deviance.

# Dimensionality Reduction

## Principle Component Analysis

Given a data matrix $\mathbf{X}$ with $n$ observations and $p$ variables, where $\mathbf{X}$ are first centered on the means of each variable - to ensure that the data is centered on the origin of the principal components. The first principal components $\mathbf{Y}_{.1}$ is given by the linear combination of the variables $\mathbf{X}_{.1}, \mathbf{X}_{.2}, \ldots, \mathbf{X}_{.p}$,

$$\mathbf{Y}_{.1} = v_{11}\mathbf{X}_{.1} + v_{12}\mathbf{X}_{.2} + \cdots + v_{1p}\mathbf{X}_{.p}$$

or equivalently

$$\mathbf{Y}_{.1} = \mathbf{X}\mathbf{v}_1$$

The first principal component is calculated such that it accounts for the greatest possible variance in the data set subject to

$$v_{11}^2 + v_{12}^2 + \cdots + v_{1p}^2 = 1$$

# Dimensionality Reduction

## Principle Component Analysis

The second principal component is calculated in the same way, with the condition that it is uncorrelated with the first principal component and that it accounts for the next highest variance.

$$\mathbf{Y}_{.2} = v_{21}\mathbf{X}_{.1} + v_{22}\mathbf{X}_{.2} + \cdots + v_{2p}\mathbf{X}_{.p}$$

This continues until we have $p$ principal components. At this point, the sum of the variances of all of the principal components will equal the sum of the variances of all of the variables, that is, all of the original information has been explained. Collectively, we can write

$$\mathbf{Y} = \mathbf{XV}$$

# Dimensionality Reduction

## Principle Component Analysis

- The fundamental assumption of PCA: the transformed variables should be as uncorrelated as possible - Cov($\mathbf{Y}$) should be as close to zero as possible.
- We choose the matrix $\mathbf{V}$ in such a way that Cov($\mathbf{Y}$) is diagonal. To achieve this, the column of the $\mathbf{V}$ is set as the eigenvector of the Cov($\mathbf{X}$).
- The eigenvectors are sorted by the eigenvalues in descending order to provide a ranking of the components.

# Dimensionality Reduction

## Factor Analysis

- Investigate whether a set of standardized variables of interest $X_1, X_2, \ldots, X_p$ are linearly related to a smaller set of unobservable factors $F_1, F_2, \ldots, F_k$,

$$
\begin{aligned}
X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + e_1 \\
X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2k}F_k + e_2 \\
&\cdots \\
X_p &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pk}F_k + e_p
\end{aligned}
$$

  Equivalently, $\mathbf{X} = \mathbf{\Lambda F} + \mathbf{e}$.
- We want $k << p$.
- The error terms **e** serve to indicate that the hypothesized relationships are not exact.
- $\mathbf{\Lambda}$ is referred to as loading matrix. For e.g., $\lambda_{12}$ is called the loading of variable $X_1$ on factor $F_2$.

# Dimensionality Reduction

## Factor Analysis

- Factor analysis is a model for the covariance matrix of $\mathbf{X}$.

$$\hat{\mathbf{\Sigma}} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$$

where $\mathbf{\Psi}$ =diag$(\psi_1, \ldots, \psi_p)$ is called uniqueness, satisfying the following assumptions:

  - The common factors: $\mathrm{E}(F_j) = 0$ and $\mathrm{var}(F_j) = 1$ for $j = 1, \ldots, k$.
  - The specific errors: $\mathrm{E}(e_j) = 0$ and $\mathrm{var}(e_j) = \psi_j$ for $j = 1, \ldots, p$.
  - The common factors are uncorrelated with one another: $\mathrm{cov}(F_i, F_j) = 0$ for $i \neq j$.
  - The specific errors are uncorrelated with one another: $\mathrm{cov}(e_i, e_j) = 0$ for $i \neq j$.
  - The specific errors are uncorrelated with the common factors: $\mathrm{cov}(e_i, F_j) = 0$ for $i = 1, \ldots, p$, $j = 1, \ldots, k$.

The fit is done by optimizing the log likelihood assuming multivariate normality over the uniquenesses. However, the loadings are not unique, that is, there exists another set of values of the $\lambda_{ij}$s yielding the same theoretical covariance matrix.

# Dimensionality Reduction

## Factor Analysis

- When the first factor solution does not reveal the loadings, it is customary to apply rotation in an effort to find another set of loadings that fit the data equally well but can be more interpretable.
- As it impossible to examine all possible rotations, we can use algorithms to find rotations satisfying certain criteria:
  - varimax criterion: maximize the variance of the squared loadings for each factor; the goal is to make some loadings as large as possible and the rest as small as possible in absolute value. Hence, it encourages the detection of factors related to only a few variables and discourages the factors influencing all variables.
  - quartimax criterion: maximize the variance of the squared loadings for each variable. It tends to produce factors with high loadings for all variables

# Thank you!