Analisis Komponen Utama
(Principal Components Analysis)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Jika dalam suatu penelitian dimana setiap individu (satuan pengamatan) diamati dengan 2 peubah, maka banyak metode analisis sederhana yang dapat diterapkan, misalnya dengan menggunakan plot dalam ruang berdimensi dua, atau regresi linear sederhana.
- Diberbagai bidang ilmu, banyak penelitian mengamati setiap individu dengan banyak peubah (3 peubah atau lebih), sehingga sulit dilakukan plot individu dalam ruang dimensi 3, bahkan tidak mungkin untuk ruang dimensi lebih dari 3.
- Banyaknya peubah yang diamati pada setiap individu sering kali antar peubah memiliki hubungan (tidak saling bebas) yang dapat menimbulkan penyimpangan terhadap asumsi pada penerapan metode analisis tertentu, misalnya dalam regresi linear berganda akan mengakibatkan terjadinya multikolinieritas.
- Analisis Komponen Utama (AKU) menjadi salah satu yang dapat digunakan untuk memecahkan permasalah diatas

Apa itu AKU (1)

Beberapa literatur menyebutkan bahwa:

- AKU merupakan teknik transformasi data untuk mereduksi himpunan data berdimensi besar menjadi dimensi lebih kecil sehingga dapat memudahkan melihat distribusi data dalam ruang berdimensi lebih kecil, misalnya dalam ruang dimensi 2.
- AKU merupakan teknik transformasi data yang menghasilkan peubah baru (selanjutnya disebut dengan KU1, KU2, … dst) yang saling bebas
- AKU merukan teknik antara yang hasilnya dapat digunakan untuk menerapkan teknik analisis lanjutan yang misalnya memerlukan asumsi kebebasan antar peubah

Apa itu AKU (2)

- Jika setiap individu diamati sebanyak p peubah, maka akan terdapat sebanyak p KU sebagai peubah baru.
- Setiap KU (KU_1, KU_2…, KU_p) merupakan kombinasi linear dari peubah yang diamati (peubah asal)
- Antar KU bersifat saling bebas (orthogonal)
- KU_1 merupakan peubah dengan ragam paling besar, artinya menjelaskan keragaman data yang terbesar dibandingkan KU lainnya
- KU_2 merupakan peubah yang menjelaskan ragam terbesar ke-2, dan seterusnya untuk KU_3, KU_4…
- Dipilih beberapa KU pertama yang mana secara kumulatif dapat menjelaskan sebagian besar keragaman total.

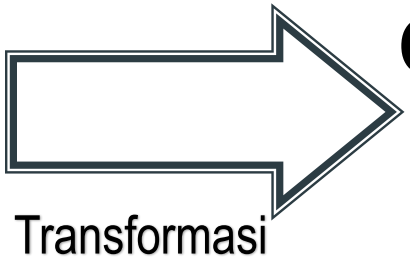
Struktur Data Amatan

Individu	Peubah				
	X1	X2	X3	...	Xp
1	x11	x12	x13		x1p
2	x21	x22	x23		x2p
3	x31	x32	x33		x3p
4	x41	x42	x43		x4p
5	x51	x52	x53		x5p
...	...	...	...	...	...
...	...	...	...	...	...
n	xn1	xn2	xn3		xnp

Gambaran Umum AKU

Gugus Peubah Asal

$\{X_1, X_2, \ldots, X_p\}$



Gugus Peubah KU

$\{KU_1, KU_2, \ldots, KU_p\}$

Hanya dipilih $k < p$ KU saja, namun mampu menjelaskan sebagian besar informasi

Cara Menentukan Komponen Utama (1)

- Didefinisikan Σ adalah matriks ragam-peragam berukuran (pxp) dari matriks pengamatan X berukuran (nxp), dimana n=jumlah individu dan p=jumlah peubah pengamatan.
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ adalah akar ciri yang berpadanan dengan vektor ciri $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ dari Σ, dengan panjang dari setiap vektor ciri masing masing adalah 1, atau $\mathbf{a}_i'\mathbf{a}_i = 1$ untuk $i = 1, 2, \ldots, p$.
- Setelah diperoleh akar ciri dan vector ciri dari matriks Σ selanjutnya dapat ditentukan KU1, KU2,… KUp

$$KU_1 = \mathbf{a}_1'\mathbf{x} = a_{11}x_1 + \ldots + a_{1p}x_p$$

$$KU_2 = \mathbf{a}_2'\mathbf{x} = a_{21}x_1 + \ldots + a_{2p}x_p$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$KU_p = \mathbf{a}_p'\mathbf{x} = a_{p1}x_1 + \ldots + a_{pp}x_p$$

Cara Menentukan Komponen Utama (2)

- Besar ragam dari setiap $KU_i = Var(KU_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i = \lambda_i$

  $Var(KU_1) = \lambda_1$

  $Var(KU_2) = \lambda_2$

  .....

  $Var(KU_p) = \lambda_p$

- Total ragam peubah asal X = tr(Σ), dan ini sama dengan penjumlahan dari seluruh akar ciri
- Jadi kontribusi setiap KU ke-j adalah sebesar

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

- Cov $(KU_i, KU_j)=0$, $i \neq j=1,2,\ldots p$

Menggunakan matriks korelasi atau ragam peragam?

Dalam menentukan KU, selain menggunakan matriks ragam-peragam Σ dapat juga menggunakan matriks korelasai R. Jika satuan pengukuran dari peubah yang diamati berbeda maka akan mempengaruhi besarnya keragaman peuban. Oleh karena itu dalam kasus seperti ini direkomendasikan menggunakan matriks korelasi.

Cara menentukan Banyaknya KU yang akan digunakan (1)

- Pada prinsipnya jumlah KU yang akan digunakan tidak mengorbankan terlalu banyak informasi yang hilang
- Pada umumnya didasarkan pada kumulatif proporsi keragaman total yang mampu dijelaskan oleh KU. Misalnya kontribusi KU_1=70% dan KU_2=23%, maka dengan KU_1 dan KU_2 sudah mampu menjelaskan sebesar 93% dari keragaman total.
- Tidak ada patokan baku berapa batas minimum tersebut, sebagian bukau menyebutkan 70%, 80%, bahkan ada yang 90%.

Cara menentukan Banyaknya KU yang akan digunakan (2)

- Cara lain yang juga sering digunakan adalah grafik plot scree.
- Plot scree merupakan plot antara akar ciri λ_k (ordinat) dengan k (absis).
- Dengan menggunakan metode ini, banyaknya komponen utama yang dipilih, yaitu k, adalah jika pada titik k tersebut plotnya curam ke kiri tapi tidak curam di kanan. Ide yang ada di belakang metode ini adalah bahwa banyaknya komponen utama yang dipilih sedemikian rupa sehingga selisih antara akar ciri yang berurutan sudah tidak besar lagi.

Ilustrasi 1

Berikut adalah data catatan waktu hasil tujuh nomor cabang lari atletik peserta yang berasal dari 55 negara pada salah satu event olimpiade yaitu lari 100 meter, 200 meter, 400 meter, 800 meter, 1500 meter, 3000 meter, dan maraton. Tiga nomor cabang lari pertama dicatat dalam satuan detik, sedangkan empat nomor yang lain dalam menit.

Berdasarkan data tersebut ingin dianalisis performa 7 cabang lari dari 55 negara tersebut

Permasalahan dan Solusi

- Untuk melihat performa 55 negara dari 7 cabang lari sulit hanya dilihat dari rata-rata, karena ketujuh cabang tersebut memiliki satuan yang berbeda dan performannya berbeda untuk setiap cabang
- Karena terdapat 7 cabang lari (7 peubah), sulit untuk dilihat melalui grafik
- Salah satu metode yg dapat digunakan adalah AKU untuk mereduksi data dari 7 dimensi kedalam 2 dimensi
- Karena satuan peubah tidak sama, maka AKU dilakukan melalui matriks korelasi

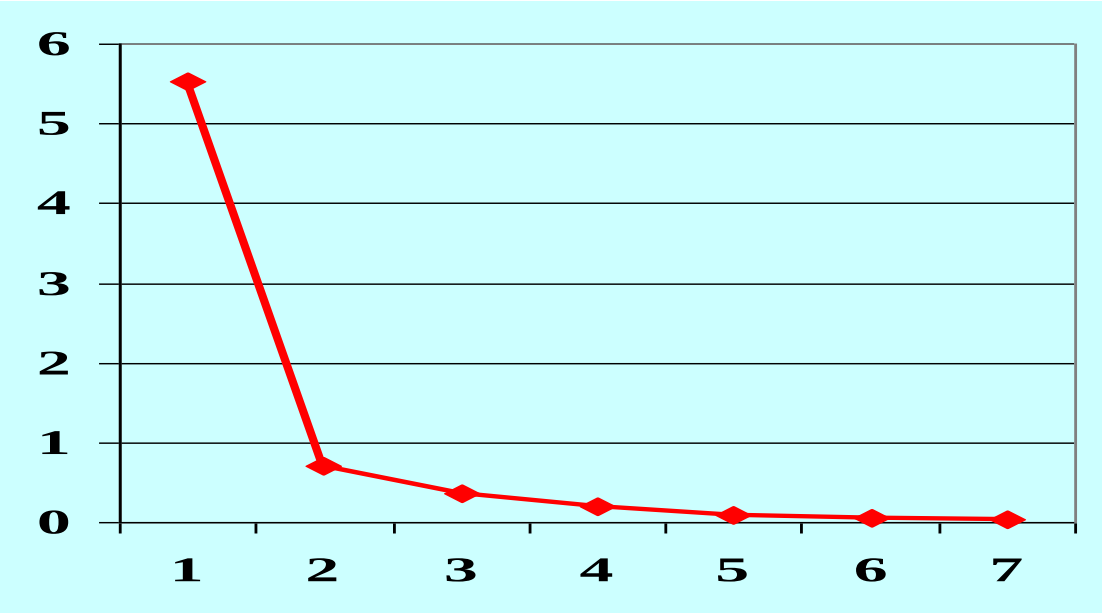
Matriks Korelasi

	m100	m200	m400	m800	m1500	m3000	marathon
m100	1.0000	0.9528	0.8350	0.7277	0.7163	0.7417	0.5423
m200	0.9528	1.0000	0.8572	0.7241	0.7029	0.7099	0.5444
m400	0.8350	0.8572	1.0000	0.8981	0.7757	0.7776	0.5507
m800	0.7277	0.7241	0.8981	1.0000	0.8260	0.8636	0.6545
m1500	0.7163	0.7029	0.7757	0.8260	1.0000	0.9031	0.6996
m3000	0.7417	0.7099	0.7776	0.8636	0.9031	1.0000	0.7966
marathon	0.5423	0.5444	0.5507	0.6545	0.6996	0.7966	1.0000

Akar Ciri Dari Matriks Korelasi

	Eigenvalue	Difference	Proportion	Cumulative
1	5.53319890	4.81746883	0.7905	0.7905
2	0.71573007	0.35411502	0.1022	0.8927
3	0.36161505	0.15335511	0.0517	0.9444
4	0.20825995	0.11607781	0.0298	0.9741
5	0.09218213	0.04086896	0.0132	0.9873
6	0.05131317	0.01361245	0.0073	0.9946
7	0.03770072		0.0054	1.0000

Plot Scree



Penentuan Banyaknya KU

- Dengan menggunakan 2 KU sudah mencapai proporsi keragaman 89.27%, artinya dengan 2 KU sudah mampu menjelaskan 89,27% keragaman data
- Pada k = 2 terlihat gambar scree plot sangat curam di kiri tapi landai di kanan. Jadi 2 KU yang digunakan sudah mencukupi.

Vektor Ciri dari setiap akar ciri

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
m100	0.378202	-.426104	0.359297	-.165099	-.331229	0.225902	0.598584
m200	0.376416	-.452874	0.363819	-.011005	0.175249	0.037974	-.698982
m400	0.391311	-.272232	-.325636	0.378804	0.371464	-.556664	0.274544
m800	0.390624	0.067673	-.512111	0.402954	-.250932	0.579870	-.137794
m1500	0.385043	0.230072	-.245359	-.680608	0.481480	0.195655	0.072641
m3000	0.395890	0.308242	-.074146	-.249112	-.615938	-.509888	-.203317
marathon	0.323383	0.621855	0.551857	0.376128	0.217762	0.056004	0.110204

KU Pertama dan KU Kedua

- KU1 mampu menerangkan keragaman data sebesar 79.05% dengan persamaan:

KU1=0.378202X1+0.376416X2+0.391311X3+0.390624X4+ 0.385043X5+0.395890X6+0.323383X7

- KU2 menerangkan keragaman data sebesar 10.22% dengan persamaan:

KU2=-0.426104X1-0.452874X2-0.272232X3+0.067673X4+ 0.230072X5+ 0.308242X6+0.621855X7

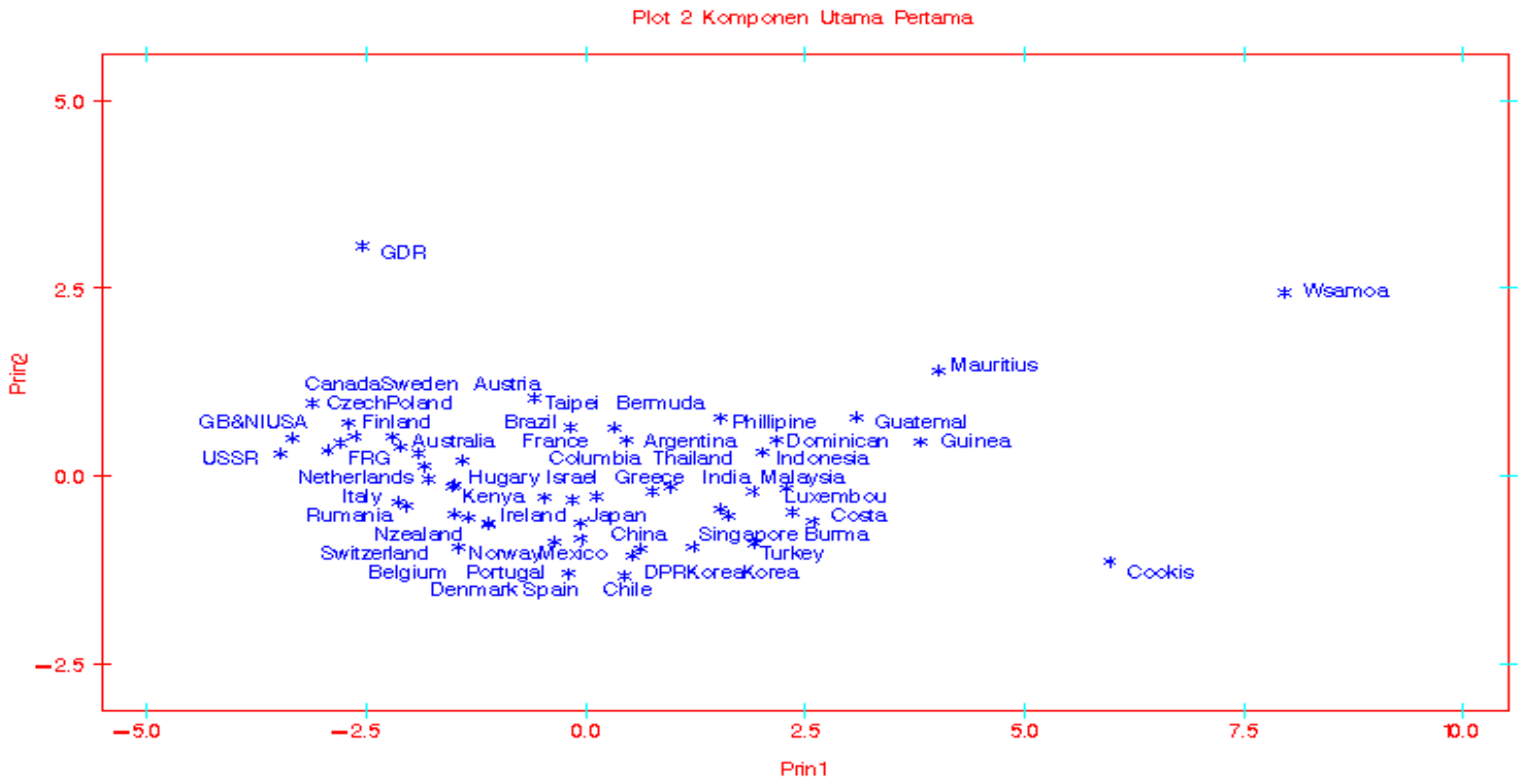
- Sehingga jika digunakan dua komponen utama akan didapatkan keragaman total yang mampu dijelaskan keduanya adalah 89.27%.

Koordinat baru 55 negara berdasarkan KU1 dan KU2

Jika skor komponen pertama ini diurutkan maka diperoleh hasil 10 terbaik adalah

Obs	country	Prin1	Prin2
1	USSR	-3.46947	0.29798
2	USA	-3.33124	0.50401
3	Czech	-3.10484	0.97537
4	FRG	-2.93434	0.34671
5	GB&NI	-2.79248	0.44274
6	Poland	-2.69963	0.70626
7	Canada	-2.61758	0.53196
8	GDR	-2.54492	3.07144
9	Finland	-2.19832	0.52134
10	Italy	-2.12838	-0.34299

Plot Skor performa 55 negara berdasarkan KU1 dan KU2



Ilustrasi 2

Penerapan AKU pada Regresi Linear Berganda untuk melihat pengaruh 7 sifat agronomis suatu tanaman (X1, X2,….X7) terhadap produksi (Y)

Permasalahan dan Solusi

- Dalam analisis regresi linear berganda salah satu asumsi yang diperlukan adalah tidak terjadi multikolinieritas (terjadi korelasi) antar peubah bebas
- Multikolinieritas akan menimbulkan terjadinya kesalahan dalam pengujian hipotesis terhadap koefisien regresi dan besar serta tanda dari penduga koefisien regresi
- Salah satu solusinya adalah data peubah bebas ditransformasi ke KU yang saling bebas. Peubah bebas terlebih dahulu datanya ditransformasi ke peubah normal baku Z

Korelasi Antar Peubah Bebas

	X1	X2	X3	X4	X5	X6	X7
X1	1.000	0.8061	0.8511	0.9015	0.9157	-0.8397	0.7843
	0.0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
X2	0.8061	1.000	0.6279	0.7361	0.8448	-0.6624	0.7592
	0.0001	0.0	0.0053	0.0005	0.0001	0.0027	0.0003
X3	0.8511	0.6279	1.000	0.84244	0.70182	-0.8079	0.70844
	0.0001	0.0053	0.0	0.0001	0.0012	0.0001	0.0010
X4	0.9015	0.7361	0.84244	1.000	0.8538	-0.7767	0.8297
	0.0001	0.0005	0.0001	0.0	0.0001	0.0001	0.0001
X5	0.9157	0.8448	0.70182	0.8538	1.000	-0.7792	0.8536
	0.0001	0.0001	0.0012	0.0001	0.0	0.0001	0.0001
X6	-0.8397	-0.6624	-0.8079	-0.7767	-0.7792	1.000	-0.6512
	0.0001	0.0027	0.0001	0.0001	0.0001	0.0	0.0034
X7	0.7843	0.7592	0.70844	0.8297	0.8536	-0.6512	1.000
	0.0001	0.0003	0.0010	0.0001	0.0001	0.0	0.0

Terjadi Multikolinieritas antar peubah bebas

Nilai VIF (deteksi multikolinearitas)

Peubah Bebas (Xi)	Varians Inflantion Factor (VIF)
X1	16.40
X2	3.70
X3	6.80
X4	7.60
X5	14.20
X6	4.20
X7	5.40

Terjadi Multikolinieritas antar peubah bebas sehingga jika dilakukan analisis regresi langsung akan bias

Analisis Komponen Utama Terhadap peubah bebas

Peubah	Komponen Utama						
	K1	K2	K3	K4	K5	K6	K7
Z1	0.403	0.083	0.134	0.063	0.447	0.410	-0.664
Z2	0.358	-0.521	0.439	0.556	-0.227	-0.216	0.006
Z3	0.365	0.541	-0.261	0.506	-0.216	0.308	0.329
Z4	0.392	0.096	-0.339	0.024	0.473	-0.702	0.069
Z5	0.393	-0.293	0.142	-0.387	0.294	0.357	0.613
Z6	-0.364	-0.453	-0.493	0.451	0.384	0.254	0.082
Z7	0.368	-0.368	-0.588	-0.279	-0.493	0.074	-0.253
Akar ciri (Ragam)	57,345	0.5038	0.2993	0.1890	0.1502	0.0897	0.0336
Proporsi	0.819	0.072	0.043	0.027	0.021	0.013	0.005
Proporsi kumulatif	0.819	0.891	0.934	0.961	0.982	0.995	1,000

Analisis Regresi dengan 4 KU Pertama

$$Y = 6.66 + 0.634\ K1 - 0.424\ K2$$

Peubah	Koef	St.dev	t-student	P
Konstan	6.665	0.0932	71.53	0.000
K1	-0.6339	0.0400	15.83	0.000
K2	-0.4239	0.1351	-3.14	0.011
K3	-0.0783	0.1753	-0.45	0.664
K4	-0.4100	0.2206	-1.86	0.093

Transformasi ke peubah Z

$$Y = 6.66 + 0.112\ Z1 + 0.351\ Z2 + 0.096\ Z3 + 0.102\ Z4 + 0.267\ Z5 - 0.059\ Z6 + 0.286\ Z7$$

Transformasi ke peubah asal X

$$Y = 18.47 + 0.0166\ X_1 + 0.139\ X_2 + 0.013\ X_3 + 0.059\ X_4 + 0.0158\ X_5 - 0.009\ X_6 + 0.140\ X_7$$

Terimakasih

Principal Component Analysis

gdito

4/9/2020

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- factoextra
- ggcorrplot
- openxlsx

Silahkan install jika belum ada

```
install.packages("factoextra")
install.packages("ggcorrplot")
install.packages("openxlsx")

library(factoextra)
library(ggcorrplot)
```

Data Pelari Wanita

Berikut adalah data catatan waktu hasil tujuh nomor cabang lari atletik wanita yang berasal dari 55 negara pada salah satu event olimpiade yaitu lari 100 meter, 200 meter, 400 meter, 800 meter, 1500 meter, 3000 meter, dan maraton. Tiga nomor cabang lari pertama dicatat dalam satuan detik, sedangkan empat nomor yang lain dalam menit.

Tahap 1 Menyiapkan data di R

```
data_women_records <- openxlsx::read.xlsx("E:/APG/R
APG/women_track_records.xlsx")
head(data_women_records)

##    l00m  200m  400m 800m 1500m 3000m Marathon   country
## 1 11.61 22.94 54.50 2.15  4.43  9.79   178.52 argentina
## 2 11.20 22.35 51.08 1.98  4.13  9.08   152.37 australia
## 3 11.43 23.09 50.62 1.99  4.22  9.34   159.37   austria
## 4 11.41 23.04 52.00 2.00  4.14  8.88   157.85   belgium
## 5 11.46 23.05 53.30 2.16  4.58  9.81   169.98   bermuda
## 6 11.31 23.17 52.80 2.10  4.49  9.77   168.75    brazil
```

Note : `openxlsx::read.xlsx` berarti menggunakan fungsi `read.xlsx` yang berada pada package `openxlsx` tanpa memanggil packagenya terlebih dahulu menggunakan `library`. Fungsi `head` digunakan untuk menampilkan data 6 baris pertama.

Untuk keperluan analisis selanjutnya nama negara (country) akan dijadikan nama baris pada data. Hal ini dilakukan dengan menggunakan fungsi `rownames`.

```
rownames(data_women_records) <- data_women_records$country
data_women_records <- data_women_records[,-8]
```

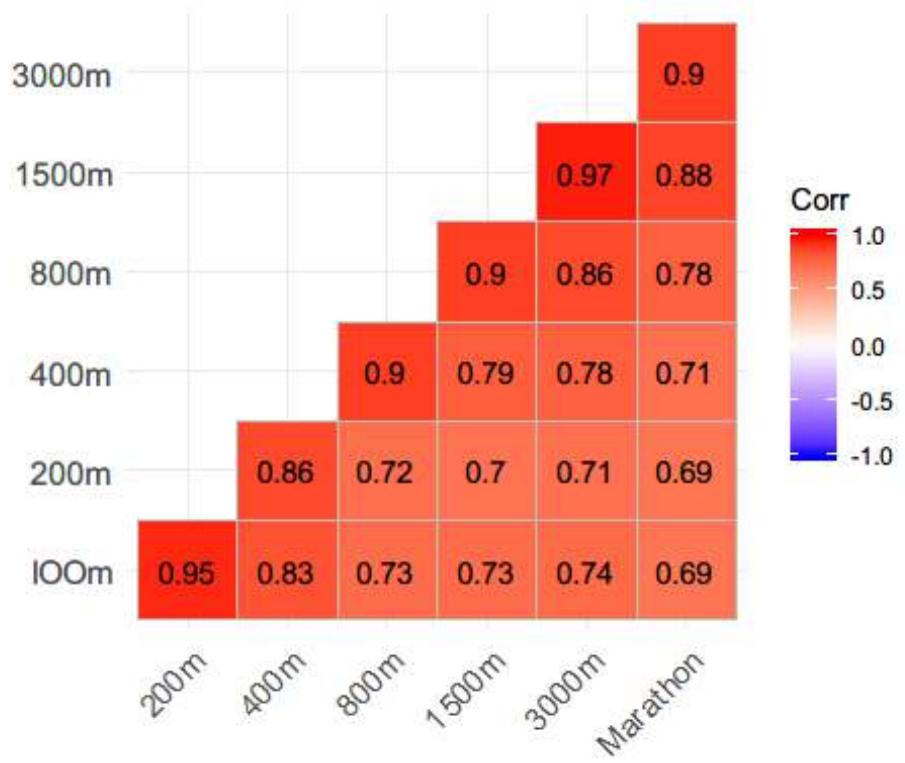
`data_women_records[,-8]` berarti kita menghilangkan kolom kedelapan pada data (kolom country).

Tahap 2 Eksplorasi dengan menggunakan matrix korelasi

```
cor_women <- cor(data_women_records)
cor_women

##               lOOm      200m      400m      800m     1500m     3000m
## lOOm     1.0000000 0.9527911 0.8346918 0.7276888 0.7283709 0.7416988
## 200m     0.9527911 1.0000000 0.8569621 0.7240597 0.6983643 0.7098710
## 400m     0.8346918 0.8569621 1.0000000 0.8984052 0.7878417 0.7776369
## 800m     0.7276888 0.7240597 0.8984052 1.0000000 0.9016138 0.8635652
## 1500m    0.7283709 0.6983643 0.7878417 0.9016138 1.0000000 0.9691690
## 3000m    0.7416988 0.7098710 0.7776369 0.8635652 0.9691690 1.0000000
## Marathon 0.6863358 0.6855745 0.7054241 0.7792922 0.8779334 0.8998374
##           Marathon
## lOOm     0.6863358
## 200m     0.6855745
## 400m     0.7054241
## 800m     0.7792922
## 1500m    0.8779334
## 3000m    0.8998374
## Marathon 1.0000000
```

Agar mudah dilihat matrix korelasi ini bisa dibuat dalam bentuk grafik dengan cara berikut.

```
ggcorrplot(cor_women,type="lower",lab = TRUE)
```



Tahap 3 Menerapkan PCA (AKU)

Dalam R, Penerapan PCA ini dapat dilakukan dengan menggunakan `fungsi prcomp`. Fungsi ini memiliki argumen `scale.` dan `center`. Jika kedua argumen ini `TRUE` maka matrix yang digunakan untuk menghitung PCA adalah matrix korelasi. Namun, jika kedua argumen ini `FALSE` atau `scale.=FALSE`, maka matrix yang digunakan adalah matrix covariance.

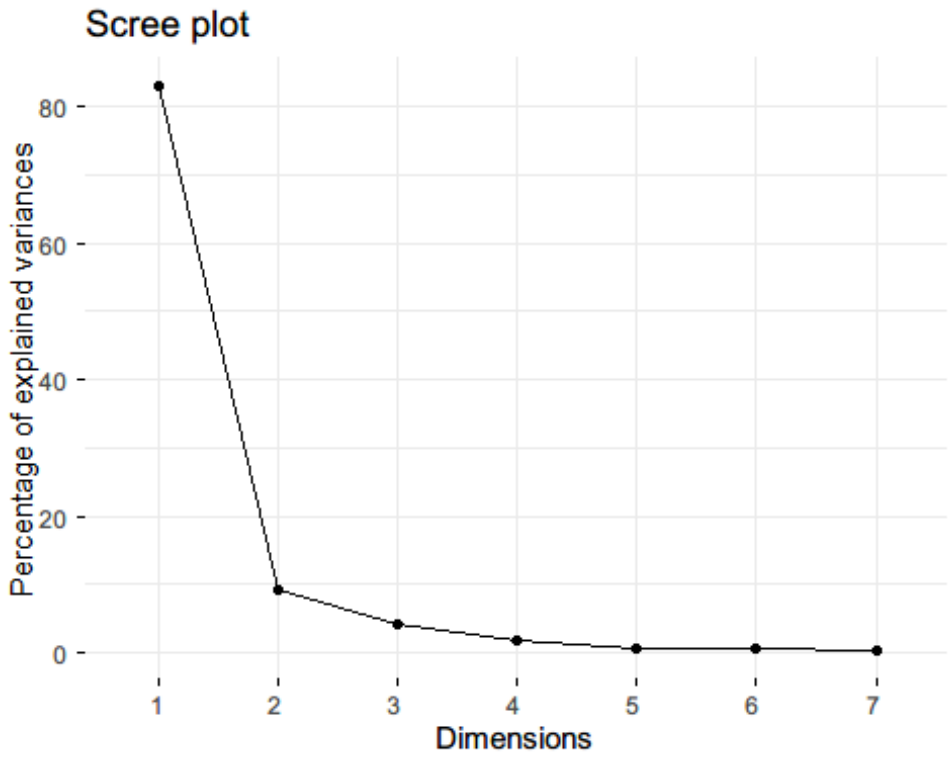
```
pca_women_records <- prcomp(data_women_records,scale.=TRUE,center=TRUE)
summary(pca_women_records)

## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.4095 0.80848 0.54762 0.35423 0.23198 0.19761
## Proportion of Variance 0.8294 0.09338 0.04284 0.01793 0.00769 0.00558
## Cumulative Proportion  0.8294 0.92276 0.96560 0.98353 0.99122 0.99679
##                           PC7
## Standard deviation     0.14981
## Proportion of Variance 0.00321
## Cumulative Proportion  1.00000
```

Hasil yang dikeluar dari sintaks diatas ada tiga macam yaitu Standard deviation, Proportion of Variance dan Cumulative Proportion dari masing-masing Komponen Utama (Principal Component). Standard deviaton merupakan akar dari akar ciri (eigenvalue). Dalam hal ini akar ciri berperan sebagai variance dari masing-masing komponen utama. Proportion of Variance didapatkan dari akar ciri pada masing-masing komponen dibagi dengan total akar ciri. Proportion of Variance menjelaskan seberapa besar keragaman peubah asal yang

dapat dijelaskan oleh masing-masing komponen utama. Semakin besar nilainya berarti semakin baik pula komponen utama tersebut untuk merepresentasikan peubah asal.

Cumulative Proportion menjelaskan seberapa besar keragaman yang dapat dijelaskan oleh komponen utama secara kumulatif. Misalnya saja dengan menggunakan dua komponen utama saja (PC1 dan PC2), sudah bisa menjelaskan lebih dari 92% keragaman dari data. Berdasarkan hal ini, kita akan memilih menggunakan dua komponen utama saja.

```
fviz_screeplot(pca_women_records,geom="line")
```



Hal lain yang bisa dilakukan untuk menentukan berapa banyak komponen utama yang digunakan adalah dengan screeplot. Fungsi untuk menampilkan screeplot pada R adalah `fviz_screeplot` yang didapat dari package `factoextra`.

Banyaknya komponen utama bisa ditentukan dengan screeplot dengan melihat di komponen utama yang mana garisnya berbentuk seperti siku (elbow). Pada gambar diatas garis membentuk siku saat berada di komponen utama kedua (dimension kedua). Sehingga banyaknya komponen utama yang digunakan sebanyak dua (Komponen Utama 1 dan Komponen Utama 2)

Tahap 4 Interpretasi PCA (AKU)

Interpretasi metode PCA dapat dilakukan dengan menggunakan vektor ciri pada masing-masing komponen utama. Semakin besar vektor ciri pada komponen utama tertentu maka semakin besar pula kontribusi dari peubah asal untuk membangun komponen utama tersebut. Catatan lain yang perlu diperhatikan adalah nilai negatif pada vektor ciri menandakan peubah asal memberikan kontribusi yang berkembalikan pada pembentukan

komponen utama. Dalam konteks vektor ciri negatif, semakin besar nilai peubah asal semakin kecil nilai pada komponen utama.

```
pca_women_records$rotation

##                PC1        PC2         PC3         PC4         PC5
## lOOm     0.3683561  0.4900597 -0.28601157  0.31938631  0.23116950
## 200m     0.3653642  0.5365800 -0.22981913 -0.08330196  0.04145457
## 400m     0.3816103  0.2465377  0.51536655 -0.34737748 -0.57217791
## 800m     0.3845592 -0.1554023  0.58452608 -0.04207636  0.62032379
## 1500m    0.3891040 -0.3604093  0.01291198  0.42953873  0.03026144
## 3000m    0.3888661 -0.3475394 -0.15272772  0.36311995 -0.46335476
## Marathon 0.3670038 -0.3692076 -0.48437037 -0.67249685  0.13053590
##                  PC6         PC7
## lOOm      0.619825234  0.05217655
## 200m     -0.710764580 -0.10922503
## 400m      0.190945970  0.20849691
## 800m     -0.019089032 -0.31520972
## 1500m    -0.231248381  0.69256151
## 3000m     0.009277159 -0.59835943
## Marathon  0.142280558  0.06959828
```

Karena kita hanya menggunakan dua komponen saja, maka vector ciri yang akan dinterpretasikan hanya pada PC1 dan PC2. PC1 memiliki vektor ciri yang relatif sama yaitu berkisar di 0.3 untuk semua cabang lomba. Vektor ciri yang relatif sama ini menandakan bahwa kontribusi peubah asal untuk membangun komponen utama ini relatif sama. Artinya nilai-nilai yang ada di PC1 (score value) dapat menggambarkan waktu lari untuk semua cabang lomba. Oleh karena itu kita dapat dapat menggunakan PC1 untuk menentukan negara mana yang memiliki pelari tercepat untuk semua kategori lomba. Vektor ciri di PC2 memiliki nilai positif untuk cabang lari jarak dekat (100m -400m) dan nilai negatif untuk cabang lari jarak jauh(800m-marathon). Hal ini berarti semakin besar score value pada PC2 maka waktu lari cabang jarak dekat semakin lambat namun waktu lari untuk cabang jarak jauh semakin cepat. Oleh karena itu, PC2 dapat digunakan untuk menentukan negara mana yang pada cabang lari jarak dekat waktunya mirip seperti cabang lari jarak jauh.

Note: Interpretasi komponen utama memiliki subjektifitas yang tinggi, oleh karena itu setiap orang menginterpretasikanya berbeda

Hal terakhir yang bisa diinterpretasikan adalah score value pada PC1 dan PC2. Score value merupakan observasi/koordinat baru pada peubah komponen utama. Dalam konteks data pelari diatas, observasinya adalah negara, sehingga kita dapat memberi insight cabang perlombaan lari dari setiap negara.

Untuk melihat score value pada komponen utama dapat dilihat dengan menggunakan sintaks berikut.

```
pca_women_records$x
```

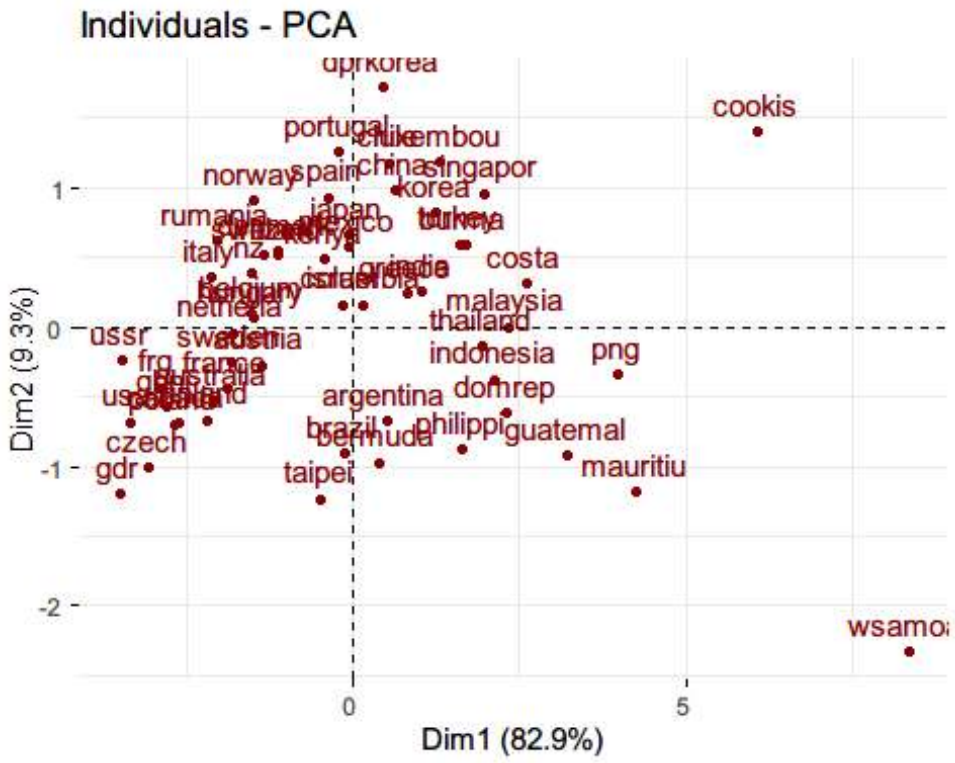
```
##                 PC1          PC2         PC3         PC4          PC5
## argentina  0.52726229 -0.674719057  0.61558926  0.04552931 -0.0025575859
## australia -2.09355119 -0.532798687 -0.04317487  0.18737792 -0.2183555124
## austria   -1.38044331 -0.274995495 -0.53231305  0.42623519 -0.0255039187
## belgium   -1.50998970  0.090963946 -0.08345684 -0.03942270 -0.0303263716
## bermuda    0.38781757 -0.976409482  0.64887210  0.47445652  0.2043216118
## brazil    -0.11839732 -0.911515514  0.32214131  0.34096557 -0.0959612766
## burma      1.68203844  0.586191540  0.07582590 -0.14511731  0.6034322854
## canada    -2.60813211 -0.695287897  0.10954223  0.03328484  0.1410820801
## chile      0.54783013  1.171926564 -0.23733316 -0.09612578 -0.2156317054
## china      0.64127320  0.980047440  0.05370738  0.02293887 -0.0579061799
## columbia   0.14157212  0.154468463  0.21456812  0.17614694  0.1895221521
## cookis     6.07727834  1.399503624 -0.21282983 -0.28085726 -0.0529187535
## costa      2.61922856  0.305956420  1.09460087  0.41203614 -0.5847203364
## czech     -3.05379905 -1.012730429 -1.04879114  0.37317120 -0.0258660649
## denmark   -1.11637538  0.540131452  0.41098527 -0.17591883 -0.1041309755
## domrep     2.29543640 -0.616097039  0.64633093 -0.26243482  0.3964092177
## finland   -2.18183995 -0.670248707  0.08087437  0.08706670  0.3299422783
## france    -1.89216992 -0.443832046  0.14465457 -0.05322917 -0.1896245118
## gdr       -3.50601681 -1.202500275 -0.52603497 -0.12430576  0.0884050075
## frg       -2.92577741 -0.437137537 -0.20120561 -0.02584472  0.0480344770
## gbni      -2.78315609 -0.578349737  0.13304952 -0.13024804  0.0417404253
## greece     0.81424542  0.233714829 -0.15991295 -0.08693903 -0.4548804436
## guatemal   3.22729824 -0.919057694  0.44304609 -0.09073846  0.3545878042
## hungary   -1.47721337  0.059384256 -0.15006131  0.12504871  0.0924438246
## india      1.01453673  0.253605564 -0.51071114  0.05275360  0.0509072296
## indonesia  2.11236466 -0.378269923  0.33669289 -0.18769360  0.3751519942
## ireland   -1.11735099  0.513043238  0.44713600 -0.08797261 -0.0255638127
## israel    -0.14296749  0.155867013  0.75136633 -0.16605776 -0.2905821673
## italy     -2.13954076  0.351991902 -0.07731676 -0.29052538 -0.2244816386
## japan     -0.05923268  0.657973909  0.40042678  0.42998430  0.1230753651
## kenya     -0.43089430  0.480611126 -0.75309705 -0.32602370 -0.0643810184
## korea      1.23386149  0.814398564  0.55640268  0.23959883  0.0129823046
## dprkorea   0.46229683  1.717885467 -1.91963681  0.34183702  0.3375496932
## luxembou   1.30174495  1.182174513 -0.10512035 -0.03151618 -0.4495547106
## malaysia   2.34053535 -0.001259817  0.11819610  0.84655411  0.1277242627
## mauritiu   4.23384717 -1.180202966 -0.08772947 -1.39411471 -0.1640608518
## mexico    -0.06348187  0.568969717 -0.09040891  0.45214535 -0.2961853317
## netherla  -1.79442661 -0.047085355  0.14270838 -0.10800877 -0.3625892311
## nz        -1.51125893  0.377920885  0.06112304  0.36901628  0.2627172276
## norway    -1.48300990  0.904195203  0.38741591 -0.14529920  0.1311735490
## png        3.98086034 -0.340244237 -0.28834865 -0.29398974  0.1595586421
## philippi   1.64018955 -0.876042797  0.22214354 -0.02227534  0.2056514197
## poland    -2.67209659 -0.702323548 -0.59597055 -0.02384024  0.0364171238
## portugal  -0.22428415  1.252662130  0.46217831 -0.02349218  0.2384703479
## rumania   -2.02982623  0.618046803 -0.83844940 -0.46958197 -0.0740123635
## singapor   1.97013151  0.951982007 -0.38929662  0.40329233  0.0732785524
## spain     -0.35565287  0.925478452 -0.05010366 -0.10341725  0.0927143013
## sweden    -1.82775494 -0.254820864  0.24805824 -0.14902201 -0.0006283576
## switzerl  -1.34665382  0.514180988  0.24518048 -0.22415740 -0.0858444392
```

```
## taipei    -0.50011940 -1.234653297  0.31159932 -0.04781148  0.0094018502
## thailand   1.95317730 -0.139873925  0.81360772  0.65551039 -0.1576314482
## turkey     1.60820411  0.594342560  0.16105018 -0.81523551  0.1477445097
## usa       -3.33581190 -0.685104574  0.38123893 -0.27057989 -0.1954822627
## ussr      -3.46468721 -0.245078447 -0.64637481 -0.20743799 -0.0824770130
## wsamoa     8.33288156 -2.326979228 -1.49263486  0.40428466 -0.3425812545
##                   PC6          PC7
## argentina  0.457907931 -0.079962873
## australia  0.137966927 -0.009815157
## austria   -0.081682704 -0.106172559
## belgium    0.062877332  0.138490873
## bermuda   -0.049426348  0.047829472
## brazil    -0.300448073 -0.006724259
## burma      0.277326099  0.055825566
## canada    -0.116436013 -0.117243747
## chile      0.128882585  0.003967955
## china      0.046618764  0.172340012
## columbia  -0.324561612 -0.122442477
## cookis    -0.055706058 -0.289275038
## costa     -0.065656582 -0.044687998
## czech      0.047409019  0.188230868
## denmark   -0.179964399  0.322413538
## domrep    -0.006777189  0.321472376
## finland   -0.031636754 -0.182734885
## france    -0.035807689  0.053223788
## gdr       -0.047152707 -0.176067724
## frg       -0.191479813  0.086729269
## gbni       0.012368224  0.059974475
## greece     0.093873143 -0.121106747
## guatemal  -0.279702884 -0.030512752
## hungary    0.061704372 -0.002880404
## india      0.134490600 -0.442230795
## indonesia -0.013434970 -0.058448424
## ireland   -0.149414709 -0.043637389
## israel    -0.090944777 -0.094991490
## italy      0.012049601  0.080196477
## japan     -0.182028982  0.141996717
## kenya      0.119151878 -0.011054538
## korea     -0.028182131 -0.027699883
## dprkorea  -0.561883421 -0.072075631
## luxembou  -0.115327186  0.123404789
## malaysia   0.374021886 -0.140160623
## mauritiu  -0.321676670 -0.209056067
## mexico     0.402826018 -0.127280417
## netherla   0.051567793 -0.071966309
## nz         0.078135023  0.198553009
## norway     0.226377016  0.086297791
## png        0.111686449  0.039358192
## philippi   0.267820061 -0.094783185
## poland     0.144920475 -0.248593496
```

```
## portugal  -0.041084319  0.041047319
## rumania   -0.050795933  0.167824160
## singapor   0.091437114  0.018007615
## spain      0.124420761 -0.021530131
## sweden    -0.159842868  0.036333974
## switzerl   0.047330468  0.068598802
## taipei     0.024237079 -0.093082005
## thailand  -0.482014923 -0.032198783
## turkey     0.287104548  0.191270368
## usa       -0.047049486  0.040368337
## ussr       0.097960158  0.017756855
## wsamoa     0.087647875  0.376903191
```

Agar lebih mudah dalam menginterpretasikan score value maka digunakaan grafik dibawah ini.

```
fviz_pca_ind(pca_women_records,col.ind = "darkred")
```



Berdasarkan grafik score value dapat diketahui bahwa negara yang memiliki catatan waktu pelari terlambat untuk semua cabang lomba adalah negara wsamoa. Hal ini dikarenakan wsamoa score value wsakoa untuk PC1 (Dim1) paling besar diantara yang lain. Walaupun negara wsamoa memiliki cabang lari terlama disemua cabang lomba, namun perbedaan waktu terkecil antara pelari jarak jauh dan jarak dekat adalah negara wsamoa. Hal ini berarti pelari untuk lomba jarak dekat sangat lambat karena memiliki waktu yang hampir mirip seperti pelari jarak jauh. Sedangkan negara yang memiliki pelari tercepat untuk semua cabang lomba adalah gdr.

Demikianlah contoh interpretasi dari PCA, apakah anda punya interpretasi lain?

ANALISIS FAKTOR (FACTOR ANALYSIS)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Dalam bidang penelitian tertentu, misalnya psikometri, sering kali ingin menggambarkan karakteristik individu tetapi tidak dapat diukur secara langsung (unobservable), misalnya intelegensi seseorang, prestasi siswa, bentuk ideal tubuh, dls.
- Karakteristik individu tersebut, yang selanjutnya disebut faktor, kemungkinan dapat dicirikan oleh segugus peubah yang dapat diukur (observable).
- Analisis Faktor merupakan suatu metode untuk menggambarkan (jika ada) pola hubungan internal banyak peubah sehingga membentuk beberapa kelompok unobservable faktor yang memiliki makna.
- Peubah-peubah yang membentuk suatu faktor tersebut memiliki korelasi tinggi didalam faktor itu sendiri dan berkorelasi rendah dengan faktor lainnya.
- Analisis faktor ini sering dikatakan sebagai pengembangan dari AKU

Struktur Data Amatan

Individu	Peubah				
	X1	X2	X3	...	Xp
1	x11	x12	x13		x1p
2	x21	x22	x23		x2p
3	x31	x32	x33		x3p
4	x41	x42	x43		x4p
5	x51	x52	x53		x5p
...	...	...	...	...	...
...	...	...	...	...	...
n	xn1	xn2	xn3		xnp

Model Faktor Ortogonal (1)

- Didefinisikan vektor peubah acak **observable X** dengan p komponen memiliki nilai tengah **μ** dan matriks peragam **Σ**.
- Model factor mendefinisikan bahwa vector X merupakan fungsi linear dari beberapa peubah acak unobservable F1, F2,…Fm (disebut factor umum/common factor) dan p sumber keragaman lainnya (disebut error).
- Model faktor dapat dituliskan dalam bentuk:

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2$$

:

:

:

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p$$

Model Faktor Ortogonal (2)

- Model factor dapat ditulis dalam bentuk matriks:

$$\underset{(px1)}{(\underline{X} - \underline{\mu})} = \underset{(pxm)}{L} \quad \underset{(mx1)}{\underline{F}} + \underset{(px1)}{\underline{\varepsilon}}$$

Dimana koefisien l_{ij} dikatakan sebagai loading dari peubah ke-j pada factor ke-i sehingga matriks L adalah matriks dari loading faktor.

F adalah vektor acak dari $F_1, F_2, \ldots F_m$. dan ε vektor galat/error dari $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_p$ dimana kedua vector tersebut unobservable.

Model Faktor Ortogonal (3)

- Yang membedakan antara model factor dan regresi linear berganda adalah bahwa dalam regresi vector peubah F observable sehingga koefisien L dapat dengan mudah diduga.
- Meskipun vector peubah F dalam model factor unobservable, melalui beberapa asumsi tambahan terhadap vector acak F dan **ε** maka dapat dilakukan pendugaan terhadap model factor

Asumsi-Asumsi Model Faktor (1)

- $E(\underline{F}) = 0, \quad E(\underline{\varepsilon}) = 0$
- $Cov(\underline{F}) = E(\underline{F}\underline{F}') = I$
- $Cov(\underline{\varepsilon}) = E(\underline{\varepsilon}\,\underline{\varepsilon}') = \psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$
- $\underline{F}$ dan $\underline{\varepsilon}$ saling bebas
- $Cov(\underline{\varepsilon},\underline{F}) = E(\underline{\varepsilon},\underline{F}) = 0$

Berdasarkan asumsi dan model factor diatas maka struktur peragam model factor dapat dinyatakan:

1. $Cov(X) = LL' + \psi$ atau

$$Var(X_i) = l_{i1}^2 + \ldots + l_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = l_{i1}\, l_{k1} + \ldots + l_{im}\, l_{km}$$

2. $Cov(X,F) = L$ atau

$$Cov(X_i, F_j) = l_{ij}$$

Asumsi-Asumsi Model Faktor (2)

- Porsi ragam peubah X ke-i yang dapat dijelaskan oleh m faktor umum disebut dengan komunalitas ke-i sedangkan porsi yang dijelaskan oleh factor spesifik disebut ragam spesifik. Struktur ragam peubah X dapat ditulis sbb:

$$\sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 + \psi_i ;$$

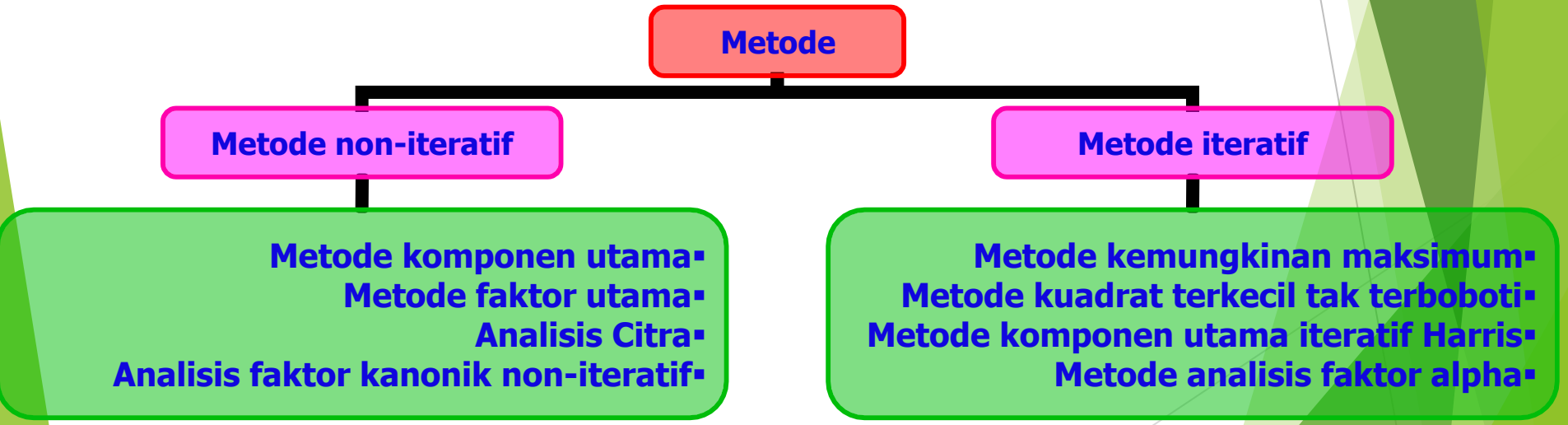
Var (X_i)=komunalitas+ragam spesifik

Atau dapat juga ditulis $\sigma_{ii} = h_i^2 + \psi_i ; i = 1,2, \ldots,p$

dengan $h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2$

Pendugaan Parameter

- Ada beberapa metode pendugaan parameter model factor, yang dapat dikelompokkan dalam metode non-iterati dan metode iterative
- Metode non-iteraif yang paling banyak digunakan adalah metode komponen utama
- Metode iterative yang banyak digunakan adalah metode kemungkinan maksimum



Metode Komponen Utama

- Misal **Σ** merupakan matriks peragam dari matriks pengamatan **X** yang memiliki pasangan nilai akar ciri (eigenvalue) dan vektor cirinya (λ_i , $\underline{e}_i$) dengan $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Misalkan m< p merupakan jumlah peubah dari faktor umum (*common factor*), maka penduga parameter sebagai berikut:
- Matriks penduga faktor loadingnya $\{l_{ij}\}$ yaitu:

  $\acute{L} = [\sqrt{\lambda_1}\,\underline{e}_1 \mid \sqrt{\lambda_2}\,\underline{e}_2 \mid \sqrt{\lambda_3}\,\underline{e}_3 \mid \ldots \mid \sqrt{\lambda_m}\,\underline{e}_m]$
- Penduga ragam spesifik adalah $\Psi = S - \acute{L}\acute{L}'$
- Nilai komunalitas untuk peubah ke-i: $h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2$

Seperti pada Analisis Komponen Utama, dalam analisis factor juga dapat menggunakan matriks korelasi R

Ilustrasi 1

Tersedia data harga saham 100 mingguan (n=100) dari 5 jenis saham (p=5). Dari data 5 jenis saham tersebut ingin factor yang mencirikan kondisi ekonomi. Analisis dilakukan dengan menggunakan matriks korelasi R

Tabel pendugaan loading faktor,komunalitas dan total proporsi keragaman yang dijelaskan dari setiap faktor untuk m=1 dan m=2

Jenis Saham (X)	Solusi satu faktor		Solusi dua faktor		
	F_1	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$	F_1	F_2	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$
Allied Chemical .1	0.783	0.39	0.783	-0.217	0.34
DuPont	0.773	0.40	0.773	-0.458	0.19
Union Carbide .2	0.794	0.37	0.794	-0.234	0.31
Exxon .3	0.713	0.49	0.713	0.472	0.27
Texaco .4	0.712	0.49	0.712	0.524	0.22
Total proporsi kumulatif keragaman yang dapat dijelaskan	0.571		0.571	0.733	

Penjelasan Hasil Analisis

- Jika menggunakan 1 factor maka terdapat 57,1% keragaman X yang dapat dijelaskan Faktor 1, sedangkan jika menggunakan 2 factor sebesar 73,3%
- Faktor pertama merepresentasikan kondisi ekonomi secara umum dan dapat disebut faktor pasar.
- Faktor kedua merupakan kontras antara saham perusahaan kimia dengan saham perusahaan minyak (pada faktor perusahaan kimia memiliki loading negatif yang relatif besar dan perusahaan minyak memiliki loading positif yang relatif besar).
- Dengan demikian faktor kedua dapat disebut faktor industri karena sebagai pembeda harga saham di industri yang berbeda.

Metode Kemungkinan Maksimum

- Metode kemungkinan maksimum (MKM) mengasumsikan bahwa matriks ragam-peragam atau matriks korelasi semua peubah bersifat non-singular.
- Fungsi kepekatan peluang bagi S adalah L(S) yaitu:

$$L(S) = c.|\boldsymbol{\Sigma}|^{-\frac{n-1}{2}}|\mathbf{S}|^{\frac{n-1}{2}-\frac{p+1}{2}} e^{\frac{n-1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S})}$$

dengan *c* adalah konstanta. Sehingga log-likelihood dari **L** dan $\boldsymbol{\psi}$, jika $\boldsymbol{\Sigma} = \mathbf{LL'} + \boldsymbol{\psi}$ adalah:

$$\ln c - \left(\frac{n-1}{2}\right)\left\{tr[(\mathbf{LL'}+\boldsymbol{\psi})^{-1}\mathbf{S}] - ln|\mathbf{LL'}+\boldsymbol{\psi})^{-1}\mathbf{S}|\right\}$$

Penduga kemungkinan maksimum bagi **L** dan $\boldsymbol{\psi}$ diperoleh dengan memaksimumkan diatas dengan kendala $k(k-1)/2$ persyaratan kenunikan (Johnson & Wichern, 1998).

- Penentuan banyaknya faktor bersama

  Uji Nisbah Kemungkinan (likelihood ratio test)

  Hipotesis nol yang diuji pada uji nisbah kemungkinan ini adalah:

  $H_0 : \Sigma = LL' + \psi$, $r(L) = k$ diketahui

  Misalkan $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\psi}}$, dan $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\psi}}$ adalah penduga kemungkinan maksimum bagi $\mathbf{L}$, $\boldsymbol{\psi}$ dan $\boldsymbol{\Sigma}$, jika H_0 benar, maka nilai maksimum untuk log dari fungsi kemungkinannya adalah:

$$\ln L_{H0} = c^* - \left(\frac{n-1}{2}\right)\{tr[\mathbf{S}^{-1}\mathbf{S}] - ln|\mathbf{S}^{-1}\mathbf{S}|\}$$
$$= c^* - \frac{n-1}{2}p$$

Statistik uji nisbah kemungkinan, yaitu:

$$-2\ln\lambda = -2\ln\left(\frac{L_{H_0}}{L}\right)$$

Menyebar khi-kuadrat dengan db=½ $[(p-k)^2-(p+k)]$. Jadi H0 ditolak jika,

$$-2\ln\left(\frac{L_{H_0}}{L}\right) \geq \chi^2_{\alpha;db=\left[(p-k)^2-(p+k)\right]/2}$$

Akaike's information Criterion(AIC)

Statistik AIC untuk model dengan k faktor didefinisikan sebagai berikut:

$$AIC(k)=-2\ln L(k)+[2p(k+1)-k(k-1)]$$

Model berfaktor k dengan k adalah nilai yang berpadanan dengan AIC (k) yang paling kecil dianggap sebagai model terbaik

Data harga saham dianalisa kembali dengan menggunakan metode maksimum likelihood dengan tetap memakai dua model faktor

Variabel		Maksimum likelihood			Komponen utama		
		Penduga faktor		$\tilde{\psi}_i = 1 - \tilde{h}_i^2$	Penduga faktor		$\tilde{\psi}_i = 1 - \tilde{h}_i^2$
Allied Chemical	.1	0.684	0.189	0.50	0.783	-0.217	0.34
DuPont	.2	0.694	0.517	0.25	0.773	-0.458	0.19
Union Karbide	.3	0.681	0.248	0.47	0.794	-0.234	0.31
Exxon	.4	0.621	-0.073	0.61	0.713	0.412	0.27
Texaco	.5	0.792	-0.442	0.18	0.712	0.524	0.22
Total proporsi kumulatif keragaman contoh yang dapat dijelaskan		0.485	0.598		0.571	0.733	

Dalam kasus data tersebut total proporsi kumulatif keragaman dengan metode komponen utama lebih besar dibandingkan dengan maximum likelihood.

Rotasi Faktor

- Dalam banyak kasus, hasil dari analisis factor sulit untuk diintepretasikan makna dari loading setiap factor
- Sebagai salah satu cara untuk membantu memudahkan intepretasi adalah melalui rotasi faktor
- Rotasi factor merupakan transformasi ortogonal dari loading factors dengan :

  L*= LT dimana **TT'=T'T=I**

- Beberapa jenis transformasi yaitu, varimax, oblique, quartimax, dan lain-lain

- **Rotasi Varimax**

Merupakan rotasi yang paling sering dipergunakan pada aplikasi yang merupakan transformasi ortogonal yang diperoleh dengan cara memaksimumkan:

$$\sum_{j=1}^{k}\left\{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{{l^*}_{ij}^{\,2}}{h_i}\right)^2-\left[\frac{1}{p}\sum_{i=1}^{p}\frac{{l_{ij}}^{2}}{h_i}\right]^2\right\}$$

- **Rotasi Oblique**

Digunakan apabila transformasi ortogonal terhadap matriks loading faktor menghasilkan faktor yang masih sulit diinterpretasikan.

- Rotasi quartimax

Transformasi ortogonal dengan tujuan memperoleh Γ yang memaksimumkan

$$\sum_i \sum_j l^*_{ij}{}^4$$

L Adalah matriks loading faktor yang ingin ditransformasi menggunakan matriks ortogonal Γ menjadi $L^* = L\Gamma$

sehingga

$$\frac{1}{Pk}\sum_i \sum_j l^*_{ij}{}^4 - \left(\frac{1}{Pk}\sum_i \sum_j l^*_{ij}{}^2\right) = \frac{1}{Pk}\sum_i \sum_j l^*_{ij}{}^4 - \left(\frac{1}{Pk}\sum_i l^*_{ij}{}^2\right)$$

Mencapai maximum.

Terimakasih

Analisis Faktor

gdito

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- psych
- ggcorrplot
- openxlsx

Silahkan install jika belum ada

```
install.packages("psych")
install.packages("ggcorrplot")
install.packages("openxlsx")
```

Tahapan analisis faktor

1. Ekplorasi data dengan melihat korelasi antar peubah
2. Menentukan banyaknya faktor

Beberapa hal yang dapat dilakukan untuk menentukan banyaknya faktor:

a. Melihat proportion of the sample variance explained atau proporsi keragaman yang bisa jelaskan oleh faktor
b. Scree plot (seperti pada analisis komponen utama)
c. Kemudahan interpretasi hasil analisis faktor

3. Estimasi Faktor Loading
4. Rotasi Faktor (Jika dibutuhkan)
5. Interpretasi Faktor

```
library(psych)
library(ggcorrplot)
```

Data Pelamar Kerja

Seorang HRD ingin mengindentifikasi faktor-faktor yang dapat menjelaskan 12 peubah yang telah dikumpulkan oleh departemen mereka untuk mengukur setiap pelamar kerja. Pegawai HRD menilai pelamar kerja dengan menggunakan skala 1 (rendah) sampai 10(tinggi). Mereka menggumpulakan penilaian untuk 50 pelamar kerja.

Menyiapkan data di R

```
data_jobApp <- openxlsx::read.xlsx("E:/APG/R APG/jobApplicants.xlsx")
head(data_jobApp)
```

```
##   Academic.record Appearance Communication Company.Fit Experience Job.Fit
## 1               6          8             7           5          6       5
## 2               9          8             8           8         10       9
## 3               6          7             7           6          6       7
## 4               7          8             6           5          8       5
```

```
## 5              4          7             8           6          6       6
## 6              7          7             7           5          5       6
##   Letter Likeability Organization Potential Resume Self-Confidence
## 1      7           7            7         6      7               7
## 2      8           9            8         9      9               9
## 3      7           8            8         6      6               8
## 4      9           8            7         8      7               7
## 5      6           7            8         5      4               6
## 6      5           7            8         7      4               6
```

Tahap 1 Ekplorasi data dengan melihat korelasi antar peubah

```
cor_jobApp <- cor(data_jobApp)
ggcorrplot(cor_jobApp,type="lower",lab = TRUE)
```

Jika kita perhatikan hasil korelasi tersebut akan didapati beberapa kelompok peubah yang memiliki korelasi yang besar dalam kelompok peubah tersebut namun korelasinya dengan peubah diluar kelompok tersebut kecil. Sebagai ilustrasi, peubah Potential dengan Academic, Experience, dan Job Fit memiliki korelasi yang besar dibandingkan dengan peubah lainnya. Peubah-peubah dengan korelasi tinggi ini bisa diukur dengan baik oleh suatu peubah latent yang disebut factor.

Tahap 2 Menentukan banyaknya factor

Menggunakan proportion of the sample variance explained

Analisis Faktor di R dapat dilakukan dengan menggunakan fungsi `fa` dari package `psych`. fungsi `fa` memiliki argumen `fm`, menyatakan metode pendugaaan dan juga `rotate`, menyatakan jenis rotasi yang digunakan. Daftar yang bisa diisi dalam argumen `fm` adalah sebagai berikut:

isi_argumen	nama_metode
minres	Komponen Utama
ols	Kuadrat terkecil
wls	Kuadrat terkecil terboboti
gls	Kuadrat terkecil terampat
pa	Metode Faktor Utama
ml	Metode Kemungkinan Maksimum
alpha	Analisis faktor alpha

Sementara itu, argumen `rotate` dapat diisi dengan beberapa metode sebagai berikut:

isi_argumen	nama_metode
none	Tanpa Rotasi
varimax	Varimax
quartimax	Quartimax
equamax	Equamax
promax	Promax/Oblique

Argumen `nfactor` menyatakan banyaknya factornya.

Untuk kasus data pelamar kerja ini, kita akan gunakan metode pendugaan komponen utama. Banyaknya factor yang akan diduga kita set sebanyak peubah yang ada di data pelamar kerja ini.

Note: metode pendugaan komponen utama secara default menggunakan semua peubah yang ada di data asli. Namun, pada fungsi `fa` pengguna harus menentukan terlebih dahulu jumlah factor-nya. Aplikasi lain seperti SAS SPSS dan Minitab tidak perlu menginputkan jumlah factornya

```
fa_jobApp <- fa(data_jobApp,nfactors = 12,fm="minres",rotate="none")
#menampilkan proportion of variance explained
fa_jobApp$Vaccounted
```

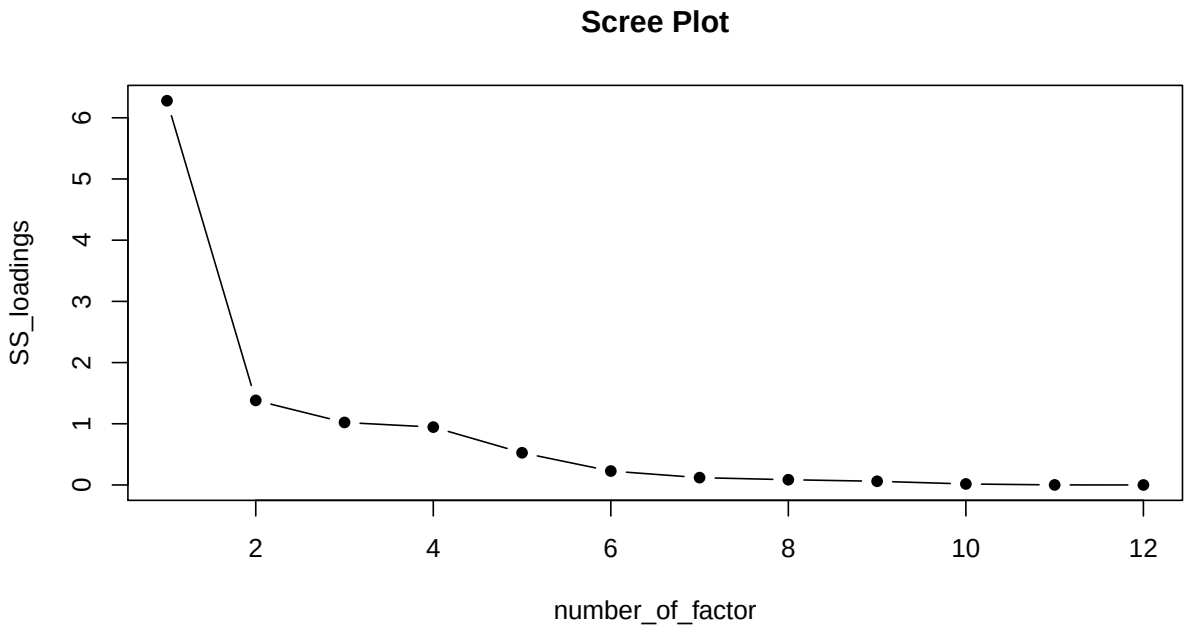
```
##                             MR1       MR2        MR3        MR4        MR5
## SS loadings           6.2782307 1.3820643 1.02167387 0.94565013 0.52627793
## Proportion Var        0.5231859 0.1151720 0.08513949 0.07880418 0.04385649
## Cumulative Var        0.5231859 0.6383579 0.72349741 0.80230159 0.84615808
## Proportion Explained  0.5886699 0.1295874 0.09579589 0.08866762 0.04934575
## Cumulative Proportion 0.5886699 0.7182573 0.81405323 0.90272085 0.95206660
##                              MR6        MR7         MR8         MR9
## SS loadings           0.22747415 0.11983572 0.085960317 0.060410456
## Proportion Var        0.01895618 0.00998631 0.007163360 0.005034205
## Cumulative Var        0.86511426 0.87510057 0.882263932 0.887298136
## Proportion Explained  0.02132881 0.01123624 0.008059954 0.005664306
## Cumulative Proportion 0.97339541 0.98463165 0.992691601 0.998355907
##                             MR10         MR11         MR12
## SS loadings           0.016491825 1.042615e-03 1.200000e-29
## Proportion Var        0.001374319 8.688458e-05 1.000000e-30
## Cumulative Var        0.888672455 8.887593e-01 8.887593e-01
## Proportion Explained  0.001546334 9.775940e-05 1.125164e-30
## Cumulative Proportion 0.999902241 1.000000e+00 1.000000e+00
```

MR1 samapai MR12 merupakan nama dari faktor-faktor yang telah diekstrasksi. Kemudian, SS loadings merupakan keragamaan dari masing-masing faktor.

Jika kita lihat Cumulative Proportion, dengan menggunakan 2 faktor saja sudah bisa menjelaskan lebih dari 71% keragamaan dari data asal. Namun, penambahan Cumulative Proportion dari penggunaan faktor 2 sampai faktor 4 masih cukup besar sehingga masih masuk akal untuk menggunakan 4 faktor.

Menggunakan Scree plot

```
SS_loadings <- fa_jobApp$Vaccounted[1,]
number_of_factor <- seq_along(SS_loadings)
plot(number_of_factor,SS_loadings,type = "b", main = "Scree Plot", pch = 16)
```



Berdasarkan screeplot, pola garis yang berbentuk seperti siku tangan berada pada faktor kedua. Sehingga menurut screeplot cukup 2 factor saja yang kita gunakan.

Karena menurut screeplot dan proportion of variance explained berbeda, kita akan coba keduanya dan melihat faktor mana yang lebih mudah diinterpretasikan.

3. Estimasi faktor loading

Menggunakan 4 faktor

```
fa_jobApp4 <- fa(data_jobApp,nfactors = 4,fm="minres",rotate="none")
print(fa_jobApp4$loadings,cut = 0)
```

```
##
## Loadings:
##                  MR1    MR2    MR3    MR4
## Academic.record  0.698  0.236 -0.304 -0.038
## Appearance       0.692 -0.222 -0.036  0.350
## Communication    0.696 -0.393  0.219 -0.211
## Company.Fit      0.786 -0.074 -0.042 -0.360
## Experience       0.623  0.489 -0.220  0.036
## Job.Fit          0.794  0.045 -0.113 -0.293
## Letter           0.633  0.411  0.599 -0.032
## Likeability      0.713 -0.240 -0.010  0.311
## Organization     0.703 -0.534  0.155 -0.231
## Potential        0.814  0.238 -0.400 -0.108
## Resume           0.708  0.353  0.424  0.216
## Self-Confidence  0.707 -0.251 -0.155  0.451
```

```
##
##                   MR1   MR2   MR3   MR4
## SS loadings     6.152 1.262 0.953 0.798
## Proportion Var  0.513 0.105 0.079 0.067
## Cumulative Var  0.513 0.618 0.697 0.764
```

Jika kita perhatikan, nilai faktor loading yang besar berkumpul di peubah faktor yang pertama, sehingga akan sulit untuk menginterpretasikan faktor 2 sampai faktor 4. Oleh karena itu, kita bisa melakukan rotasi faktor agar lebih mudah menginterpretasikanya.

Note: argumen cut=0 berarti kita menampilkan semua faktor loading, jika kita tidak tentukan isi dari argumen cut, secara default R tidak menampilkan nilai yang berkisar antara $[-0.1, 0.1]$

Menggunakan 2 faktor

```
fa_jobApp2 <- fa(data_jobApp,nfactors = 2,fm="minres",rotate="none")
print(fa_jobApp2$loadings,cut = 0)
```

```
##
## Loadings:
##                   MR1    MR2
## Academic.record  0.703  0.305
## Appearance       0.683 -0.172
## Communication    0.699 -0.434
## Company.Fit      0.778 -0.067
## Experience       0.637  0.578
## Job.Fit          0.792  0.063
## Letter           0.581  0.165
## Likeability      0.707 -0.198
## Organization     0.708 -0.548
## Potential        0.807  0.304
## Resume           0.670  0.187
## Self-Confidence  0.682 -0.143
##
##                   MR1   MR2
## SS loadings     5.993 1.169
## Proportion Var  0.499 0.097
## Cumulative Var  0.499 0.597
```

Untuk dua faktor pun, nilai faktor loading yang besar terkumpul pada faktor pertama sehingga akan sulit diinterpretasikan. Oleh karena itu, kita akan melakukan rotasi faktor untuk penggunaan dua factor ini.

Note: Ukuran besar atau kecil dari faktor loading itu relatif. Beberapa buku menyebutkan nilai diatas 0.6 sudah besar

Tahap 4 Rotasi Faktor

```
fa_jobApp4_rotate <- fa(data_jobApp,nfactors = 4,fm="minres",rotate="varimax")
print(fa_jobApp4_rotate$loadings,cut = 0)
```

```
##
## Loadings:
##                   MR1    MR2    MR4    MR3
## Academic.record  0.713  0.178  0.269  0.154
## Appearance       0.230  0.272  0.704  0.171
## Communication    0.117  0.760  0.309  0.208
## Company.Fit      0.521  0.652  0.151  0.187
```

```
## Experience       0.732 -0.032  0.180  0.327
## Job.Fit          0.614  0.531  0.170  0.206
## Letter           0.223  0.243  0.059  0.904
## Likeability      0.226  0.321  0.688  0.185
## Organization     0.093  0.843  0.360  0.087
## Potential        0.855  0.257  0.285  0.118
## Resume           0.286  0.129  0.331  0.802
## Self-Confidence  0.273  0.210  0.816  0.085
##
##                  MR1   MR2   MR4   MR3
## SS loadings    2.757 2.392 2.211 1.806
## Proportion Var 0.230 0.199 0.184 0.151
## Cumulative Var 0.230 0.429 0.613 0.764
```

Setelah dirotasi menggunakan metode varimax, terlihat bahwa faktor loadings tidak yang bernilai besar tidak lagi bekumpul pada satu faktor saja sehingga memungkinkan diinterpretasikan untuk setiap faktor yang terbentuk.

Fitur lain yang menarik untuk dibahas dalam analisis faktor adalah Communality. communality menjelaskan tentang banyaknya keragamaan yang dapat dijelaskan oleh faktor untuk masing-masing peubah asal. Semakin nilainya mendekati satu semakin baik keragamaan yang dapat dijelaskan.

```
fa_jobApp4_rotate$communalities
```

```
## Academic.record      Appearance   Communication     Company.Fit
##       0.6361963       0.6523914       0.7303190       0.7542383
##      Experience          Job.Fit          Letter     Likeability
##       0.6772272       0.7309871       0.9299102       0.6622484
##    Organization       Potential          Resume Self-Confidence
##       0.8570245       0.8914205       0.8520531       0.7911206
```

Dalam penggunan 4 faktor ini nilai communality untuk semua peubah asal bernilai besar (lebih dari 0.6) sehingga dengan penggunaan 4 faktor sudah tepat

```
fa_jobApp2_rotate <- fa(data_jobApp,nfactors = 2,fm="minres",rotate="varimax")
print(fa_jobApp2_rotate$loadings,cut = 0)
```

```
##
## Loadings:
##                  MR1   MR2
## Academic.record 0.282 0.712
## Appearance      0.605 0.361
## Communication   0.802 0.187
## Company.Fit     0.598 0.503
## Experience      0.042 0.859
## Job.Fit         0.516 0.604
## Letter          0.294 0.527
## Likeability     0.640 0.359
## Organization    0.888 0.112
## Potential       0.356 0.785
## Resume          0.342 0.606
## Self-Confidence 0.584 0.381
##
##                  MR1   MR2
## SS loadings    3.583 3.579
## Proportion Var 0.299 0.298
## Cumulative Var 0.299 0.597
```

Untuk penggunaan dua faktor juga akan nilai faktor loadings yang besar sudah tersebar. Oleh karena itu memungkinkan diinterpretasikan.

```
fa_jobApp2_rotate$communalities
```

```
## Academic.record      Appearance   Communication     Company.Fit
##       0.5867947       0.4963696       0.6774295       0.6104035
##      Experience         Job.Fit          Letter     Likeability
##       0.7404683       0.6313752       0.3644644       0.5383605
##    Organization       Potential          Resume Self-Confidence
##       0.8015258       0.7437662       0.4843980       0.4859770
```

Dalam penggunan 2 faktor ini nilai communality untuk terdapat peubah asal bernilai keci (kurang dari 0.4) sehingga dengan penggunaan 2 faktor kurang tepat

Dalam hal ini lebih baik menggunakan 4 faktor saja, namun dalam ilustrasi ini 2 faktor akan tetap digunakan untuk memperlihatkan perbedaan interpretasi dalam pemilihan banyaknya faktor digunakan

Note: Communalities bisa diperiksa saat sebelum kita rotasi

Tahap 5 Interpretasi Faktor

Menggunakan 4 faktor Untuk mempermudah interpretasi, faktor loading yang ditampilkan selain $[-0.6, 0.6]$.

```
print(fa_jobApp4_rotate$loadings,cut = 0.6)
```

```
##
## Loadings:
##                  MR1    MR2    MR4    MR3
## Academic.record  0.713
## Appearance                      0.704
## Communication           0.760
## Company.Fit             0.652
## Experience       0.732
## Job.Fit          0.614
## Letter                                 0.904
## Likeability                     0.688
## Organization            0.843
## Potential        0.855
## Resume                                 0.802
## Self-Confidence                 0.816
##
##                  MR1   MR2   MR4   MR3
## SS loadings    2.757 2.392 2.211 1.806
## Proportion Var 0.230 0.199 0.184 0.151
## Cumulative Var 0.230 0.429 0.613 0.764
```

Pada penggunaan 4 faktor faktor laoding hasil rotasi dapat diinterpretasikan sebagai berikut:

a. Accademic Records, Experience, Job Fit, dan Potential memiliki nilai faktor loading yang besar dan positif untuk faktor 1, kita bisa menyebut faktor 1 sebagai ketepatan penempatan dan potensi berkembang bagi pegawai dalam perusahaan
b. Communication, Company Fit dan Organization memiliki nilai faktor loading yang besar dan positif untuk faktor 2 sehingga kita bisa menyebut faktor 2 dengan kemampuan dalam bekerja (work skills) c.Letter dan Resume memiliki nilai faktor loading yang besar dan positif untuk faktor 3 sehingga kita bisa menyebut faktor 3 dengan kemampuan menulis

c. Appearance ,Likeability and Self Confidence memiliki nilai faktor loading yang besar dan positif untuk faktor 4 sehingga kita bisa menyebut faktor 4 dengan kualitas personal pegawai

Menggunakan 4 faktor

```
print(fa_jobApp2_rotate$loadings,cut = 0.6)
```

```
##
## Loadings:
##                 MR1   MR2
## Academic.record       0.712
## Appearance      0.605
## Communication   0.802
## Company.Fit
## Experience            0.859
## Job.Fit               0.604
## Letter
## Likeability     0.640
## Organization    0.888
## Potential             0.785
## Resume                0.606
## Self-Confidence
##
##                  MR1   MR2
## SS loadings    3.583 3.579
## Proportion Var 0.299 0.298
## Cumulative Var 0.299 0.597
```

Pada penggunaan 2 faktor faktor laoding hasil rotasi dapat diinterpretasikan sebagai berikut: a. Appearance, Communication,Likeability, dan Organization memiliki nilai faktor loading yang besar dan positif untuk faktor 1, kita bisa menyebut faktor 1 sebagai soft-skill dari pegawai b. Accademic Record, Experience, Job Fit, Potential dan Resume memiliki nilai faktor loading yang besar dan positif untuk faktor 1, kita bisa menyebut faktor 1 sebagai hard-skill dari pegawai

Berdasarkan penggunaan 4 faktor dan 2 faktor terdapat interpretasi diperoleh interpretasi yang berbeda dari faktor-faktor yang dihasilkan. Sehingga penggunaan 4 faktor atau dua faktor bisa disesuaikan dengan kebutuhan perusahaan tersehut.

Berikut adalah contoh penerapan analisis faktor menggunakan R. Jika kalian mencoba metode pendugaan yang berbeda dan rotasi berbeda maupun jumlah faktor berbeda kalian akan menemukan interpretasi yang berbeda pula. Tidak ada benar dan salah dalam interpretasi faktor pada faktor analisis

==========Selamat Mencoba==========

ANALISIS GEROMBOL (CLUSTER ANALYSIS)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Dalam beberapa hal, tujuan suatu penelitian adalah untuk mengelompokkan individu-individu berdasarkan banyak peubah penciri
- Individu-individu dalam satu kelompok memiliki kemiripan atau keragaman yang kecil dibandingkan individu-individu di kelompok yang berbeda
- Analisis gerombol merupakan metode yang dapat menggabungkan beberapa individu ke dalam kelompok-kelompok berdasarkan sifat kemiripan atau sifat ketidakmiripan antar objek, sehingga objek dalam kelompok lebih mirip dibandingkan dengan objek antar kelompok
- Kemiripan/ketakmiripan antar objek dalam analisis gerombol menggunakan konsep jarak
- Sebagai contoh ingin menggerombolkan 34 provinsi di Indonesia berdasarkan indicator kesejahteraan rakyat
- Terdapat beberapa metode penggerombolan

Konsep Jarak Antar Objek

- Objek yang berada pada gerombol yang sama memiliki kemiripan yang lebih besar dibandingkan objek yg ada dalam gerombol lainnya.
- Kemiripan antar objek diukur dengan konsep jarak
- Jarak antar 2 objek *a* dan *b,* dinotasikan dengan d*(a,b),* dimana :

- $d(a, b) \geq 0$
- $d(a, a) = 0$
- $d(a, b) = d(b, a)$
- d(*a*, *b*) meningkat seiring semakin tidak mirip kedua objek a dan b
- $d(a,c) \leq d(a,b) + d(b,c)$

Asumsi : semua pengukuran bersifat numerik

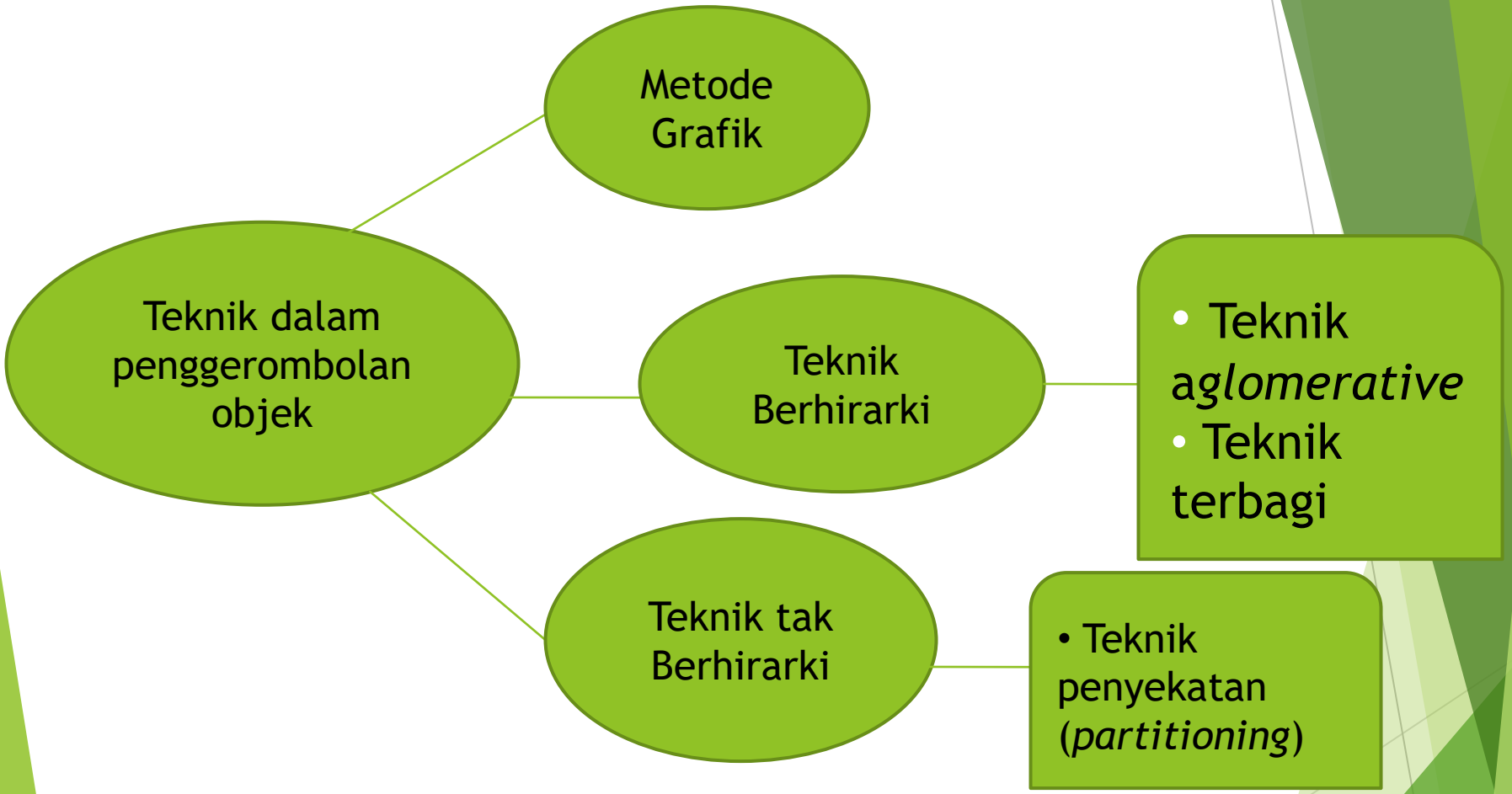
Struktur Data Amatan

Individu	Peubah				
	X1	X2	X3	...	Xp
1	x11	x12	x13		x1p
2	x21	x22	x23		x2p
3	x31	x32	x33		x3p
4	x41	x42	x43		x4p
5	x51	x52	x53		x5p
...	...	...	...	...	...
...	...	...	...	...	...
n	xn1	xn2	xn3		xnp

Beberapa konsep jarak

Jarak	Formula
Jarak Euclidean	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$ $= \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Jarak Minkowski / Jarak city-block / Jarak Manhattan	$d(x, \mathbf{y}) = \left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert^k\right]^{\frac{1}{k}}$
Jarak Mahalanobis	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$

Beberapa Teknik Penggerombolan



Metode grafik sangat subjetif untuk menarik kesimpulan

Perbedaan antara teknik berhirarki dengan teknik tak berhirarki

Teknik berhirarki

- banyaknya gerombol yang akan dihasilkan belum diketahui
- hasil penggerombolan ditampilkan dalam bentuk dendrogram

Teknik tak berhirarki

- banyaknya gerombol sudah ditentukan dulu
- Beberapa metode: K-rataan Macqueen, metode Chernoff dan kurva Andrews

Metode berhirarkhi lebih popular digunakan

Beberapa metode penggerombolan berhirarkhi:

- Pautan Tunggal
- Pautan Lengkap
- Pautan Centroid
- Pautan Median
- Pautan Rataan

Pautan Tunggal (Single Linkage)

Jarak antar dua gerombol diukur dengan jarak terdekat antara sebuah objek dalam gerombol yang satu dengan sebuah objek dalam gerombol yang lain.

$$h(Br, Bs) = \min \{ d(\mathbf{x}i, \mathbf{x}j); \mathbf{x}i \text{ anggota } Br, \text{ dan } \mathbf{x}j \text{ anggota } Bs \}$$

- **Pautan Lengkap (Complete Linkage)**

  Jarak antar dua gerombol diukur dengan jarak terjauh antara sebuah objek dalam gerombol yang satu dengan sebuah objek dalam gerombol yang lain.

$h(Br, Bs) = \max \{ d(\mathbf{x}i, \mathbf{x}j); \mathbf{x}i \text{ anggota } Br, \text{ dan } \mathbf{x}j \text{ anggota } Bs \}$

Pautan Centroid (Centroid Linkage)

Jarak antara dua buah gerombol diukur sebagai jarak Euclidean antara kedua rataan (centroid) gerombol.

Jika $\bar{\mathbf{x}}_r$ dan $\bar{\mathbf{x}}_s$ adalah vektor rataan (centroid) dari gerombol *Br* dan *Bs*, maka jarak kedua gerombol tersebut didefinisikan sebagai :

$$h(B_r, B_s) = d(\bar{\mathbf{x}}_r, \bar{\mathbf{x}}_s)$$

Centroid cluster yang baru didefinisikan sebagai :

$$\frac{n_r \bar{\mathbf{x}}_r + n_s \bar{\mathbf{x}}_s}{n_r + n_s}$$

Pautan Median (Median Linkage)

Jarak antar gerombol didefinisikan sebagai jarak antar median, dan gerombol-gerombol dengan jarak terkecil akan digabungkan.

Median untuk gerombol yang baru adalah

$$\text{Mbaru} = \frac{\mathbf{m}_r + \mathbf{m}_s}{2}$$

Pautan Rataan (Average Linkage)

Jarak antara dua buah gerombol, *Br* dan *Bs* didefinisikan sebagai rataan dari $n_r n_s$ jarak yang dihitung antara $\mathbf{x}_i$ anggota *Br* dan $\mathbf{x}_j$ anggota *Bs*

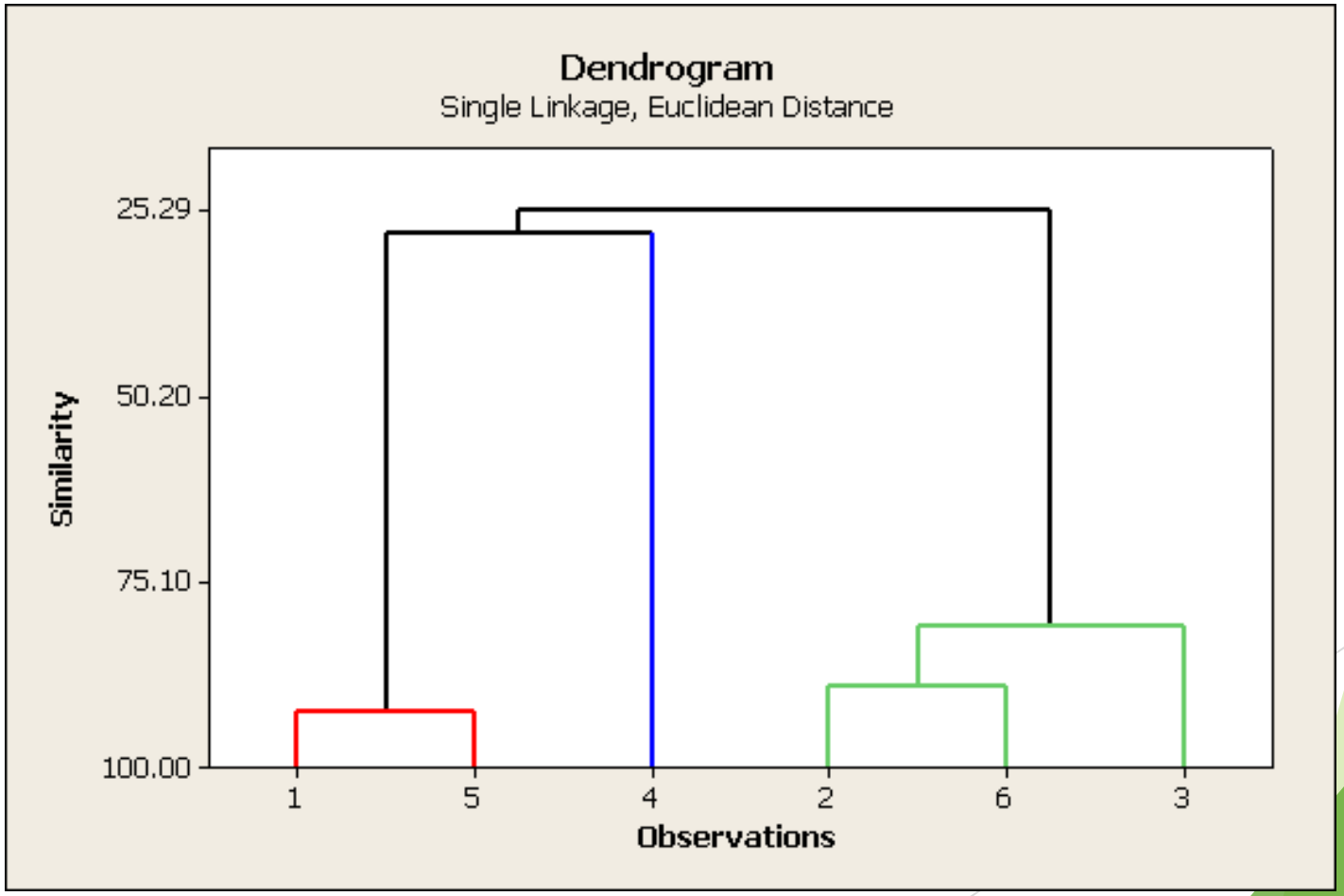
$$h(B_r, B_s) = \frac{1}{n_r n_s} \sum_{\mathbf{x}_i \in B_r} \sum_{\mathbf{x}_j \in B_s} d(\mathbf{x}_i, \mathbf{x}_j)$$

Berikut adalah data nilai 7 mata pelajaran dari 6 siswa. Berdasarkan data tersebut ingin diketahui kemiripan prestasi 6 siswa tersebut

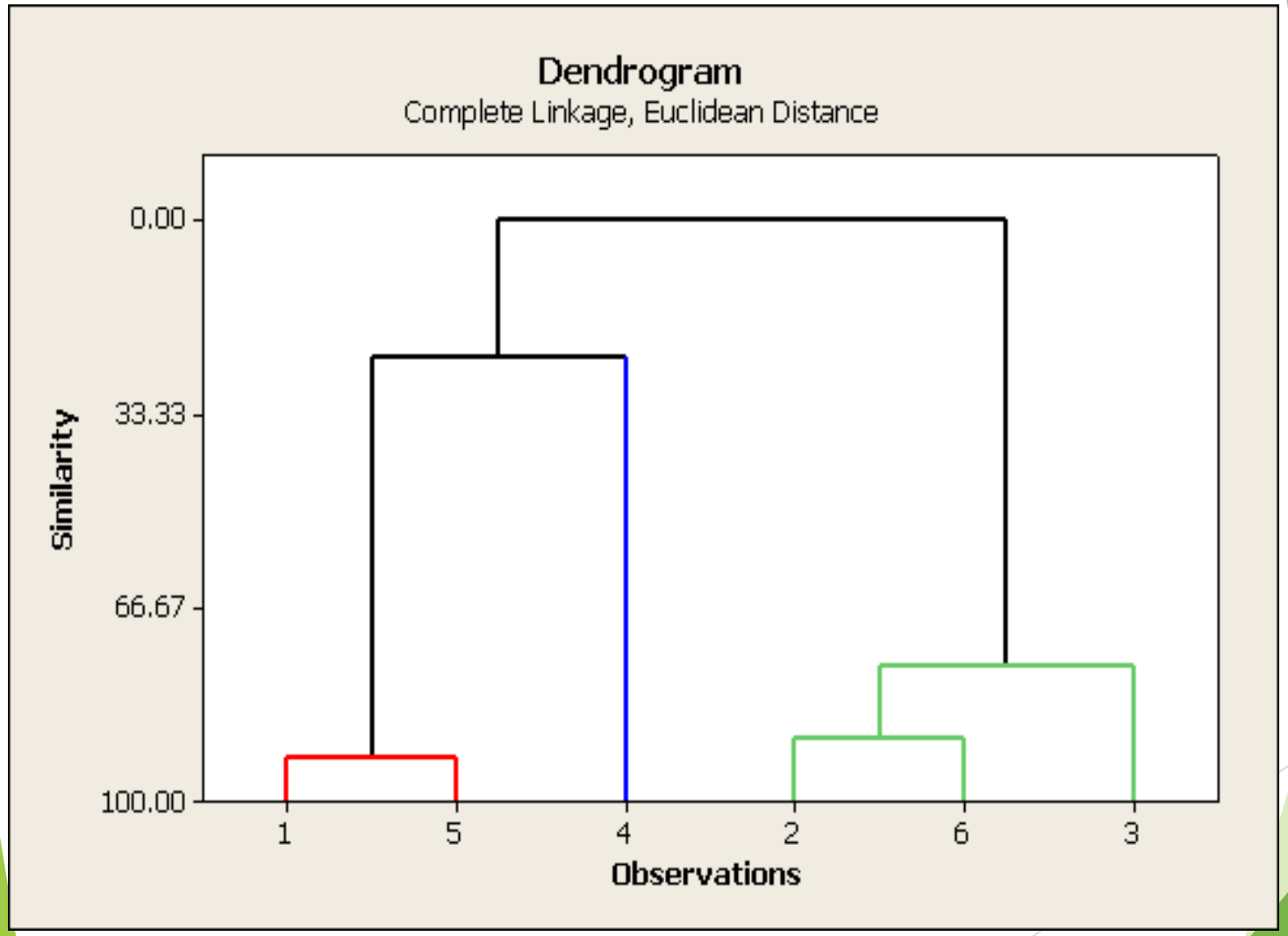
	Mat	Fis	Bio	Sej	Kew	Sos	Seni
1.Andi	8.1	8.3	7.6	6.2	5.8	5.4	6.0
2. Benny	5.6	6.3	6.1	7.3	7.4	7.6	6.0
3.Budi	5.2	5.8	5.7	7.0	6.8	7.2	5.7
4. Ika	6.7	6.8	5.6	7.4	5.3	5.4	7.9
5. Maya	8.2	8.2	7.4	6.4	5.7	5.5	6.1
6. Ana	5.7	6.4	5.9	7.1	7.2	7.3	5.8

Perbandingan hasil dendogram kelima metode penggerombolan (menghasilkan hasil yang berbeda)

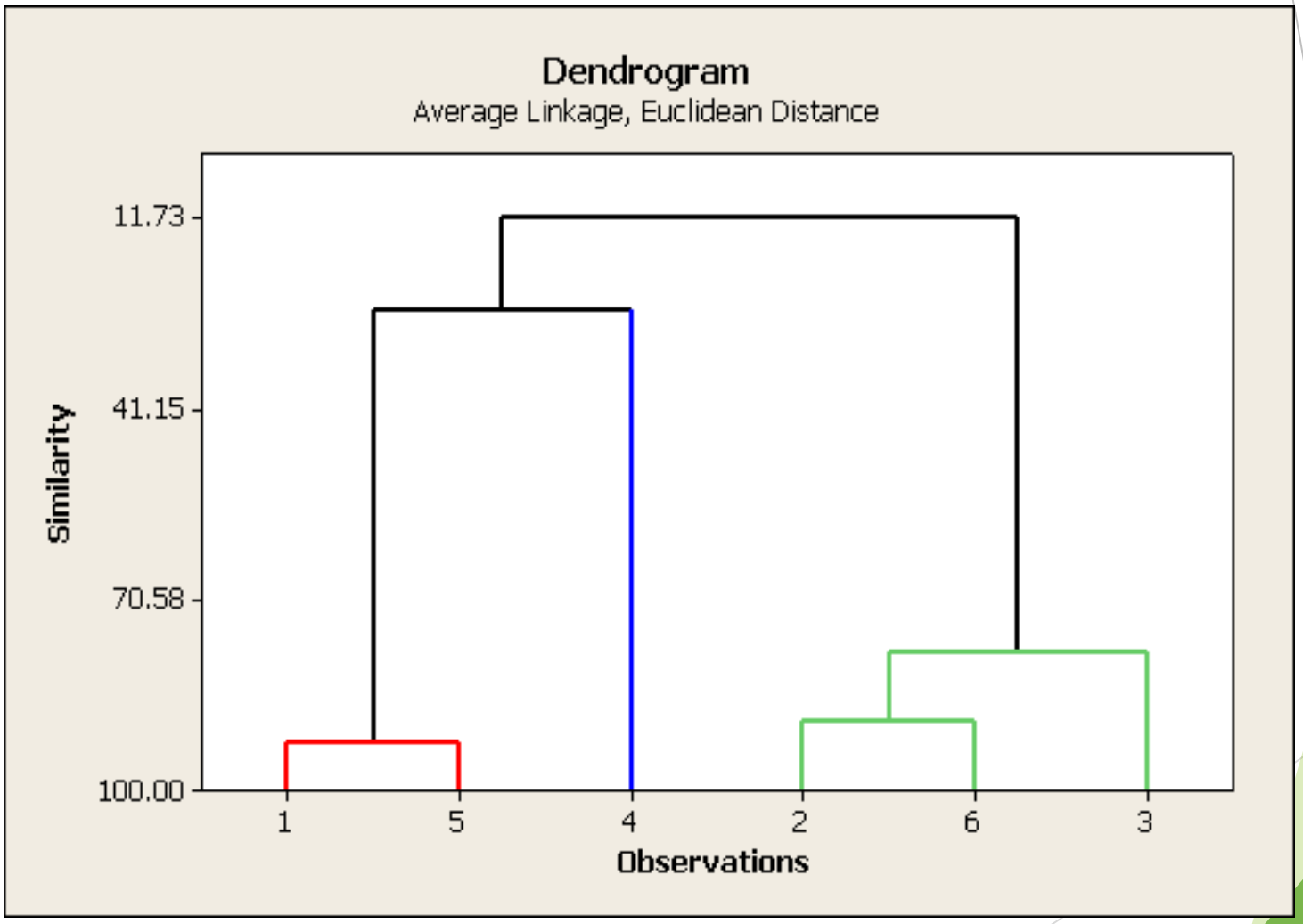
Single Linkage



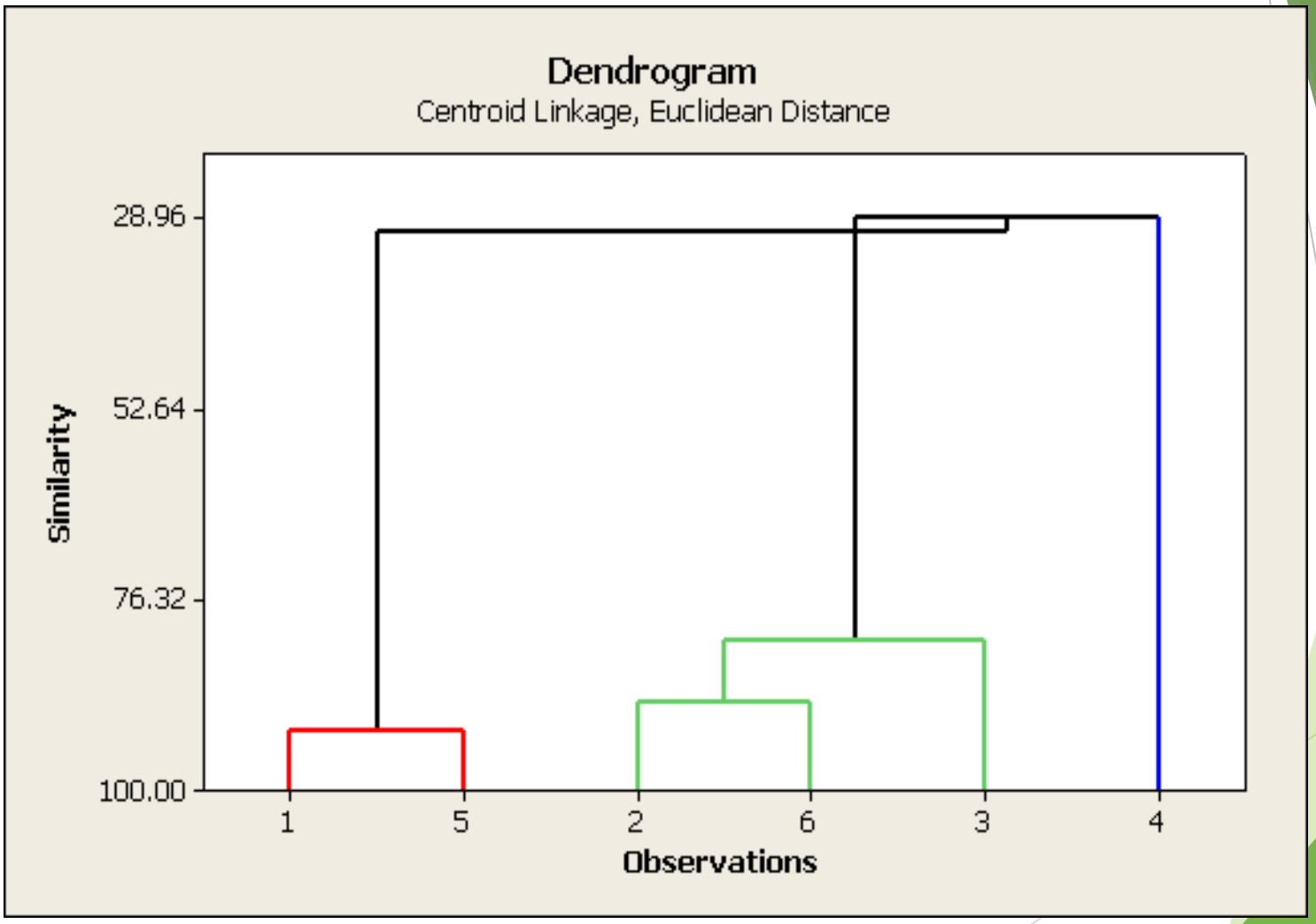
Complete Linkage



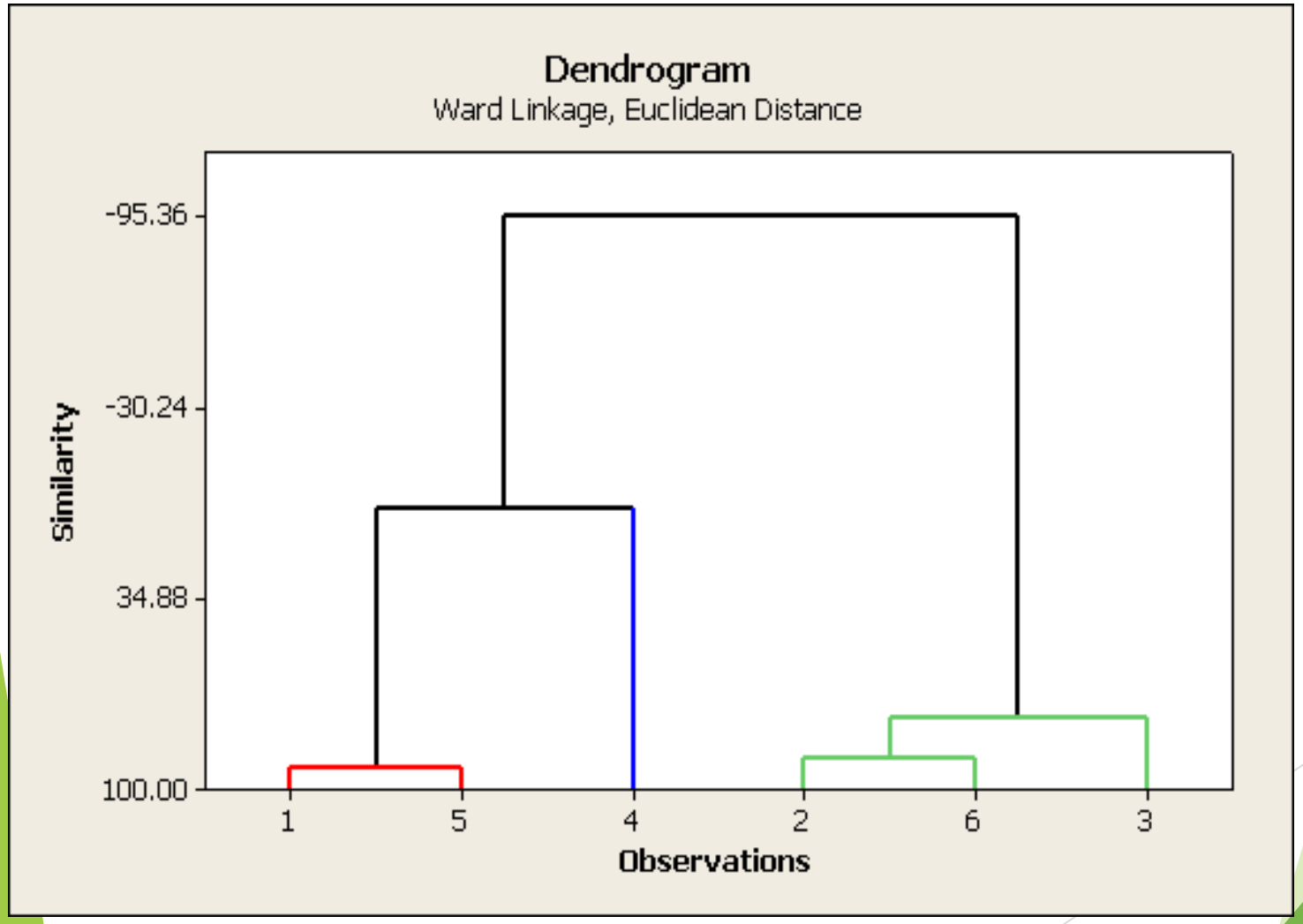
Average Linkage



Centroit Linkage



Ward Linkage



Metode Penggerombolan tak berhirarki

- Metode K rataan (*k-means*)

Algoritmanya sbb :

1. Tentukan besarnya *k*, yaitu banyaknya gerombol, dan tentukan juga centroid di tiap gerombol.
2. Hitung jarak antara setiap objek dengan setiap centroid.
3. Hitung kembali rataan (centroid) untuk gerombol yang baru terbentuk.
4. Ulangi langkah 2 sampai tidak ada lagi pemindahan objek antar gerombol.

Terimakasih

Analisis Gerombol

gdito

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- factoextra

Silahkan install jika belum ada

```
install.packages("factoextra")
```

```
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Metode K-means

Prosedur penerapan K-means

1. Pre-processing data
2. Memilih banyaknya gerombol
3. Menerapkan K-means
4. Interprestasi Gerombol yang terbentuk

Data Pelanggan Mall

Seorang pemilik Mall ingin mengelompokan customer di Mall yang ia miliki, sehingga tim marketing bisa mengembangkan strategi yang tepat untuk customer yang tepat pula. Data yang dimiliki oleh Mall tersebut adalah Customer ID, umur pelanggan (age), pendapatan tahunan dalam ribu dollar (annual income) dan spending score. Spending score merupakan nilai yang diberikan oleh Mall kepada customer berbasarkan perilaku customer (waktu kunjungan,jenis barang yang dibeli, dan banyaknya uang yang dihabiskan dalam belanja) yang memiliki rentang nilai 1-100. Semakin besar nilai Spending Score berarti customer semakin loyal pada Mall tersebut dan semakin besar pula uang belanja yang digunakan.

Menyiapkan data di R

```
data_mall <- read.csv("Mall_Customers.csv")
head(data_mall)
```

```
##   CustomerID  Genre Age Annual.Income Spending.Score
## 1          1   Male  19            15             39
## 2          2   Male  21            15             81
## 3          3 Female  20            16              6
## 4          4 Female  23            16             77
## 5          5 Female  31            17             40
## 6          6 Female  22            17             76
```

Pre-processing data

Peubah yang digunakan untuk menerapkan k-means adalah peubah Age AnnualIncome dan Spending Score. Oleh karena itu peubah yang tidak kita gunakan akan kita hilangkan tersebih dahulu.

```
data_mall <- data_mall[,c("Age","Annual.Income","Spending.Score")]
head(data_mall)
```

```
##   Age Annual.Income Spending.Score
## 1  19            15             39
## 2  21            15             81
## 3  20            16              6
## 4  23            16             77
## 5  31            17             40
## 6  22            17             76
```

Standarisasi peubah

Standarisasi peubah merupakan proses transformasi peubah menjadi peubah yang memiliki rata-rata nol dan simpangan baku satu. Process standarisasi ini dilakukan jika kita melihat perbedanan satuan pengukuran peubah-peubah yang digunakan contoh(umur dan pendapatan). Standarisasi dilakukan karena metode k-means menggunakan konsep jarak antara objek/amatan, yang mana sensitif terhadap satuan pengukuran. Formula untuk standarisasi data adalah sebagai berikut:

$$y = \frac{y - \bar{y}}{\sigma_y}$$

dengan $\bar{y}$ merupakan rata-rata dari y dan σ_y merupakan simpangan baku dari y.

Dalam R, standarisasi data bisa dilakukan dengan menggunakan fungsi `scale`.

```
data_mall_standardize <- scale(data_mall)
apply(data_mall_standardize,2,mean)
```

```
##            Age  Annual.Income Spending.Score
##  -1.016906e-16  -8.144310e-17  -1.096708e-16
```

```
apply(data_mall_standardize,2,sd)
```

```
##            Age  Annual.Income Spending.Score
##              1              1              1
```

Jika kita perhatikan rata-rata dan simpangan baku peubah setelah distandarisasi mendekati nol dan satu.

Note: Dalam tahapan pre-processing data, kita menyiapkan data agar metode kmeans bisa diterapkan secara maksimal. Dua hal yang umumnya dilakukan pada tahap ini adalah memilih peubah yang digunakan dan melakukan standarisasi peubah.

Memilih banyaknya gerombol

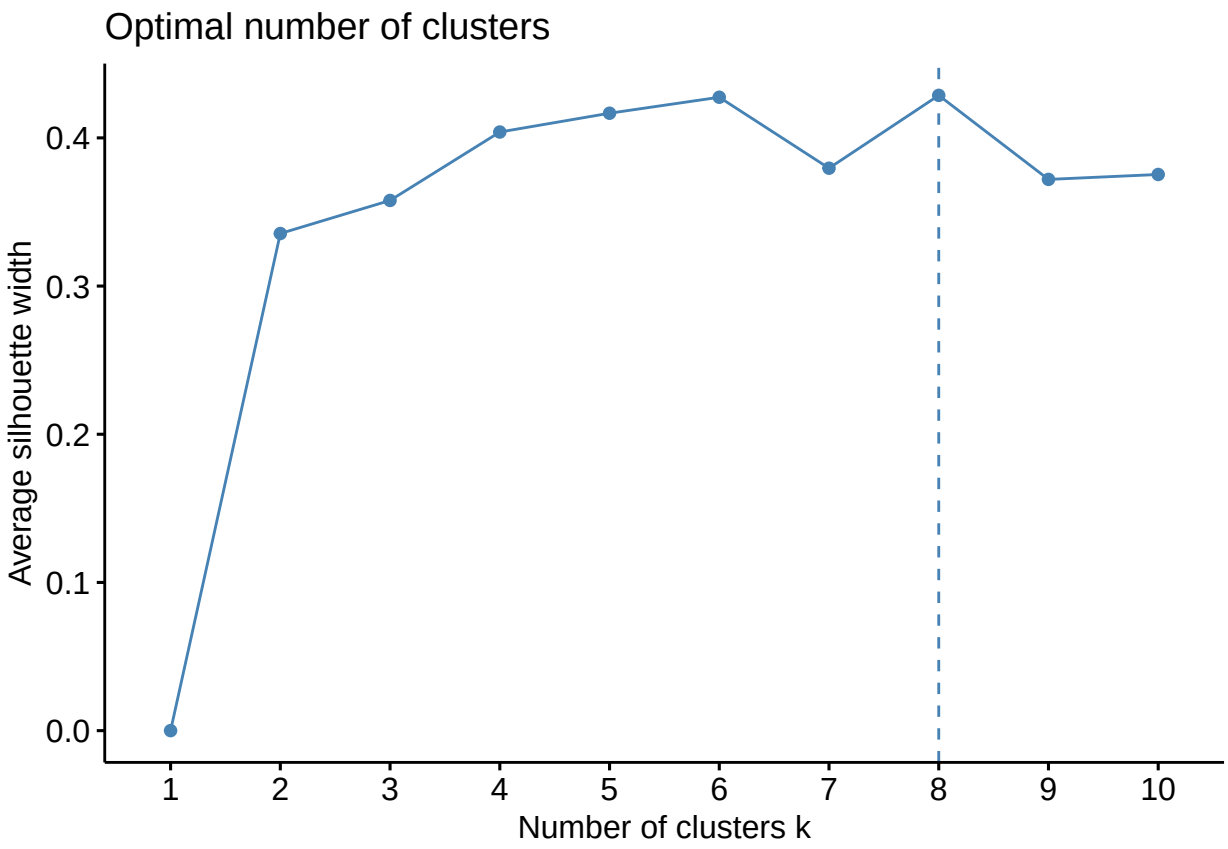
Umumnya, banyaknya gerombol dapat ditentukan dengan menggunakan beberapa kriteria statistik, seperti koefisien **silhouette** dan **WSS** atau (Within Sum of Square).

Kriteria koefisien silhouette dihitung berdasarkan jarak antar amatan. Koefisien ini mengukur seberapa dekat suatu amatan dengan amatan lain yang berada di gerombol yang sama (dikenal sebagai ukuran cohesion) dibandingkan dengan jarak terhadap amatan lain yang berada di gerombol berbeda (dikenal sebagai ukuran separation). Koefisien yang nilainya semakin besar menunjukkan bahwa gerombol yang terbentuk sudah sesuai.

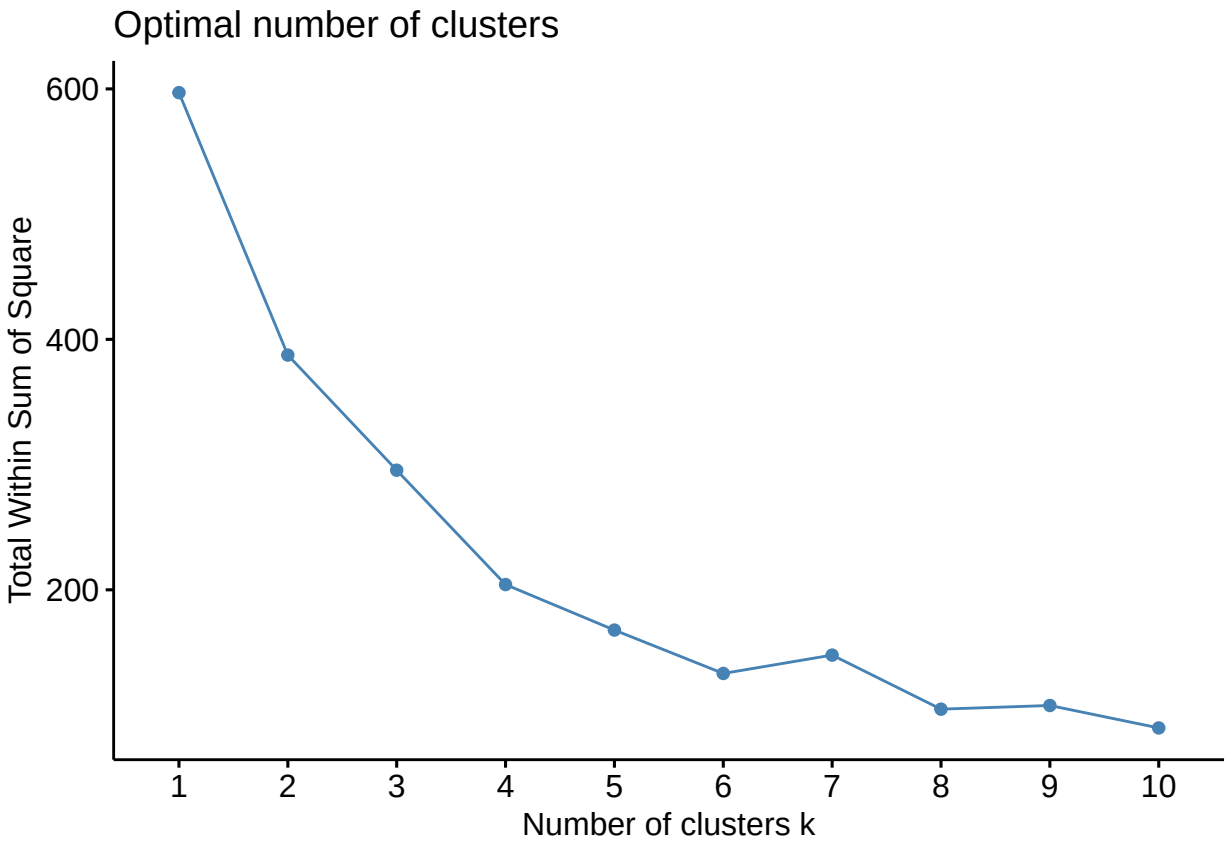
Kriteria WSS merupakan kriteria yangmenghitung keragamaan dalam gerombol yang terbentuk. Semakin kecil keragaman dalam gerombol yang terbentuk menunjukkan bahwa gerombol yang terbentuk sudah sesuai.

Dengan menggunakan kriteria tersebut, kita bisa membandingkan banyaknya gerombol yang paling sesuai pada data yang kita sedang analisis. Dalam R, fungsi `fviz_nbclust` dari package `factoextra` dapat digunakan untuk memilih banyaknya gerombol.

```
fviz_nbclust(data_mall_standardize,FUNcluster = kmeans,method = "silhouette")
```




```
fviz_nbclust(data_mall_standardize,FUNcluster = kmeans,method = "wss")
```



Untuk kriteria koefisien silhoutte, banyaknya gerombol dengan nilai koefisien tertinggi yang kita pilih. Sedangkan pada WSS, banyaknya gerombol yang kita pilih didasarkan pada banyaknya gerombol yang mana garisnya berbentuk seperti siku (elbow). Pada gambar diatas garis membentuk siku saat berada di gerombol keempat. **Karena penentuan ini berdasarkan visual,jadi setiap orang mungkin berbeda melihat pola sikunya**

Berdasarkan kedua kriteria tersebut, banyaknya gerombol terbaik yang dipilih berbeda. Jika demikian, banyaknya gerombol bisa ditentukan berdasarkan kemudahan interpretasi gerombol yang terbentuk. Pada tulisan ini kita akan menggunakan 4 gerombol saja.

Note: secara default banyaknya gerombol yang dicobakan pada fungsi `fviz_nbclust` adalah 10, jika ingin merubah hal tersebut bisa dilakukan dengan menggunakan argumen `kmax` dalam fungsi,misal `kmax=20`.

Menerapkan K-means

Setelah kita mendapatkan banyaknya gerombol terbaik, maka selajutnya kita akan menerapkan metode kmenas untuk mendapatkan label gerombol pada setiap amatan. Fungsi `eclust` dari package `factoextra` digunakan untuk menerpkan metode kmeans. Pada fungsi `eclust`, kita cukup memasukan data yang sebelum distandarisasi, karena dalam fungsi tersebut terdapat argumen `stand`, yang jika diatur `stand=TRUE` secara otomatis data yang kita gunakan akan distandarisasi.

```
kmeans_mall <- eclust(data_mall,stand = TRUE,FUNcluster = "kmeans",k=4,graph = F)
kmeans_mall$cluster
```

```
##   [1] 3 3 3 3 3 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 3 2 3 2 3 2 3 2 3 2 3 2 3 2
##  [36] 3 2 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 2 2 2 2 2 3 2 2 3 2 2 2 3 2 2 3 3
##  [71] 2 2 2 2 2 3 2 2 3 2 2 3 2 2 3 2 2 3 3 2 2 3 2 2 3 3 2 3 2 3 3 2 2 3 2
## [106] 3 2 2 2 2 2 3 1 3 3 3 2 2 2 2 3 1 4 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4
```

```
## [141] 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4 1 4 1
## [176] 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4
```

```
kmeans_mall$centers
```

```
##           Age Annual.Income Spending.Score
## 1  0.03711223     0.9876366     -1.1857814
## 2  1.08344244    -0.4893373     -0.3961802
## 3 -0.96008279    -0.7827991      0.3910484
## 4 -0.42773261     0.9724070      1.2130414
```

Label gerombol untuk setiap amatan/objek, bisa diperoleh dengan menggunakan `$cluster`. Kemudian,interpretasi setiap gerombol yang terbentuk dapat dilakukan dengan menggunakan bantuan nilai rata-rata dari masing-masing peubah dihitung berdasarkan gerombol. Informasi ini bisa diperoleh dengan menggunakan `$centers`. Karena kita melakukan standarisasi peubah, maka nilai rata-rata yang diperoleh juga dalam skala standarisasi.

4. Interprestasi Gerombol yang terbentuk

Berdasarkan nilai rata-rata dari `$centers`, berikut adalah interpretasinya

- Gerombol 1 : gerombol ini merupakan customer-customer yang cukup muda (peubah age bernilai kecil) dan berpenghasilan besar (peubah Income bernilai besar) namun sedikit sekali menghabiskan uangnya untuk berbelanja (peubah spending score bernilai kecil bahkan negatif).
- Gerombol 2 : gerombol ini merupakan customer-customer yang sudah tua (peubah age bernilai besar) dan berpenghasilan kecil (peubah Income bernilai kecil) dan sedikit sekali menghabiskan uangnya untuk berbelanja (peubah spending score bernilai kecil). Gerombol ini mungkin murupakan customer yang sudah pensiun dan hanya memiliki pemasukan dari tunjangan pensiun.
- Gerombol 3 : gerombol ini merupakan customer-customer yang masih sangat muda (peubah age bernilai kecil) dan berpenghasilan kecil (peubah Income bernilai kecil) namun menghabiskan uangnya untuk berbelanja cukup besar (peubah spending score bernilai besar). Gerombol ini mungkin murupakan customer yang aneh, karena memiliki penghasilan yang kecil namun belanjanya banyak.
- Gerombol 4 : gerombol ini merupakan customer-customer yang masih cukup muda (peubah age bernilai kecil) dan berpenghasilan besar (peubah Income bernilai besar) namun menghabiskan uangnya untuk berbelanja cukup besar (peubah spending score bernilai besar). Gerombol ini mungkin murupakan customer yang paling menarik untuk menjadi target marketing selanjutnya.

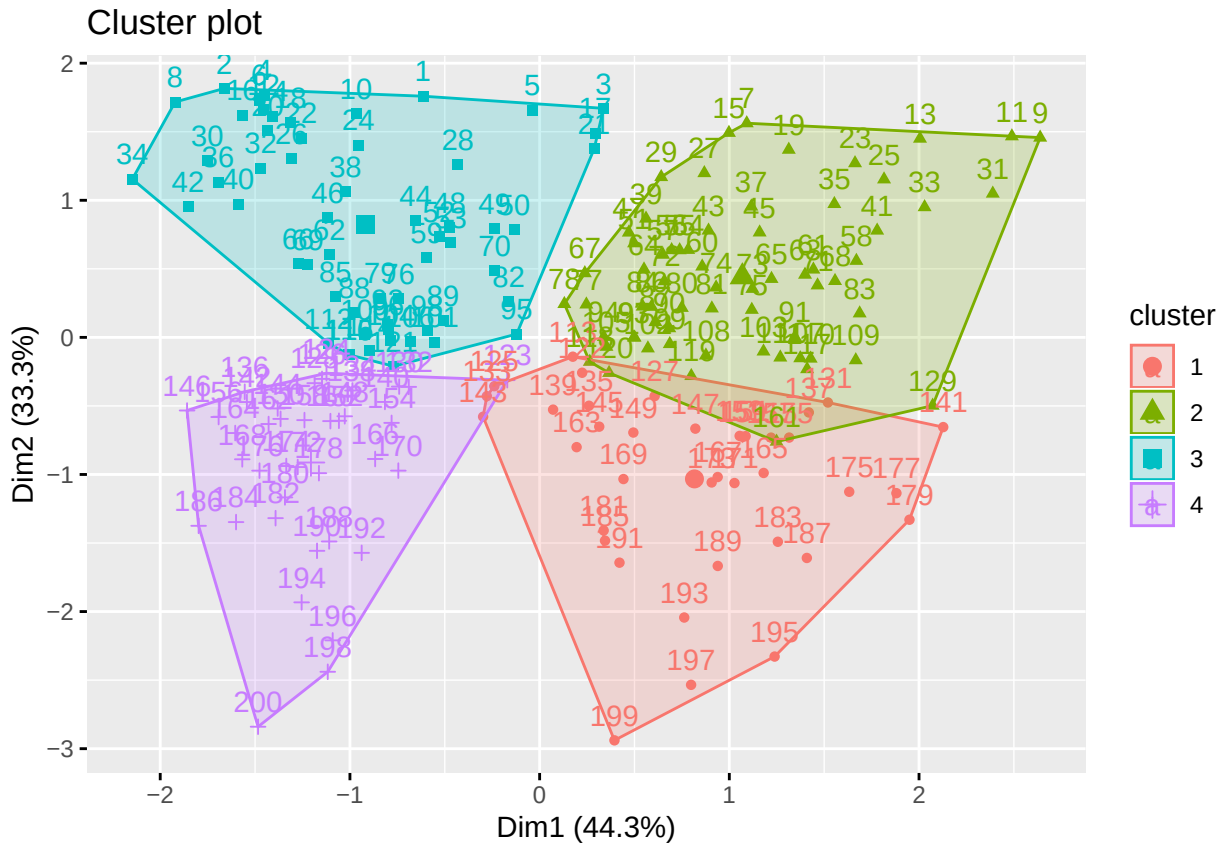
Jika sulit membaca hasil dalam bentuk skala standarisasi maka kita bisa menggunakan fungsi `aggregate` untuk melihat rata-ratanya dalam skala aslinya. Fungsi ini dapat menghitung rata-rata setiap peubah berdasarkan gerombol yang terbentuk.

```
aggregate(data_mall,by =list(gerombol=kmeans_mall$cluster),
            FUN = mean)
```

```
##   gerombol      Age Annual.Income Spending.Score
## 1        1 39.36842      86.50000       19.57895
## 2        2 53.98462      47.70769       39.96923
## 3        3 25.43860      40.00000       60.29825
## 4        4 32.87500      86.10000       81.52500
```

Cara lain untuk memnginterpretasikan hasil gerombol adalah menggunakan scatterplot. Jika peubah untuk membangun kluster lebih dari dua, maka sebelum dibentuk scatterplot peubah tersebut direduksi terlebih dahulu menggunakan analisis komponen utama menjadi dua komponen utama. Namun, untuk interpretasinya setiap gerombolnya kita harus mengetahui interpretasi dari kedua komponen utama dan belum tentu dengan dua komponen utama tersebut sudah mampu menjelaskan keragaman data asal dengan baik.

```
fviz_cluster(kmeans_mall)
```



Interpretasi dua komponen utama bisa dilihat dengan akar cirinya.

```
pca_mall <- prcomp(data_mall_standardize)
pca_mall$rotation
```

```
##                       PC1         PC2         PC3
## Age             0.70638235 -0.03014116 0.707188441
## Annual.Income  -0.04802398 -0.99883160 0.005397916
## Spending.Score -0.70619946  0.03777499 0.707004506
```

Metode Hierarchical Clustering

Prosedur Hierarchical Clustering

1. Pre-processing data
2. Memilih metode linkage dan banyaknya gerombol
3. Menerapkan Hierarchical Clustering
4. Interprestasi Gerombol yang terbentuk

Data yang digunakan untuk ilustrasi Hierarchical Clustering sama seperti Kmeans diatas, yaitu menggunakan data pelanggan Mall

Menyiapkan data di R

```
data_mall <- read.csv("Mall_Customers.csv")
head(data_mall)
```

```
##   CustomerID  Genre Age Annual.Income Spending.Score
## 1          1   Male  19            15             39
## 2          2   Male  21            15             81
## 3          3 Female  20            16              6
## 4          4 Female  23            16             77
## 5          5 Female  31            17             40
## 6          6 Female  22            17             76
```

Pre-processing data

Memilih peubah yang digunakan untuk analisis

```
data_mall <- data_mall[,c("Age","Annual.Income","Spending.Score")]
head(data_mall)
```

```
##   Age Annual.Income Spending.Score
## 1  19            15             39
## 2  21            15             81
## 3  20            16              6
## 4  23            16             77
## 5  31            17             40
## 6  22            17             76
```

Standarisasi peubah

```
data_mall_standardize <- scale(data_mall)
```

Memilih metode linkage dan banyaknya gerombol

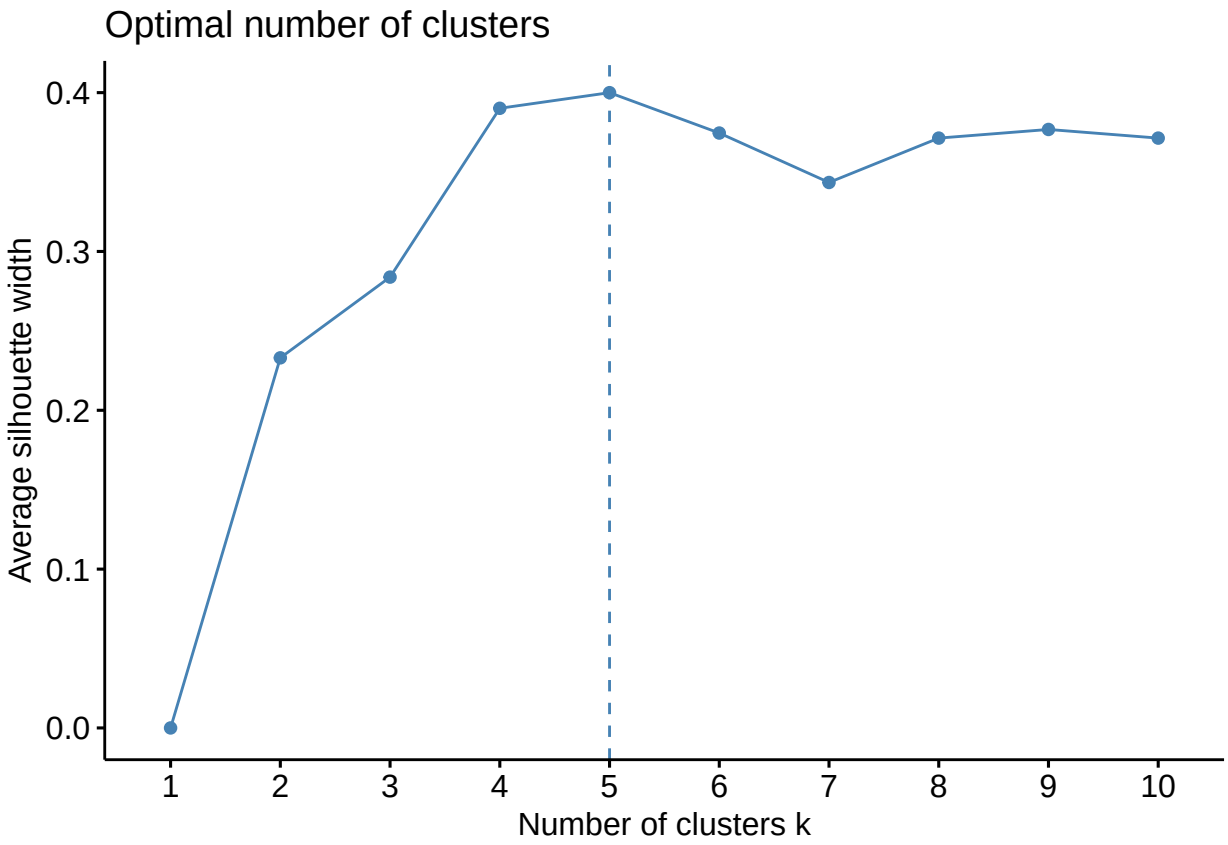
Untuk memilih metode linkage dan banyaknya gerombol bisa menggunakan

a. Koefisien silhoutte dan WSS (seperti k-means)
b. Menggunakan dendogram

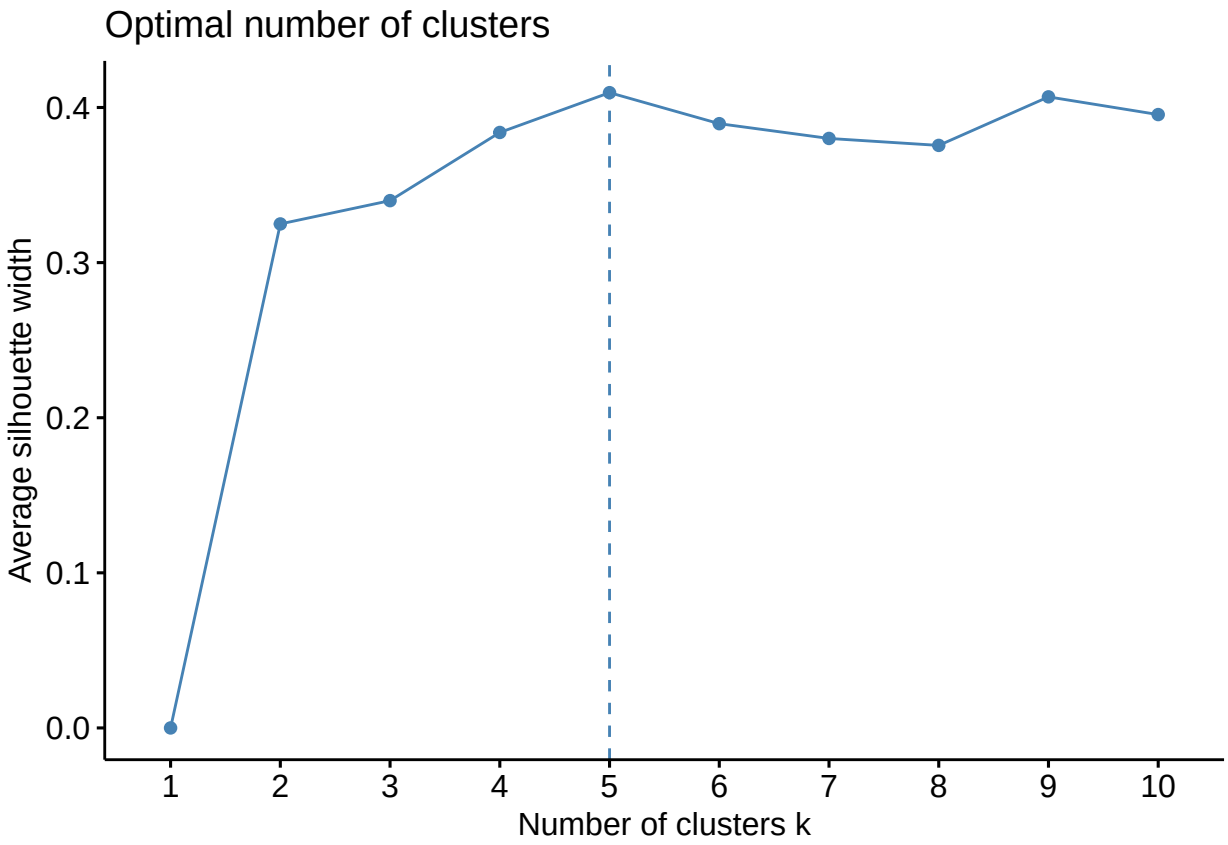
Menggunakan koefisien silhouette dan wss

Untuk ilustrasi kita akan menggunakan metode silhouette saja karena lebih mudah menentukan jumlah gerombolnya.

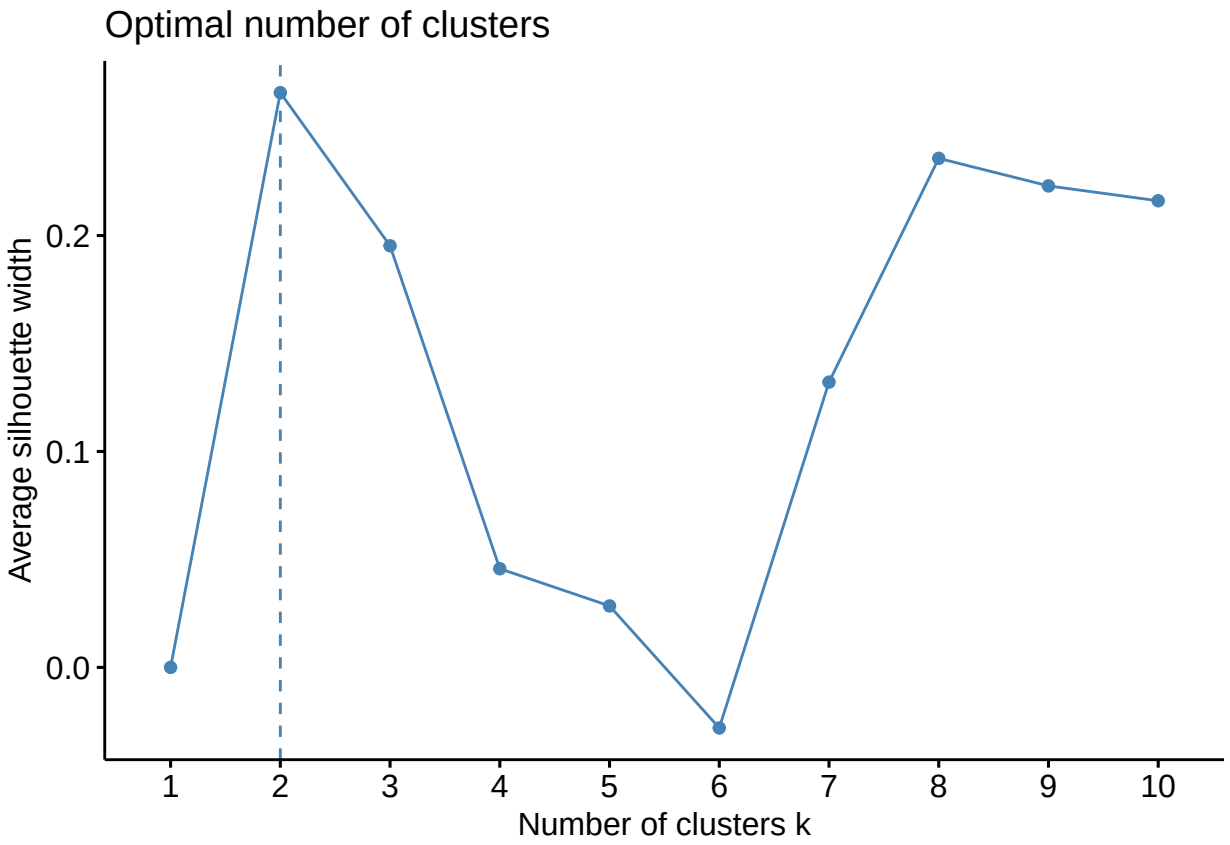
```
#complete
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "complete",hc_metric = "euclidean")
```



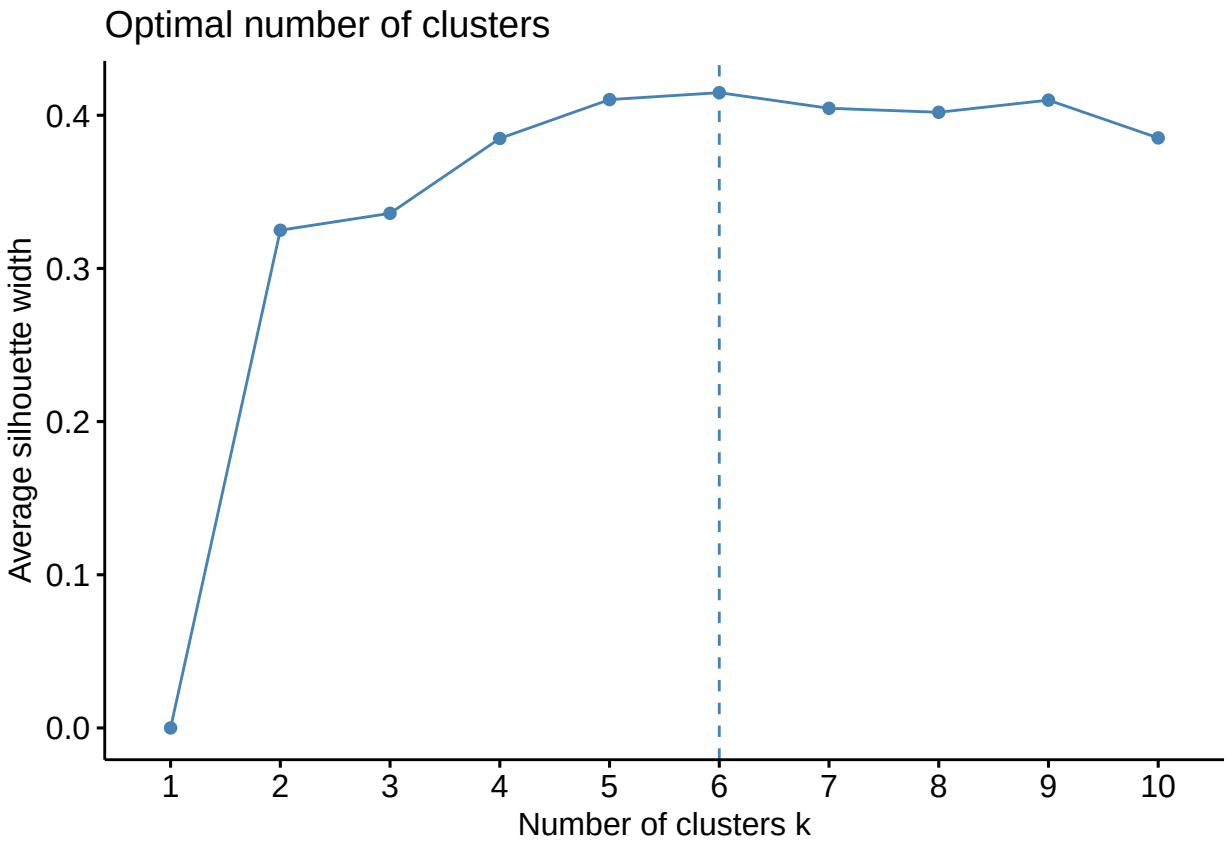

```
#average
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "average",hc_metric = "euclidean")
```




```
#centroid
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "centroid",hc_metric = "euclidean")
```




```
#ward
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "ward.D",hc_metric = "euclidean")
```

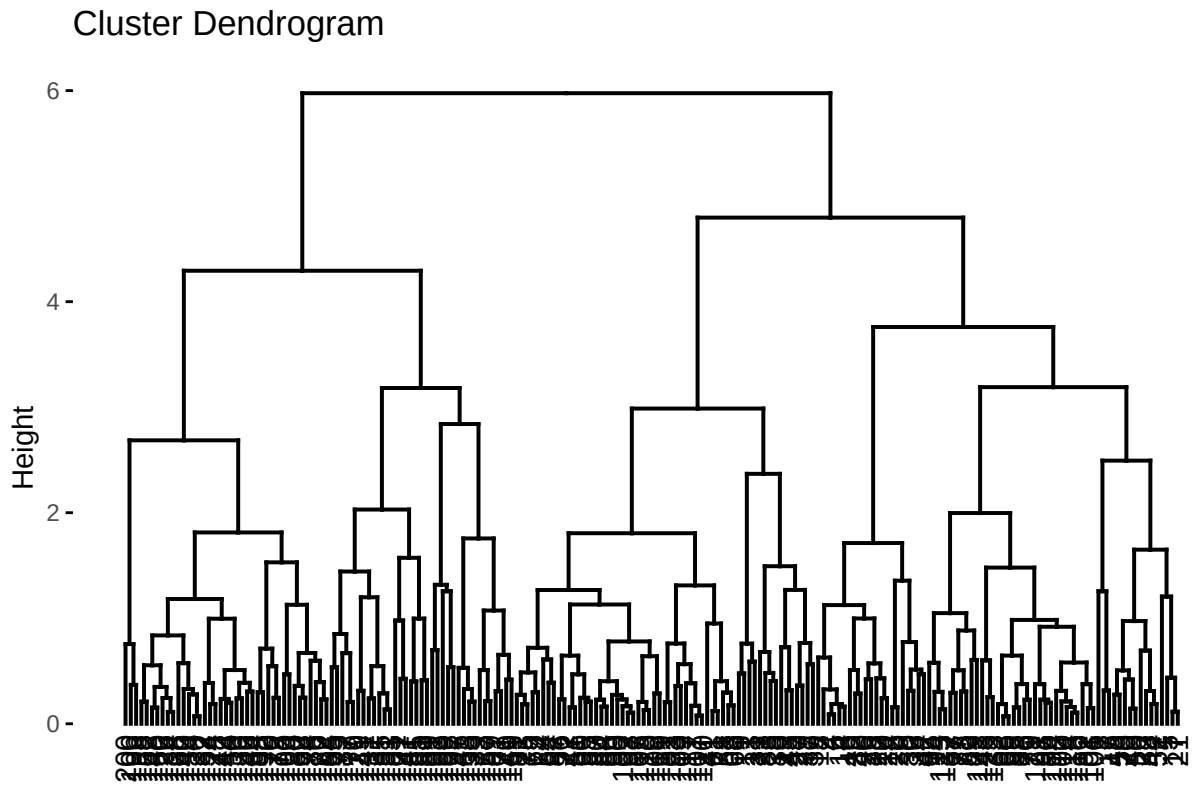


Berdasarkan koefisien silhouette, metode complete dan average memilih 5 gerombol, sedangkan metode centroid dan ward masing-masing memilih 2 dan 6 gerombol. Untuk saat ini, kita akan mencoba menggunakan 5 gerombol dengan metode complete (Jika dua metode linkage memilih banyaknya gerombol yang sama, gerombol yang terbentuk akan relatif mirip, oleh karena itu bisa pilih salah satu).

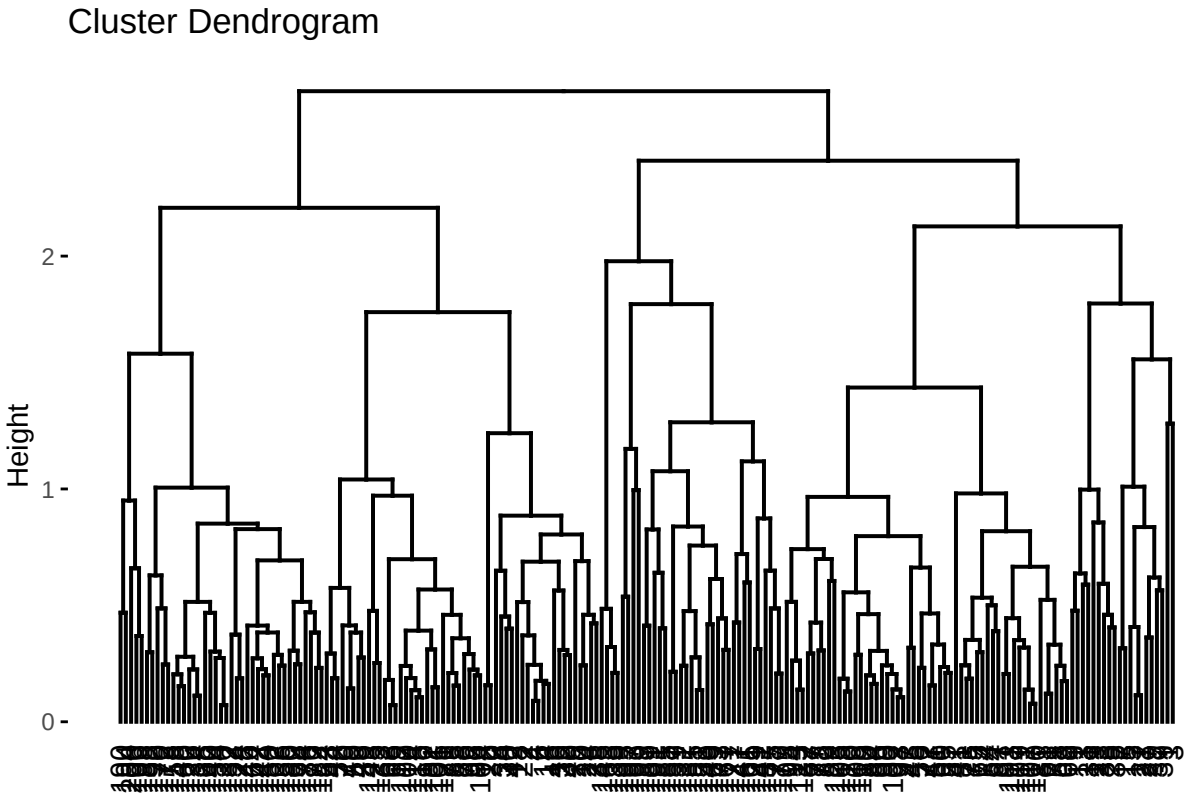
Menggunakan dendogram

Penggunaan dendogram untuk data yang memiliki amatan yang banyak mungkin tidak efektif karena memilih gerombol dengan dendogram dilakukan secara visual.

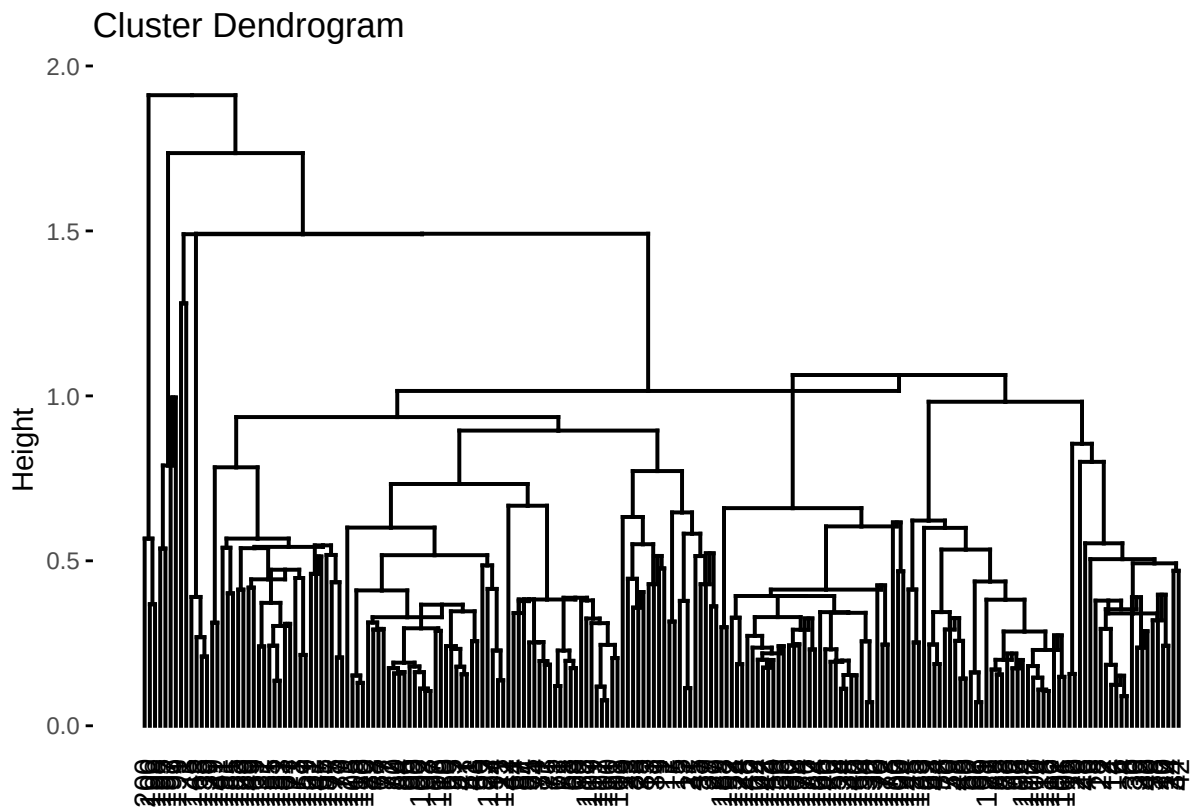
```
linkage_methods <- c("complete","average","centroid","ward.D")
hc_mall_dend <- lapply(linkage_methods, function(i)
  hclust(dist(data_mall_standardize,method = 'euclidean'),method = i)
  )
#complete
fviz_dend(hc_mall_dend[[1]])
```



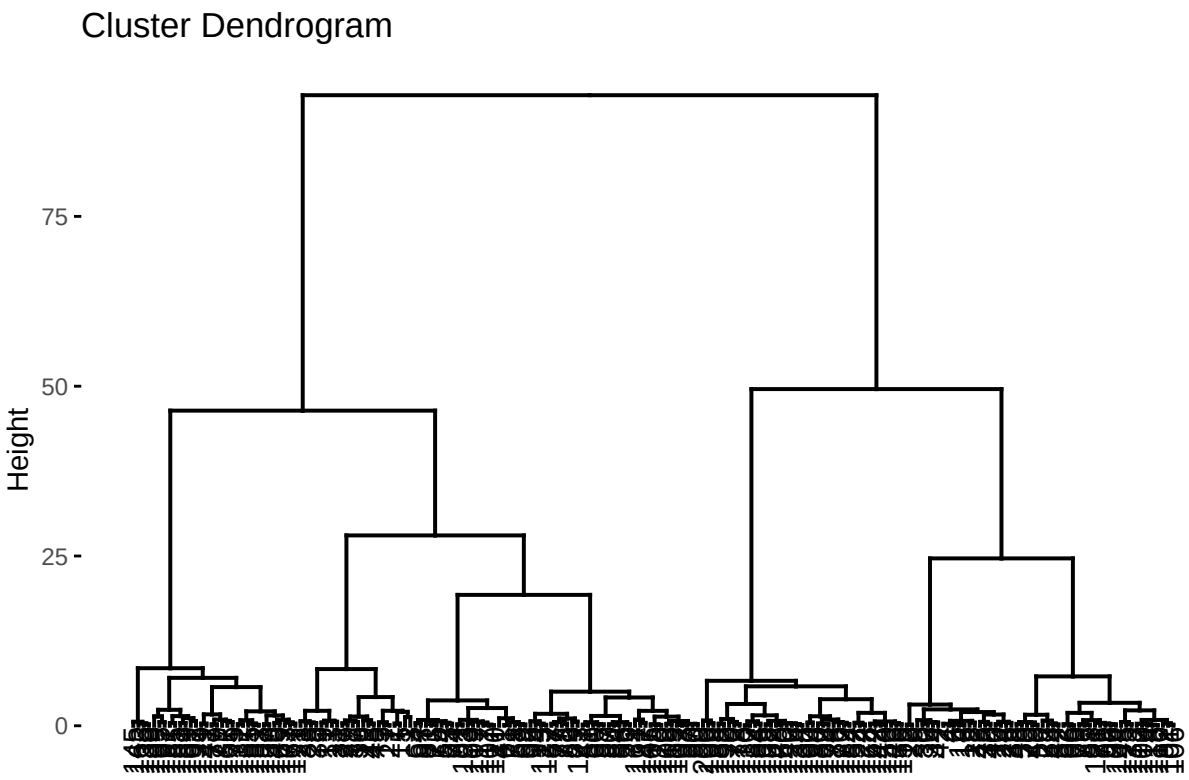

```
#average
fviz_dend(hc_mall_dend[[2]])
```




```
#centroid
fviz_dend(hc_mall_dend[[3]])
```




```
#ward
fviz_dend(hc_mall_dend[[4]])
```



Jika diperhatikan dari keempat dendogram pada masing-masing metode linkage, banyaknya gerombol yang terbentuk sama seperti menggunakan keofisien silhouette diatas.

3. Menerapkan Hierarchical Clustering

```
hc_mall <- eclust(data_mall,stand = TRUE,FUNcluster = "hclust",k=5,hc_method = "complete",hc_metric = "
hc_mall$cluster
```

```
##   [1] 1 2 1 2 1 2 1 2 3 2 3 2 3 2 1 2 1 2 3 2 1 2 3 2 3 2 3 1 3 2 3 2 3 2 3
##  [36] 2 3 2 3 2 3 2 3 1 3 2 3 1 1 1 3 1 1 3 3 3 3 3 1 3 3 1 3 3 3 1 1 3 1 1
##  [71] 3 3 3 3 3 1 1 1 1 3 3 1 3 3 1 3 3 1 1 3 3 1 3 1 1 1 3 1 3 1 1 3 3 1 3
## [106] 1 3 3 3 3 3 1 1 1 1 1 3 3 3 3 1 1 1 4 1 4 5 4 5 4 5 4 1 4 5 4 5 4 5 4
## [141] 5 4 1 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5
## [176] 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4
```

4. Interprestasi Gerombol yang terbentuk

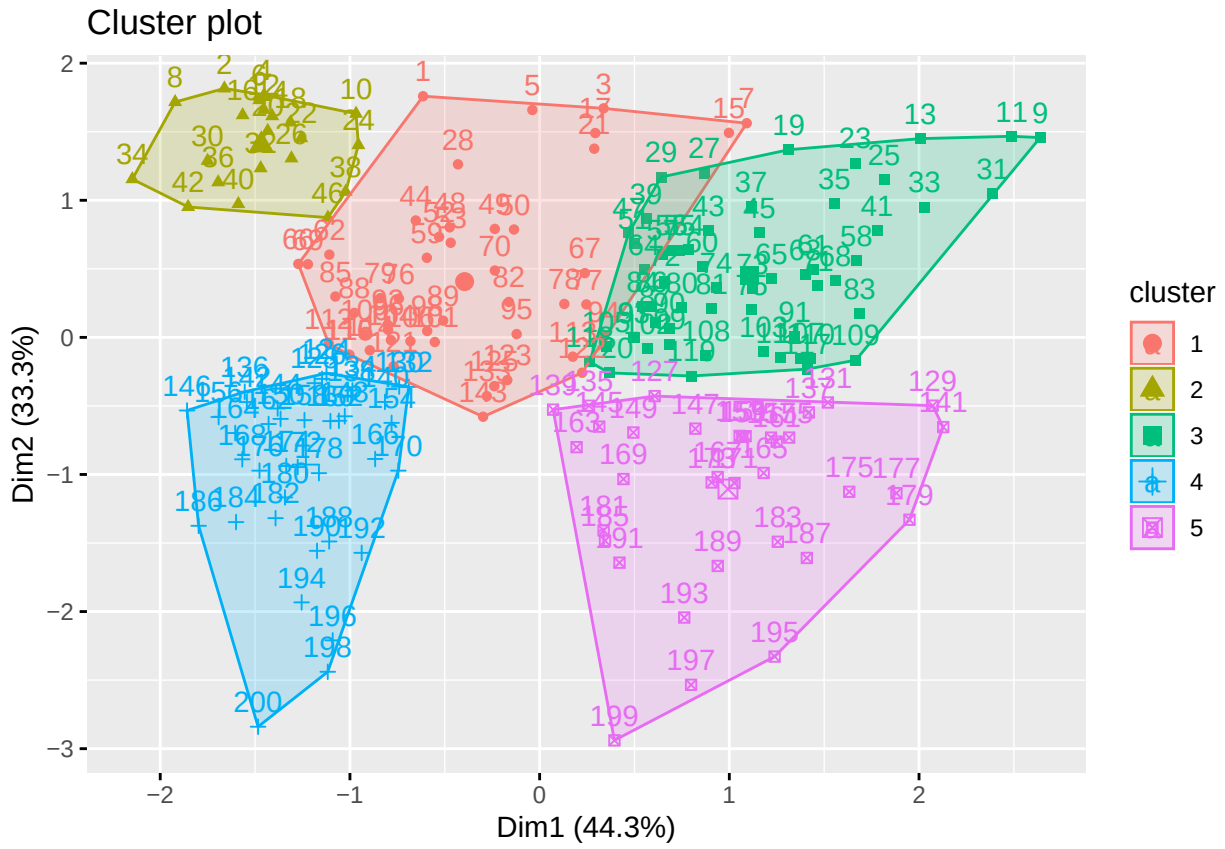
Coba lakukan interpretasi gerombol seperti metode kmeans diatas

```
aggregate(data_mall,by =list(gerombol=hc_mall$cluster),
          FUN = mean)
```

```
##   gerombol      Age Annual.Income Spending.Score
## 1        1 28.35417      50.29167       45.93750
## 2        2 24.80952      25.61905       80.23810
## 3        3 55.33333      47.31579       41.08772
## 4        4 32.69231      86.53846       82.12821
## 5        5 41.68571      88.22857       17.28571
```

Scatterplot

```
fviz_cluster(hc_mall)
```



Interpretasi dua komponen utama bisa dilihat dengan akar cirinya.

```
pca_mall <- prcomp(data_mall_standardize)
pca_mall$rotation
```

```
##                       PC1         PC2         PC3
## Age             0.70638235 -0.03014116 0.707188441
## Annual.Income  -0.04802398 -0.99883160 0.005397916
## Spending.Score -0.70619946  0.03777499 0.707004506
```

Analisis Diskriminan (Discriminant Analysisi)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Jika dalam analisis gerombol kita melakukan pengelompokan individu/objek yang ada berdasarkan kemiripan kedalam beberapa gerombol, yang menjadi pertanyaan adalah bagaimana jika ada individu baru? Individu baru tersebut termasuk dalam gerombol yang mana?
- Untuk dapat memasukkan individu baru kedalam gerombol yang ada maka harus ada suatu fungsi yang dapat membedakan antar gerombol. Fungsi tersebut disebut dengan fungsi diskriminan.
- Jadi analisisi gerombol dan analisis diskriminan adalah dua metode yang erat hubungannya dalam mengelompokkan objek dan memasukkan objek baru dalam kelompok

Contoh Penerapan Fungsi Diskriminan

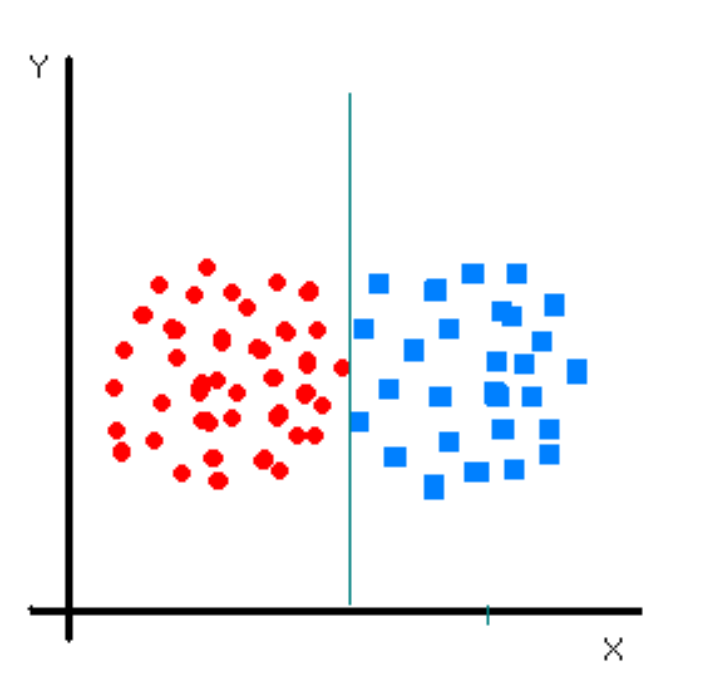
Fungsi diskriminan dipakai untuk memutuskan kasus berikut:

- Berdasarkan beberapa indicator nilai beberapa maple siswa SMA selama beberapa semester maka dapat diputuskan apakah mahasiswa ini bisa diterima atau tidak dalam salah satu PT tertentu?
- Berdasarkan data umur, pekerjaan, penghasilan, kepemilikan asset dan jumlah anggota keluarga, seseorang yang mengajukan kredit jika diterima apakah dapat dikategorikan punya potensi tidak bermasalah, sedikit bermasalah dan akan bermasalah dalam pengembalian pinjaman?
- Dengan melihat gejala-gejala yang nampak pada seseorang, bagaimana dokter bisa menduga penyakit apa yang diderita orang tersebut?

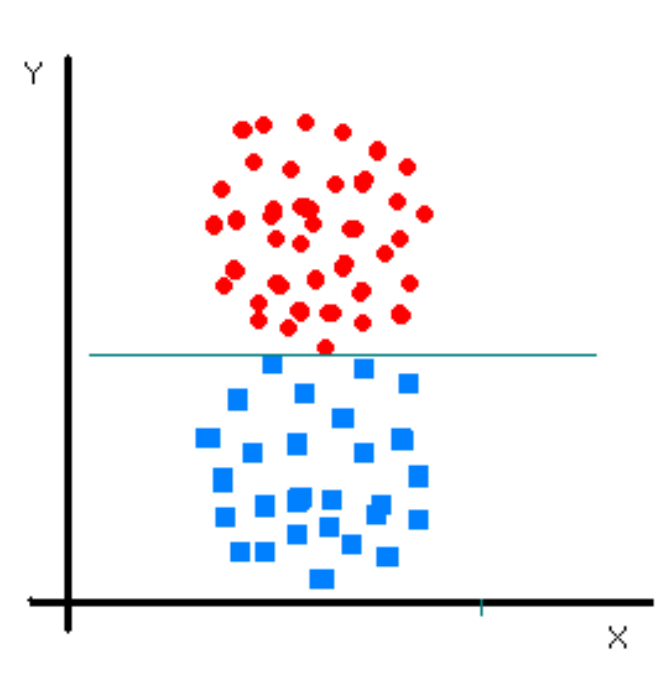
Apa Itu Fungsi Diskriminan?

- Merupakan fungsi dari beberapa peubah penciri yang dapat membedakan karakteristik individu antar gerombol
- Dari banyak peubah tersebut, fungsi diskriminan akan menghasilkan sebuah indeks
- Berdasarkan kriteria tertentu, indeks ini digunakan untuk mengklasifikasikan suatu objek masuk dalam suatu gerombol
- Tidak selalu (bahkan jarang) diperoleh fungsi diskriminan dengan tingkat ketepatan yang sempurna
- Fungsi Diskriminan memiliki ukuran yang menggambarkan tingkat ketepatan/ketidaktepatan

Ilustrasi Fungsi Diskriminan

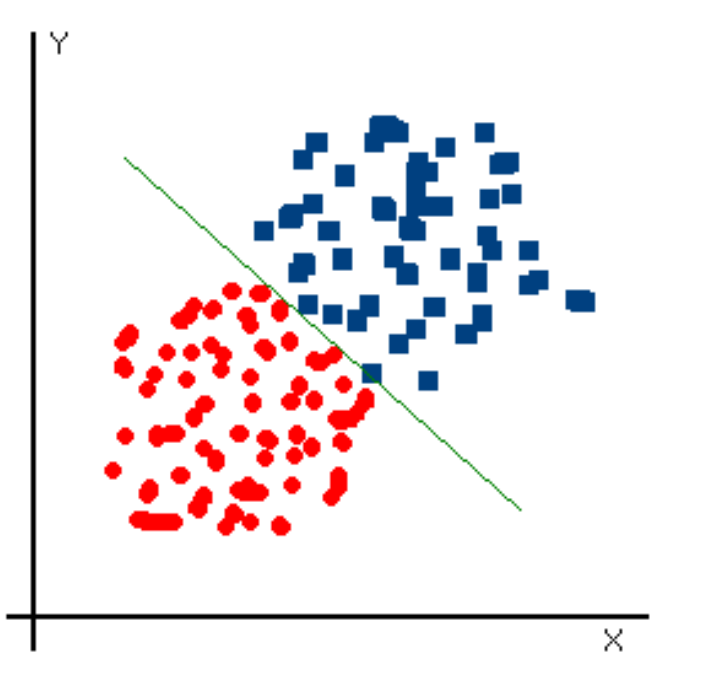


Peubah X mampu menjadi pembeda, tetapi Y tidak

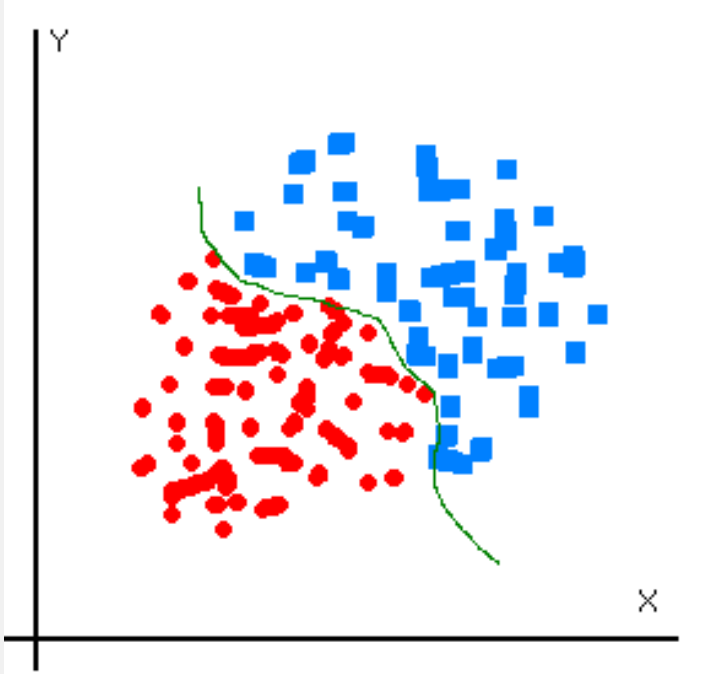


Peubah Y mampu menjadi pembeda, tetapi X tidak

Fungsi Diskriminan



Peubah X dan Y menjadi pembeda secara linear



Peubah X dan Y menjadi pembeda tetapi tidak linear

Bagaimana Cara Menduga Fungsi Diskriminan

Banyak metode untuk menduga fungsi diskriminan, yang tergantung kepada jumlah gerombol dan juga pola penggerombolan objek.

- **Pendekatan Fisher**

Pendekatn Fisher hanya digunakan jika jumlah gerombolnya hanya ada 2. Logika pendekatan Fisher bisa dituliskan sebagai berikut:

- Cari **a** sehingga jarak antara $E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\mu}_1$ di $\prod_1$ (gerombol 1) dengan $E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\mu}_2$ di $\prod_2$ (gerombol 2) maksimum, atau memaksimumkan $|\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\mu}_2|$ dengan kendala $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1$.

Pendekatan Fisher (lanjutan….)

- Diperoleh a = $\Sigma^{-1}(\mu_1 - \mu_2)$ dimana Σ adalah matriks ragam-peragam data
- Kelompokan objek yang memiliki karakteristik vector x ke gerombol Π_1 jika $\mathbf{a}'\mathbf{x} \geq h$, dan masukkan dalam gerombol Π_2 jika <h dengan $h = \mathbf{a}'(\mu_1 + \mu_2) / 2$. Dengan kata lain, **x** akan dimasukkan ke gerombol yang paling dekat kemiripannya.

Ilustrasi penerapan Pendekatan Fisher

Telah diketahui karakterisik ikan salmon yang berasal dari Alaska dan Kanada dengan melihat pertumbuhannya ketika hidup di air tawar (x1) dan ketika hidup di airl laut (x2). Dalam suatu penangkapan ikan salmon diinginkan bisa diidentifikasi apakah ikan yang tertangkap berasal dari Alaska atau Kanada. Lima puluh ikan diambil dari setiap tempat, kemudian diukur x1 dan x2 untuk mengetes model Fisher yang akan didapat(Minitab, Inc).

Ilustrasi penerapan Pendekatan Fisher

$$\mathbf{S} = \begin{bmatrix} 676.0 & -649.1 \\ -649.1 & 2138.1 \end{bmatrix},$$

serta vektor rataan untuk masing masing populasi

ikan dari Alaska $\quad \bar{\mathbf{x}}_1 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix}$

ikan dari Canada $\quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix}$

Sehingga diperoleh vektor fungsi diskriminan

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{x}_1 - \bar{x}_2) = \mathbf{a} = \mathbf{S}^{-1}\bar{\mathbf{x}}_1 - \mathbf{S}^{-1}\bar{\mathbf{x}}_2 = \begin{bmatrix} -0.0521 \\ 0.0137 \end{bmatrix},$$

dan batasan nilai bagi kedua populasi sebesar $\hat{k} = -0.5657$.

Ilustrasi penerapan Pendekatan Fisher

- Dengan demikian, jika kita memiliki suatu pengamatan baru **x** = (x_1, x_2) maka kita akan memasukkannya ke populasi 1 (ikan dari Alaska) jika

$$-0.0521\, x_1 + 0.0137\, x_2 \geq -0.5657$$

dan jika sebaliknya maka kita masukkan ke populasi ke-2 (Kanada).

- Sebagai teladan, jika diperoleh sebuah ikan dengan nilai pengamatan **x** = (103, 405), maka nilai

a'**x** = -0.0521 (103) + 0.0137 (405) = 10.918, dan kita masukkan ke dalam jenis ikan Alaska

Hasil Klasifikasi berdasarkan Pendekatan Fisher --

Bentuk tabel salah klasifikasinya adalah:

		Hasil Klasifikasi		% Salah Klasifikasi
		Alaska	Canada	
Seharusnya	Alaska	44	6	12%
	Canada	1	49	2%
Total				7%

Fungsi diskriminan yang dibentuk diatas memberikan tingkat kesalahan sebesar 7%

Metode Pendekatan Lain --

Cara lain untuk melakukan klasifikasi adalah menggunakan konsep jarak terhadap vektor rataan populasi yang paling dekat. Artinya jika ada suatu pengamatan baru **x** = (x_1, x_2), maka pengamatan atau objek baru ini akan kita masukkan ke dalam populasi ke-1 (Π_1) hanya jika jarak **x** terhadap vektor rataan populasi ke-1 lebih dekat daripada jarak **x** terhadap vektor rataan populasi ke-2. Jarak antara **x** terhadap vektor rataan diperoleh menggunakan formula mahalanobis, yaitu:

$$d_j(\mathbf{x}) = \{[\mathbf{x} - \bar{\mathbf{x}}_j]' \mathbf{S}^{-1} [\mathbf{x} - \bar{\mathbf{x}}_j]\}^{1/2}$$

- Misalkan untuk pengamatan **x** = (103, 405) seperti pada ilustrasi sebelumnya. Setelah dihitung jarak dengan populasi 1 dan populasi 2 diperoleh:

  $d_1(\mathbf{x}) = 0.5421$

  $d_2(\mathbf{x}) = 1.3322$

- Karena $d_1(\mathbf{x}) < d_2(\mathbf{x})$ maka **x** diklasifikasikan berasal dari populasi 1 (ikan dari Alaska).

Metode Pendekatan Lain --

Pendekatan lain yang juga dapat digunakan adalah menggunakan peluang posterior. Suatu pengamatan **x** = (x_1, x_2) akan diklasifikasikan ke dalam populasi Π_1 hanya jika peluang posteiornya lebih besar dari pada peluang posterior masuk ke Π_2, dan sebaliknya. Peluang posterior masuk ke dalam Π_j adalah

$$P(j \mid \mathbf{x}) = \frac{e^{-\frac{1}{2}d_j^2(\mathbf{x})}}{e^{-\frac{1}{2}d_1^2(\mathbf{x})} + e^{-\frac{1}{2}d_2^2(\mathbf{x})}}$$

- Misalkan untuk pengamatan **x** = (103, 405) seperti pada ilustrasi sebelumnya. Setelah dihitung peluang posteriornya diperoleh:

  $P(1|\mathbf{x}) = 0.677$

  $P(2|\mathbf{x}) = 0.323$.

- Karena $P(1|\mathbf{x}) > P(2|\mathbf{x})$ maka **x** sekali lagi diklasifikasikan berasal dari Alaska.

Analisis Diskriminan untuk k Populasi yang Menyebar Normal

- Ada konsep sebaran prior
- Seringkali juga perlu mempertimbangkan biaya salah klasifikasi sehingga perlu mencari fungsi yang meminimumkan expected cost of missclassification

$$\sum_{t=1}^{k} \pi_t \sum_{s=1}^{k} P(s \mid t) c(s \mid t)$$

Analisis Diskriminan Linear (1)

- Dengan asumsi bahwa setiap populasi menyebar normal ganda dan matriks ragam-peragam sama di setiap populasi serta biaya salah klasifikasi sama besar di setiap populasi, maka aturan yang paling sederhana pada klasifikasi bisa dinyatakan dalam fungsi kuadrat jarak yaitu

$$d_t(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_t)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_t) - 2\, ln(\pi_t)$$

- Suatu objek **x** diklasifikasikan kepada populasi yang terdekat, yang dihitung menggunakan formula di atas. Atau, **x** akan diklasifikasikan berasal dari populasi ke-*t* jika

$$d_t^2(\mathbf{x}) = \min_{j=1,\ldots,k} \left\{ d_j^2(\mathbf{x}) \right\}$$

Analisis Diskriminan Linear (2)

- Seperti halnya pada bagian terdahulu, mengklasifikasikan objek pengamatan ke populasi yang terdekat setara dengan mengklasifikasikan objek ke populasi dengan peluang posterior yang paling besar. Pada kasus k buah populasi, peluang tersebut besarnya diperoleh dari :

$$P(t \mid x) = \frac{e^{-\frac{1}{2}d_t^2(x)}}{\sum_{j=1}^{k} e^{-\frac{1}{2}d_j^2(x)}} \qquad t = 1, 2, \ldots, k$$

Menduga Tingkat Salah Klasifikasi

- *Error Rate*, dugaan tingkat kesalahan di populasi ke-s adalah:

$$\hat{E}R(s) = \sum_{t=1, t \neq s}^{k} P(t \mid s)$$

Menduga Tingkat Salah Klasifikasi (1)

Pendugaan Tingkat Kesalahan dengan Validasi Silang

- jika ada *n* objek pengamatan, maka hanya (*n* – 1) pengamatan yang digunakan sebagai gugus data pembentukan fungsi diskriminan
- satu pengamatan sisanya digunakan untuk evaluasi
- proses di atas diulang sebanyak n kali, satu kali untuk setiap data yang disisihkan
- proporsi kesalahan adalah dugaan tingkat kesalahan

Menduga Tingkat Salah Klasifikasi (2)

posterior probability error rate

$$\mathrm{PPER}_{1t} = 1 - \frac{1}{\pi_t \sum_{j=1}^{k} n_j} \sum_{D_t} P(t \mid \mathbf{x})$$

Simple PPER

$$\mathrm{PPER}_{2t} = 1 - \frac{1}{\pi_t} \sum_{j=1}^{k} \frac{\pi_j}{n_j} \sum_{D_{jt}} P(t \mid x),$$

Stratified PPER

Analisis Diskriminan Kuadratik

Asumsi: populasi menyebar normal ganda, namun matriks ragam-peragamnya tidak sama

Suatu objek pengamatan tertentu **x** akan dimasukkan ke populasi ke-t jika

$$d_t^2(\mathbf{x}) = \min_{j=1,\ldots,k} \left\{ d_j^2(\mathbf{x}) \right\}$$

dengan $d_j^2(\mathbf{x})$ adalah kuadrat jarak yang didefinisikan (sedikit berbeda dengan kasus fungsi diskriminan linear) sebagai:

$$d_j^2(\mathbf{x}) = [\mathbf{x} - \bar{\mathbf{x}}_j]' \mathbf{S}^{-1} [\mathbf{x} - \bar{\mathbf{x}}_j] + ln \, |\mathbf{S}_j| - 2 \, ln(\pi_t); \; j = 1, 2, \ldots, k.$$

Formula peluang posterior sama persis dengan formula peluang posterior untuk kasus diskriminan linear kecuali pada formula $d_j^2(\mathbf{x})$.

Pemilihan Peubah Yang digunakan pada Analisis Diskriminan (1)

- Agar supaya peubah yang digunakan dalam fungsi diskriminan tidak terlalu banyak, maka dapat dipilih beberapa peubah yang paling penting dan memiliki daya pembeda yang kuat
- Beberapa metode yang dapat digunakan seperti pada penerapan regresi liniear berganda:

forward selection: dimulai dengan memilih satu peubah yang paling penting, dan dilanjutkan dengan pemilihan peubah penting lain satu demi satu menggunakan suatu kriteria tertentu.

Pemilihan Peubah Yang digunakan pada Analisis Diskriminan (2)

backward selection: dimulai dengan model penuh, yaitu memuat semua peubah. Di setiap tahap dilakukan pembuangan peubah yang paling tidak penting satu demi satu dengan kriteria yang sama dengan prosedur forward. Proses diteruskan hingga tidak ada lagi peubah yang dikeluarkan. Prosedur ini dikenal sebagai prosedur.

stepwise selection: Kombinasi antara kedua prosedur di atas. Di setiap tahap dimungkinkan ada peubah yang masuk sekaligus ada peubah yang dikeluarkan, berdasarkan kriteria tertentu yang ditetapkan pada awal proses.

Terimakasih