Data Science with R

Detecting outlier Multivariate and Reduction Feature - Preprocessing Data

Statistics – ITS

Regita Putri Permata and Fitria Nur Aida

(Master Student at Statistics Department – ITS)

Meet 3: Preprocessing data with R Part 2

- **Outline**
    - Detecting Outlier Multivariate
    - Feature Selection
    - Feature Reduction using PCA

Detecting outlier MUltivariate

- Multivariate outlier detection is the important task of statistical analysis of multivariate data
- Multivariate Analysis based on Normal Multivariate distribution
- Methods for Detection Multivariate Outlier
    - **Mahalanobis Distance**
    - Cook's Distance
    - Leverage Point

Mahalanobis Square Distance

- A classical Approach for detecting outliers is to compute the Mahalanobis Distance (MDi) for each observation xi :

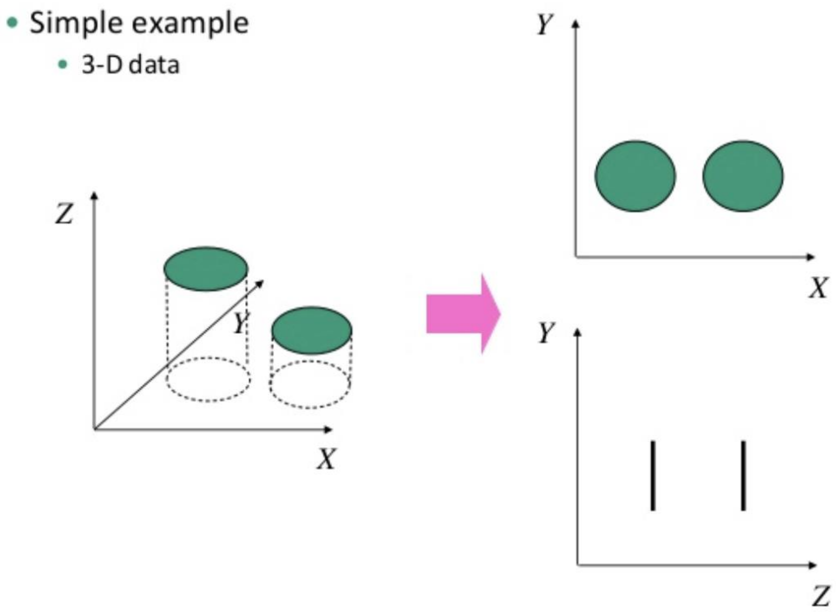
$$d_j^2 = \left(\mathbf{X_j} - \overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X_j} - \overline{\mathbf{X}}\right)$$

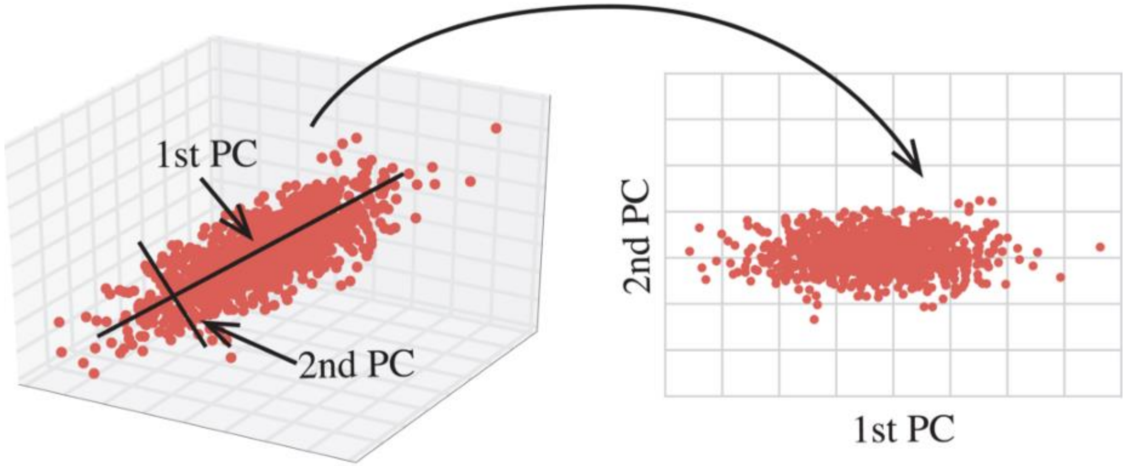
- Mahalanobis Distance is compared by fifth percentile of Chi square distribution. A more extreme percentile must serve to determine observation that do not fit the pattern of the remaining data

Data Reduction

Data Reduction

- **Dimension reduction** is the process of reducing the number of variables (also sometimes referred to as features or of course dimensions) to a set of values of variables called principal variables.
- Including insignificant variables can significantly impact your **model performance**.





in the first image, it is three dimensional data with **X,Y, Z axes**. The second image is a two dimensional space with **PC1, PC2 as axes.**

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction
  - Recursive Feature Elimination

Dimentionality Reduction

- Feature Extraction
  - Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features).
  - These new reduced set of features should then be able to summarize most of the information contained in the original set of features.
- Method
  - Principal Component Analysis
  - Multidimensional Scaling
  - Independent Component Analysis
  - t-distributed stochastic neighbor embedding (t-SNE)

Feature Selection in R

- The **stepwise regression** (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.
- There are three strategies of stepwise regression (James et al. 2014,P. Bruce and Bruce (2017)):
- **Forward selection**, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
- **Backward selection** (or **backward elimination**), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
- **Stepwise selection** (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

Principle Component Analysis in R

- Merupakan metode *interdependence* (tidak ada istilah x dan y) yang digunakan untuk mereduksi dimensi variabel
- Salah satu cara *mengatasi adanya multikolinieritas* (korelasi yang cukup tinggi antar variabel prediktor)
- Lebih banyak digunakan bersamaan dengan metode lain seperti Regresi Berganda
- Data yang digunakan berskala *interval* dan *rasio*
- Hasil PC tidak aling berkorelasi dan mempunyai varians sebesar mungkin

Tahapan PCA :

1. Standardisasi Data, karena akan digunakan matriks kovarian data yang terstandardisasi.
2. Menghitung *eigen value* dan *eigen vektor.*

```
> set.seed(101)
> pca <- prcomp(xnum,center=TRUE,scale.=TRUE)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.7544 1.9315 1.26956 1.12811 1.08124 0.92224 0.85241 0.68897 0.56441 0.56103
Proportion of Variance 0.3993 0.1963 0.08483 0.06698 0.06153 0.04476 0.03824 0.02498 0.01677 0.01657
Cumulative Proportion  0.3993 0.5957 0.68048 0.74746 0.80899 0.85375 0.89199 0.91698 0.93374 0.95031
                          PC11    PC12    PC13   PC14    PC15    PC16    PC17    PC18    PC19
Standard deviation     0.50727 0.45244 0.40492 0.3595 0.27046 0.25230 0.18201 0.11438 0.07664
Proportion of Variance 0.01354 0.01077 0.00863 0.0068 0.00385 0.00335 0.00174 0.00069 0.00031
Cumulative Proportion  0.96385 0.97463 0.98325 0.9901 0.99391 0.99726 0.99900 0.99969 1.00000
```

2. Penentuan PC yang diambil

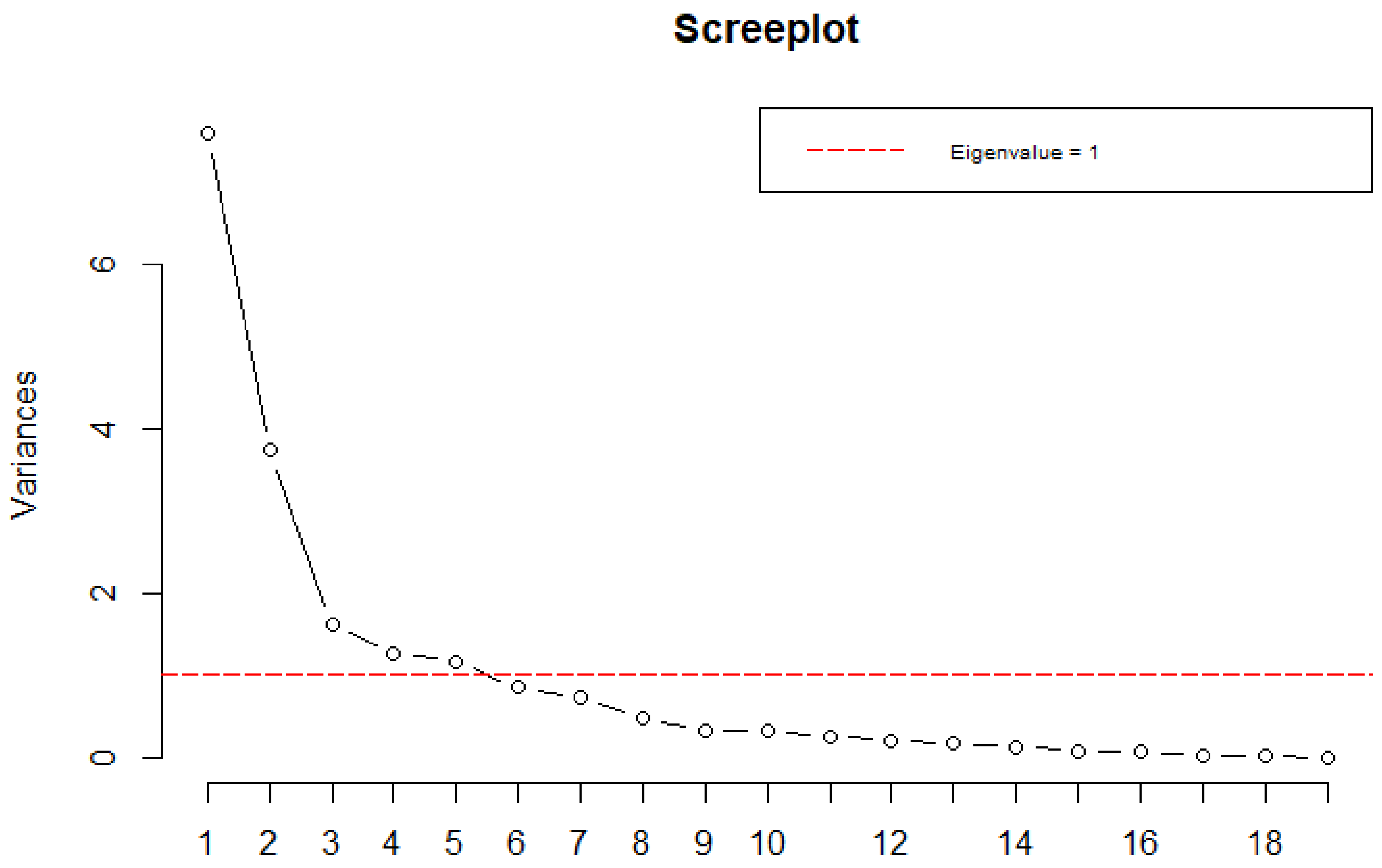
Kriteria PC yang diambil:

a. λ_i>1

b. Proporsi kumulatif *eigen value* banyaknya PC yang diambil 60-80%

```
> set.seed(101)
> pca <- prcomp(xnum,center=TRUE,scale.=TRUE)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.7544 1.9315 1.26956 1.12811 1.08124 0.92224 0.85241 0.68897 0.56441 0.56103
Proportion of Variance 0.3993 0.1963 0.08483 0.06698 0.06153 0.04476 0.03824 0.02498 0.01677 0.01657
Cumulative Proportion  0.3993 0.5957 0.68048 0.74746 0.80899 0.85375 0.89199 0.91698 0.93374 0.95031
                          PC11    PC12    PC13   PC14    PC15    PC16    PC17    PC18    PC19
Standard deviation     0.50727 0.45244 0.40492 0.3595 0.27046 0.25230 0.18201 0.11438 0.07664
Proportion of Variance 0.01354 0.01077 0.00863 0.0068 0.00385 0.00335 0.00174 0.00069 0.00031
Cumulative Proportion  0.96385 0.97463 0.98325 0.9901 0.99391 0.99726 0.99900 0.99969 1.00000
```

c. Scree plot



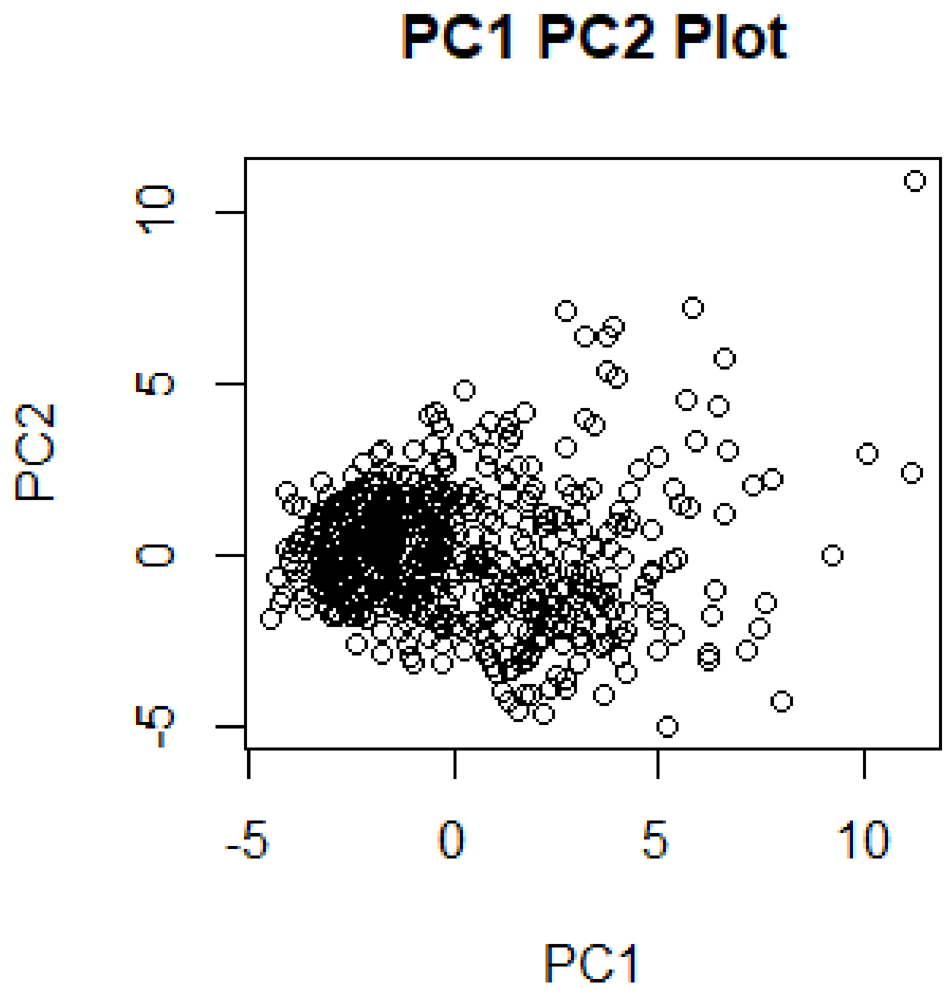
Screeplot
Eigenvalue = 1
Variances
0
2
4
6
1
2
3
4
5
6
7
8
9
10
12
14
16
18

3.Menentukan Variabel yang masuk PC1, PC2, PC3, PC4, PC5 dengan cara memilih nilai koefisien (nilai elemen *eigen vektor)* yang lebih besar

```
> View(pca$rotation)
```

	PC1	PC2	PC3	PC4	PC5
texture_mean	0.12993692	-0.19414352	0.638756166	0.15928006	-0.008578255
perimeter_mean	0.24588856	-0.34303626	-0.150768783	-0.08283581	0.044371031
area_mean	0.23454363	-0.34648129	-0.149695887	-0.09623109	0.009590223
smoothness_mean	0.20945013	0.13840330	-0.232014197	0.50483540	0.004850177
compactness_mean	0.32751669	0.04920405	-0.077838943	0.17109687	0.224357410
concavity_mean	0.34065966	-0.05118831	-0.060938749	-0.00877150	0.173914447
concave.points_mean	0.32394943	-0.16065938	-0.159782437	0.09075661	0.087733773
symmetry_mean	0.21250844	0.14925525	-0.106019295	0.40826507	-0.082135984
fractal_dimension_mean	0.15075090	0.38585979	0.013894260	0.24845657	0.167859249
radius_se	0.25969792	-0.13296101	-0.014304333	-0.02500141	-0.431296464
area_se	0.11727158	-0.12935008	-0.012310436	-0.08808386	-0.535402837
smoothness_se	0.08736619	0.28646286	0.102845442	0.02607489	-0.407965711
compactness_se	0.27882635	0.21289499	0.129854206	-0.24988173	0.122632359
concavity_se	0.25473462	0.19044365	0.111542671	-0.37702053	0.154448377
concave.points_se	0.28658046	0.11693670	0.004926141	-0.28929857	-0.010829325
symmetry_se	0.11544297	0.24840552	0.046532712	0.04656732	-0.428570594
fractal_dimension_se	0.21012152	0.31379391	0.131453002	-0.26744247	0.050822367
radius_worst	0.20635775	-0.29998388	-0.093500617	-0.03131978	-0.061791172
texture_worst	0.11822729	-0.20374369	0.614682717	0.25618807	0.080289714

- Jika ingin membuat plot (memvisualisasikan) lebih dari 2 variabel prediktor, dapat menggunakan PCA



Thankyou