ANALISIS GEROMBOL

CLUSTER ANALYSIS

Pendahuluan

- Tujuan dari analisis gerombol :

  Menggabungkan beberapa objek ke dalam kelompok-kelompok berdasarkan sifat kemiripan atau sifat ketidakmiripan antar objek

  Objek dalam kelompok lebih mirip dibandingkan dengan objek antar kelompok

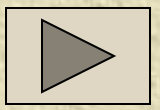
  Ketakmiripan antar objek diukur dengan jarak tertentu ⟶ jarak Euclid, dll

Hal yang perlu diperhatikan dalam membuat penggerombolan :

- Tujuan dari penggerombolan
- Kemiripan atau ketakmiripan seperti apa yang diharapkan → berhubungan dengan pemilihan peubah
- Mengkuantifikasi ukuran kemiripan antar objek

Metode Penggerombolan

- Metode Grafik
- Metode Penggerombolan Berhirarki
- Metode Penggerombolan tak Berhirarki



Metode Grafik

- Plot Profil
- Plot Andrew
- Plot Andrew termodifikasi

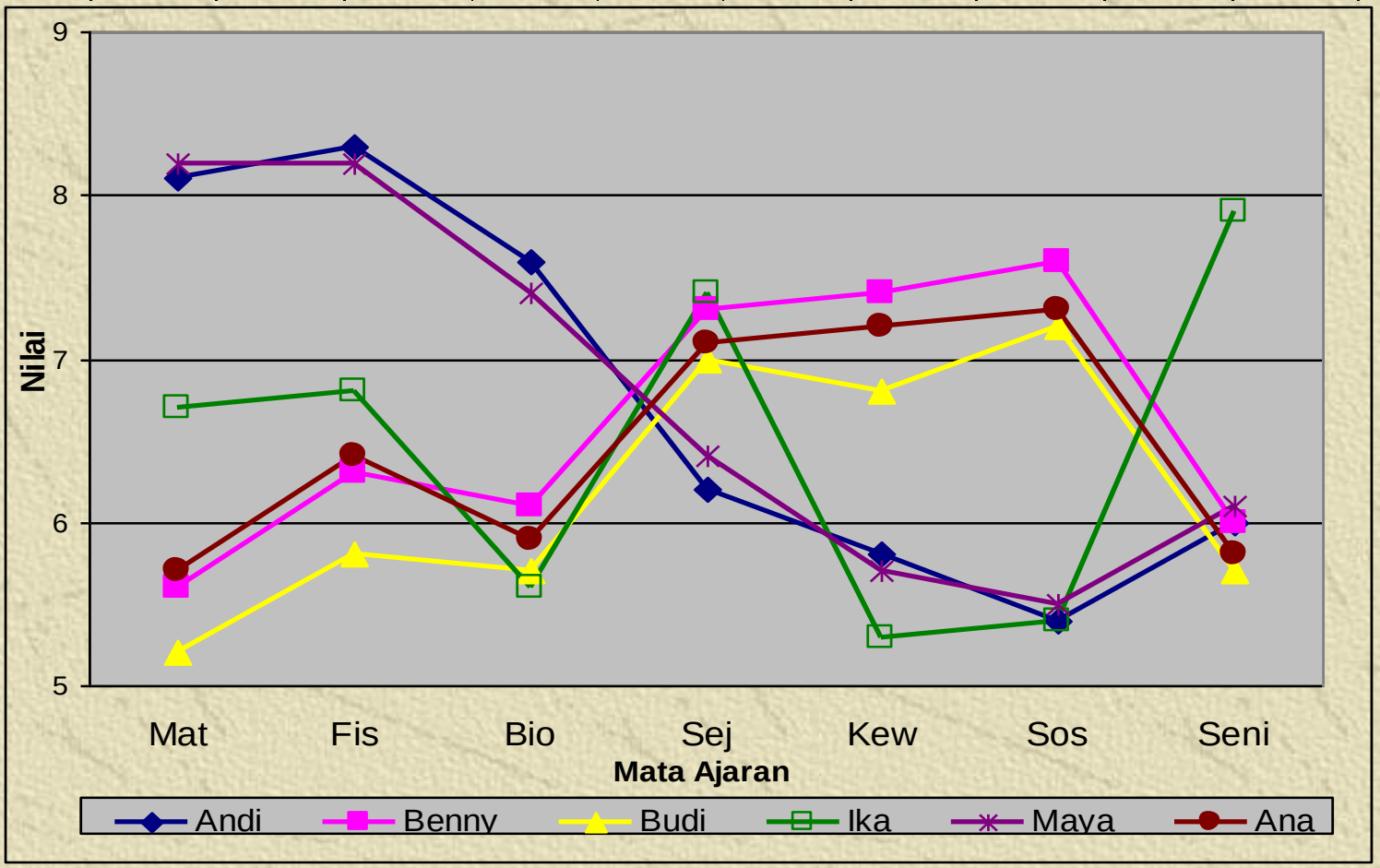
Plot Profil

- Plot profil dari setiap pengamatan
- Pembakuan data sangat membantu
- Kelemahan : tidak efektif untuk data yang terlalu banyak pengamatan.
- Ilustrasi : Diperoleh hasil ujian untuk 7 mata ajaran yaitu Matematika, Fisika, Biologi, Sejarah Nasional, Pendidikan kewiraan, dan Kesenian. Ada 6 mahasiswa yang terlibat.

Tabel datanya sebagai berikut :

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

Plot Profilnya sebagai berikut



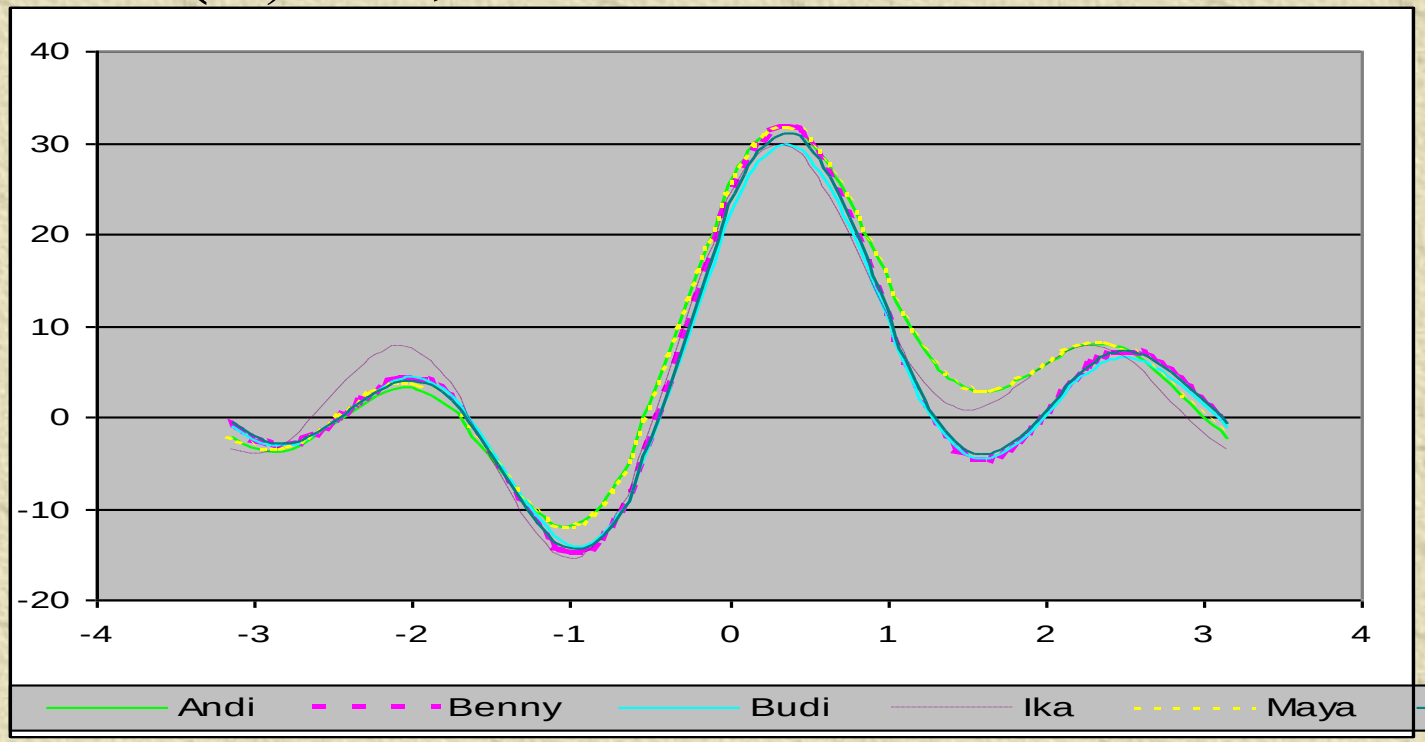
Interpretasi

- ANDI dan MAYA mempunyai profil yang mirip, keduanya mempunyai kemampuan yang tinggi di bidang IPA
- BENNY, BUDI, dan ANNA, keduanya pencinta ilmu sosial
- IKA mempunyai kearekteristik sendiri

Plot Andrews

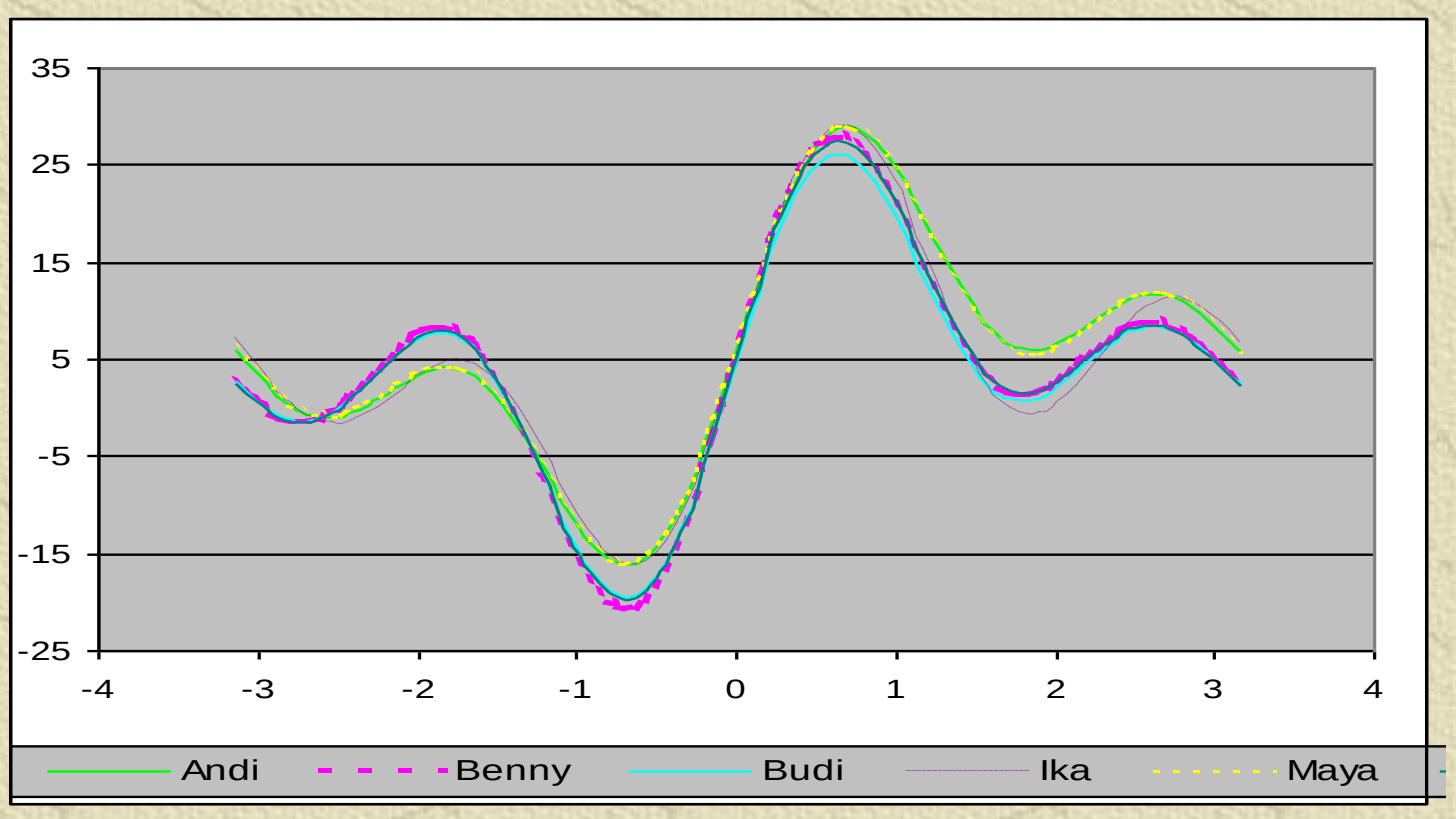
Fungsi Andrews didefinisikan sebagai berikut :

$$fx(t) = x1/\sqrt{2} + x2\ \sin(t) + x3\ \cos(t) + x4\ \sin(2t) + x5\ \cos(2t) + \ldots, \text{ untuk } -\pi \leq t \leq \pi$$



Plot Andrews Termodifikasi

$$gx(t) = (1/\sqrt{2})\ \{\ x1 + x2[\sin(t) + \cos(t) + x3[\sin(t) - \cos(t)] + x4[\sin(2t) + \cos(2t)] + x5[\sin(2t) - \cos(2t)] + \ldots\ \},\ \text{untuk}\ -\pi \le t \le \pi$$



Ukuran Kemiripan dan Ketakmiripan

Syarat jarak yang digunakan untuk mengukur ketakmiripan antar 2 objek *a* dan *b*, dinotasikan dengan d*(a,b)*, :

- $d(a, b) \geq 0$
- $d(a, a) = 0$
- $d(a, b) = d(b, a)$
- $d(a, b)$ 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

Beberapa konsep jarak yang digunakan :

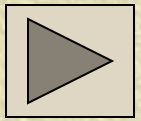
Jarak	Formula
Jarak Euclidean	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$ $= \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Jarak Minkowski / Jarak city-block / Jarak Manhattan	$d(x, \mathbf{y}) = \left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert^k\right]^{\frac{1}{k}}$
Jarak Mahalanobis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

2. Metode Penggerombolan berhirarki

1. Metode aglomeratif
2. Metode berhirarki divisif (pemisahan)

Beberapa ukuran ketakmiripan antar gerombol :

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



Lanjutan

- Pautan Tunggal (Single Linkage = Nearest Neighbor)

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

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

Lanjutan

- Pautan Lengkap (Complete Linkage = Farthest Neighbor)

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

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

Lanjutan

- 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 :

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

Lanjutan

- 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

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

Lanjutan

- 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*

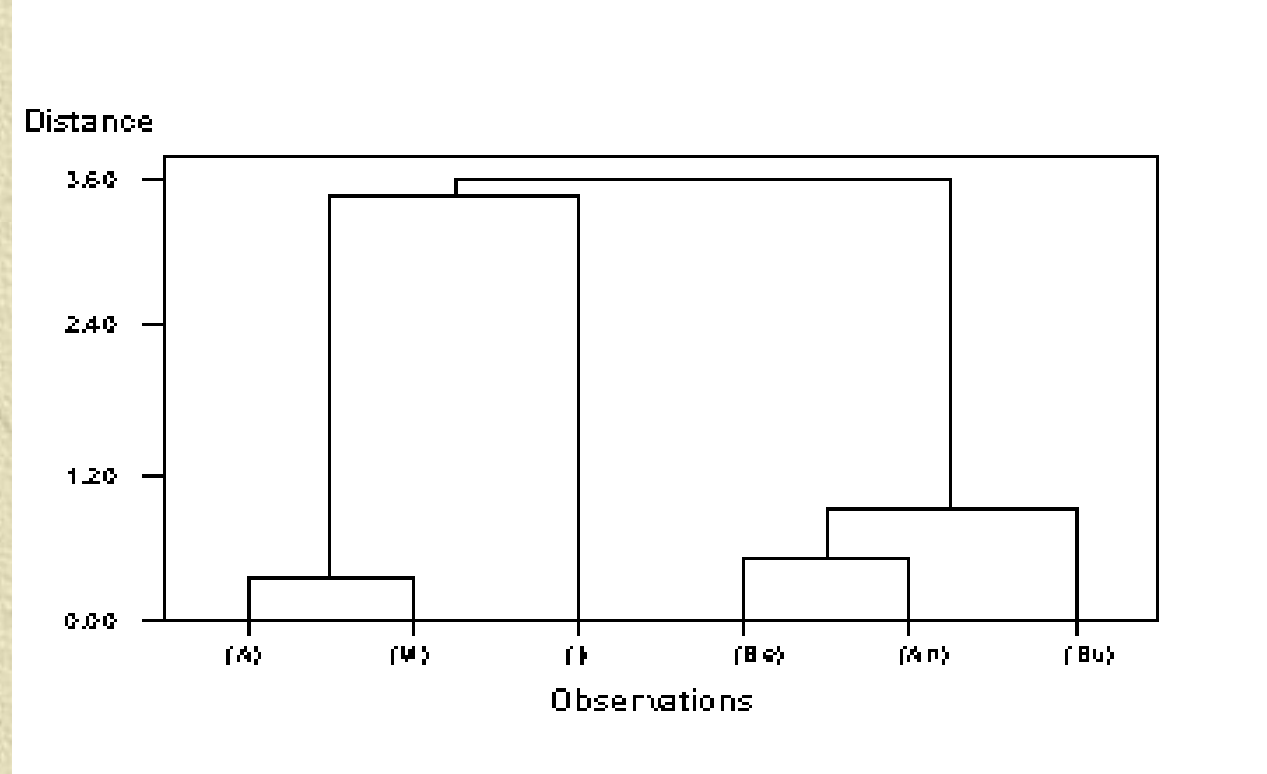
$$W(B_r, B_s) = \frac{1}{n_r n_s} \sum_{\mathbf{x}_i \in B_r} \sum_{\mathbf{x}_j \in B_s} d(\mathbf{x}_i, \mathbf{x}_j)$$

Ilustrasi :

- Dari ilustrasi sampel sebelumnya, digunakan konsep jarak Euclidian dan diperoleh matriks jarak sbb :

	Andi	Benny	Budi	Ika	Maya	Ana
Andi {A}	0.0000	4.5946	4.8208	3.6756	0.3606	4.3220
Benny {Be}	4.5946	0.0000	1.1269	3.8184	4.4922	0.5196
Budi {Bu}	4.8208	1.1269	0.0000	3.7081	4.7191	0.9165
Ika {I}	3.6756	3.8184	3.7081	0.0000	3.4438	3.6014
Maya {M}	0.3606	4.4922	4.7191	3.4438	0.0000	4.2202
Ana {An}	4.3220	0.5196	0.9165	3.6014	4.2202	0.0000

Dengan menggunakan konsep Single lingkage diperoleh hasil dalam bentuk dendogram sebagai berikut :



3. 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.

Ilustrasi

- Misalkan ada dua variabel X_1 dan X_2 yang tiap objeknya diberi nama A, B, C dan D. Datanya sebagai berikut:

Objek	Pengamatan	
	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Langkah yang dilakukan :

1. Dikelompokkan ke dalam 2 kelompok. Centroid dipilih secara acak : $\mathbf{c}_1 = (2, 2)$ dan $\mathbf{c}_2 = (-1, -2)$.
2. Jarak yang digunakan jarak Euclidian. Memasukkan objek ke gerombol berpatokan pada jarak terdekat

   Diperoleh matriks jarak sbb :

	$\mathbf{c}_1$	$\mathbf{c}_2$
A	$(5-2)^2 + (3-2)^2 = 10$	$(5+1)^2+(3+2)^2 = 61$
B	$(-1-2)^2 + (1-2)^2 = 10$	$(-1+1)^2 + (1+2)^2 = 9$
C	$(1-2)^2 + (-2-2)^2 = 17$	$(1+1)^2 + (-2+2)^2 = 4$
D	$(-3-2)^2 + (-2-2)^2 = 41$	$(-3+1)^2 + (-2+2)^2 = 4$

Lanjutan

3. Hitung centroid baru, rataan dari vektor masing-masing unsur.

$\mathbf{c}_1 = (5, 3)$

$\mathbf{c}_2 = [(-1, 1) + (1, -2) + (-3, -2)]/3 = (-1, -1)$

Diperoleh matriks yang sbb :

	$\mathbf{c}_1$	$\mathbf{c}_2$
A	$(5-5)^2 + (3-3)^2 = 0$	$(5+1)^2+(3+1)^2 = 52$
B	$(-1-5)^2 + (1-3)^2 = 40$	$(-1+1)^2 + (1+1)^2 = 4$
C	$(1-5)^2 + (-2-3)^2 = 41$	$(1+1)^2 + (-2+1)^2 = 5$
D	$(-3-5)^2 + (-2-3)^2 = 89$	$(-3+1)^2 + (-2+1)^2 = 5$

→ Diperoleh 2 gerombol : G_1 = {A} dan G_2 = {B, C, D}.