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

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

Transformasi

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 $KU_i = Var(KU_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i = \lambda_i$

  $Var(KU_1) = \lambda_1$

  $Var(KU_2) = \lambda_2$

  .....

  $Var(KU_p) = \lambda_p$

- Total ragam peubah asal X = tr($\mathbf{\Sigma}$), 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≠j=1,2,…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

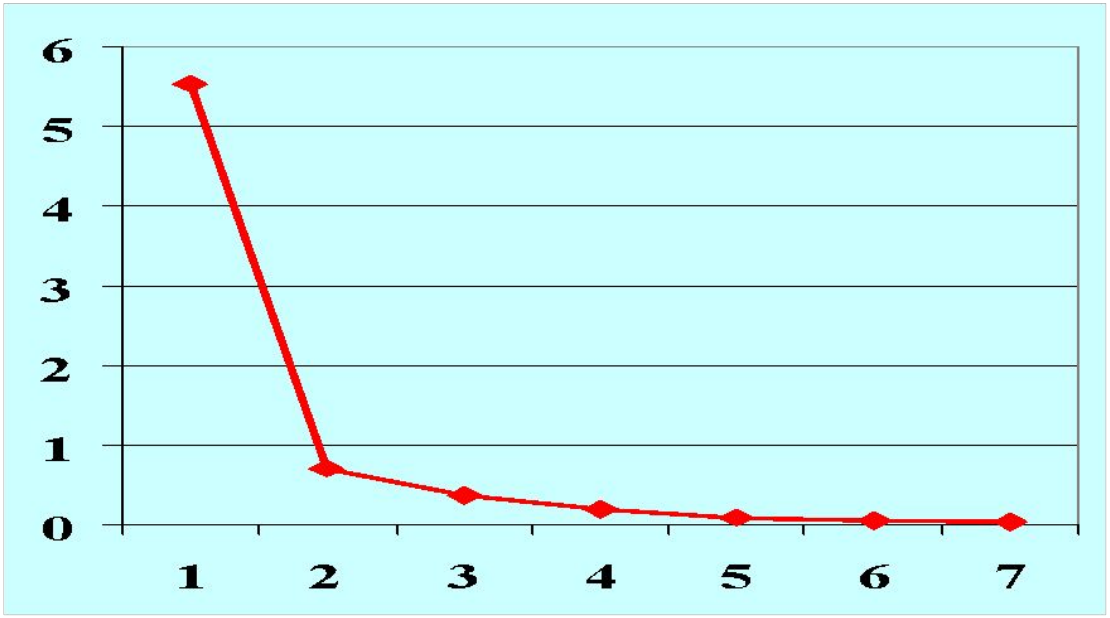
Matriks Korelasi

	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

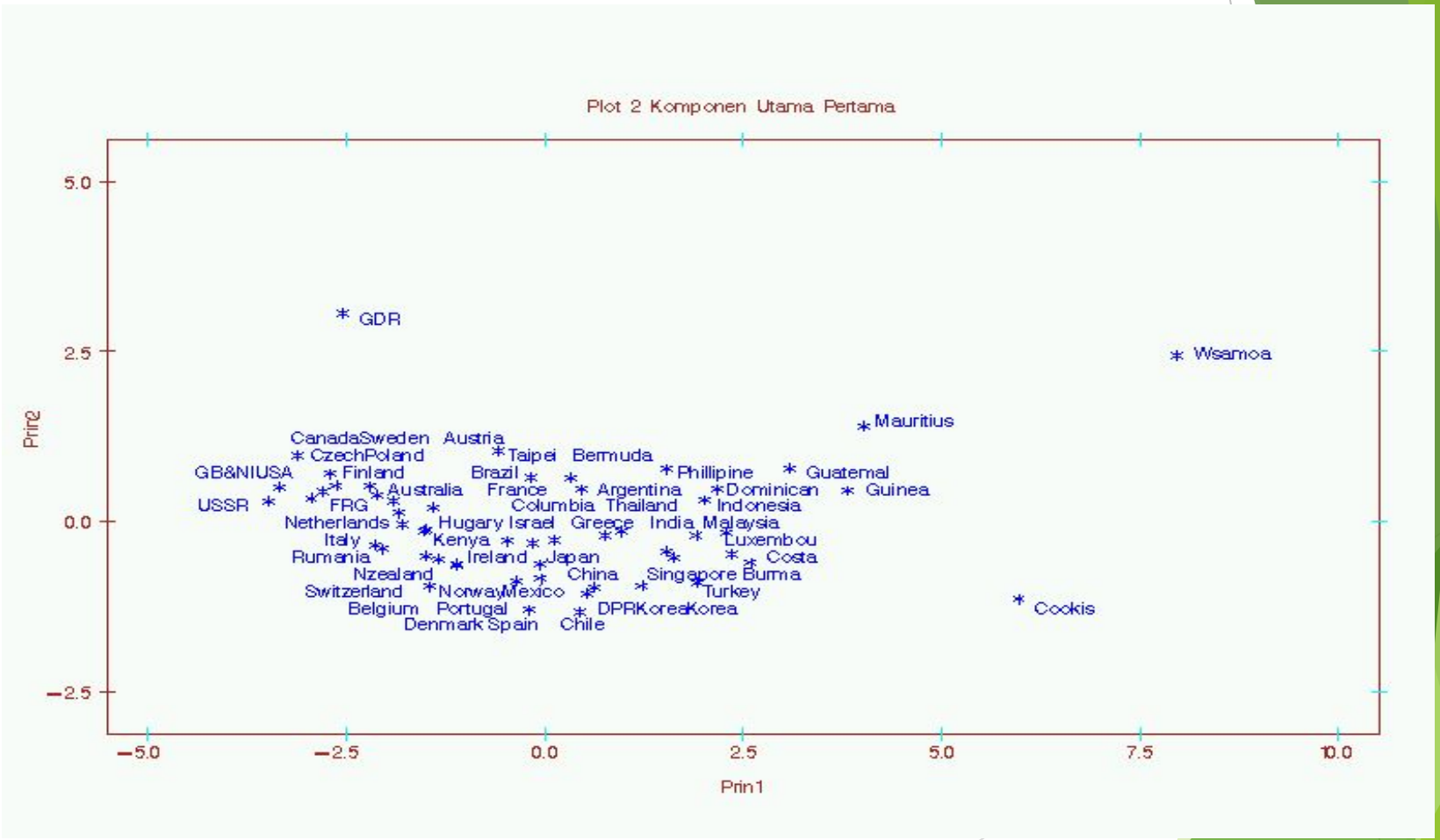
- Sehingga jika digunakan dua komponen utama akan didapatkan keragaman total yang mampu dijelaskan keduanya adalah 89.27%.

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