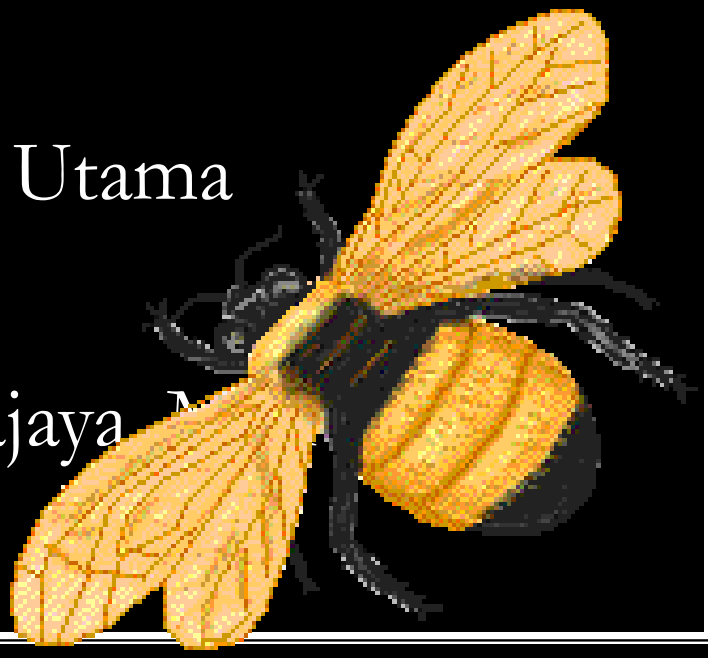
Analisis Peubah Ganda

Analisis Komponen Utama

Dr. Ir. I Made Sumertajaya M



Pengamatan Peubah Ganda

- memerlukan ‘sumberdaya’ lebih, dalam analisis
- informasi tumpang tindih pada beberapa peubah

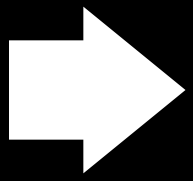
Apa itu Komponen Utama

- Merupakan kombinasi linear dari peubah yang diamati → informasi yang terkandung pada KU merupakan gabungan dari semua peubah dengan bobot tertentu
- Kombinasi linear yang dipilih merupakan kombinasi linear dengan ragam paling besar → memuat informasi paling banyak
- Antar KU bersifat ortogonal → tidak berkorelasi → informasi tidak tumpang tindih

Analisis Komponen Utama

Gugus peubah asal

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



Gugus KU

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

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

Ilustrasi Komponen Utama

Untuk menceritakan bagaimana wajah pacar kita waktu SMA, tidak perlu disebutkan hidungnya mancung, kulitnya halus, rambutnya indah tergerai dan sebagainya. Tapi cukup katakan 'Pacar saya waktu SMA orangnya cantik'. Kata 'cantik' sudah mampu menggambarkan uraian sebelumnya.

Bentuk Komponen Utama

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

Jika gugus peubah asal $\{X_1, X_2, \ldots, X_p\}$ memiliki matriks ragam peragam $\mathbf{\Sigma}$ maka ragam dari komponen utama adalah

$$\sigma^2_{KU_1} = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{1i}a_{1j}\sigma_{ij}$$

Tugas kita adalah bagaimana mendapatkan vektor $\mathbf{a}_1$ sehingga ragam di atas maksimum (vektor ini disebut vektor koefisien)

Mendapatkan KU pertama

- Vektor $\mathbf{a}_1$ merupakan vektor ciri matriks $\boldsymbol{\Sigma}$ yang berpadanan dengan akar ciri paling besar.
- Kombinasi linear dari $\{X_1, X_2, \ldots, X_p\}$ berupa $KU_1 = \mathbf{a}_1\mathbf{x} = a_{11}x_1 + \ldots + a_{1p}x_p$ dikenal sebagai KU pertama dan memiliki ragam sebesar λ_1 = akar ciri terbesar

KU kedua

- Bentuknya $KU_2 = \mathbf{a}_2\mathbf{x} = a_{21}x_1 + \ldots + a_{2p}x_p$
- Mencari vektor a_2 sehingga ragam dari KU_2 maksimum, dan KU_2 tidak berkorelasi dengan KU_1
- a_2 tidak lain adalah vektor ciri yang berpadanan dengan akar ciri terbesar kedua dari matriks $\boldsymbol{\Sigma}$.

Komponen Utama

Misalkan $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ adalah vektor ciri yang berpadanan dengan vektor ciri $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ dari matriks $\boldsymbol{\Sigma}$, dan panjang dari setiap vektor itu masing masing adalah 1, atau $\mathbf{a}_i'\mathbf{a}_i = 1$ untuk $i = 1, 2, \ldots, p$. Maka $KU_1 = \mathbf{a}_1'\mathbf{x}$, $KU_2 = \mathbf{a}_2'\mathbf{x}$, …, $KU_p = \mathbf{a}_p'\mathbf{x}$ berturut-turut adalah komponen utama pertama, kedua, …, ke-p dari $\mathbf{x}$. Lebih lanjut $var(KU_1) = \lambda_1$, $var(KU_2) = \lambda_2$, …, $var(KU_p) = \lambda_p$, atau akar ciri dari matriks ragam peragam $\boldsymbol{\Sigma}$ adalah ragam dari komponen-komponen utama.

Kontribusi setiap KU

- Ragam dari setiap KU sama dengan akar ciri $\boldsymbol{\Sigma}$, yaitu λ_i
- Total ragam peubah asal seluruhnya adalah tr($\boldsymbol{\Sigma}$), dan ini sama dengan penjumlahan dari seluruh akar ciri
- Jadi kontribusi setiap KU ke-j adalah sebesar

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

Interpretasi setiap KU

- Interpretasi setiap KU didasarkan pada nilai pada vektor $\mathbf{a}_j$, karena nilai ini berhubungan linear dengan korelasi antara X dengan KU
- Informasi pada KU didominasi oleh informasi X yang memiliki koefisien besar.

Permasalahan Umum dalam AKU

- Penentuan KU menggunakan ‘matriks ragam-peragam’ vs ‘matriks korelasi’
- Penentuan banyaknya KU

Menggunakan matriks korelasi atau ragam peragam?

Secara umum ini adalah pertanyaan yang sulit. Karena tidak ada hubungan yang jelas antara akar ciri dan vektor ciri matriks ragam peragam dengan matriks korelasi, dan komponen utama yang dihasilkan oelh keduanya bisa sangat berbeda. Demikian juga dengan berapa banyak komponen utama yang digunakan.

Menggunakan matriks korelasi atau ragam peragam?

Perbedaan satuan pengukuran yang umumnya berimplikasi pada perbedaan keragaman peubah, menjadi salah satu pertimbangan utama penggunaan matriks korelasi. Meskipun ada juga beberapa pendapat yang mengatakan gunakan selalu matriks korelasi.

Menggunakan matriks korelasi atau ragam peragam?

Penggunaan matriks korelasi memang cukup efektif kecuali pada dua hal.

Pertama, secara teori pengujian statistik terhadap akar ciri dan vektor ciri matriks korelasi jauh lebih rumit.

Kedua, dengan menggunakan matriks korelasi kita memaksakan setiap peubah memiliki ragam yang sama sehingga tujuan mendapatkan peubah yang kontribusinya paling besar tidak tercapai.

Penentuan Banyaknya KU

Metode 1

- didasarkan pada kumulatif proporsi keragaman total yang mampu dijelaskan.
- Metode ini merupakan metode yang paling banyak digunakan, dan bisa diterapkan pada penggunaan matriks korelasi maupun matriks ragam peragam.
- Minimum persentase kergaman yang mampu dijelaskan ditentukan terlebih dahulu, dan selanjutnya banyaknya komponen yang paling kecil hingga batas itu terpenuhi dijadikan sebagai banyaknya komponen utama yang digunakan.
- Tidak ada patokan baku berapa batas minimum tersebut, sebagian bukau menyebutkan 70%, 80%, bahkan ada yang 90%.

Penentuan Banyaknya KU

Metode 2

- hanya bisa diterapkan pada penggunaan matriks korelasi. Ketika menggunakan matriks ini, peubah asal ditransformasi menjadi peubah yang memiliki ragam sama yaitu satu.
- Pemilihan komponen utama didasarkan pada ragam komponen utama, yang tidak lain adalah akar ciri. Metode ini disarankan oleh Kaiser (1960) yang berargumen bahwa jika peubah asal saling bebas maka komponen utama tidak lain adalah peubah asal, dan setiap komponen utama akan memiliki ragam satu.
- Dengan cara ini, komponen yang berpadanan dengan akar ciri kurang dari satu tidak digunakan. Jollife (1972) setelah melakukan studi mengatakan bahwa *cut off* yang lebih baik adalah 0.7.

Penentuan Banyaknya KU

Metode 3

- penggunaan grafik yang disebut plot scree.
- Cara ini bisa digunakan ketika titik awalnya matriks korelasi maupun ragam peragam.
- Plot scree merupakan plot antara akar ciri λ_k dengan k.
- 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. Interpretasi terhadap plot ini sangat subjektif.

Kegunaan Lain KU

- Plot skor KU dua dimensi sebagai alat awal diagnosis pada analisis gerombol
- KU yang saling bebas mengatasi masalah multikolinear dalam analisis regresi

Contoh Penerapan AKU

Ilustrasi berikut menggunakan catatan waktu pada olimpiade Los Angeles tahun 1984 untuk berbagai nomor lari putri di cabang atletik. Ada tujuh nomor yang dicatat, yaitu lari 100 meter, 200 meter, 400 meter, 800 meter, 1500 meter, 3000 meter, dan maraton. Tiga nomor pertama catatan waktu dalam satuan detik, sedangkan empat nomor yang lain dalam menit. Data yang tersedia ada 55 negara peserta.

Masalah yang ingin dipecahkan adalah memeringkatkan negara berdasarkan performa dari keseluruhan nomor. Cara yang paling sederhana sebenarnya adalah dengan cara merata-ratakan catatan ketujuh nomor, setelah terlebih dahulu menyamakan satuan menjadi detik (atau menit). Namun seperti yang dibahas sebelumnya, rata-rata tidak mampu memberikan informasi sebanyak jika menggunakan komponen utama. Pemilihan komponen utama pertama, namapaknya cukup beralasan.

Yang menjadi permasalah dalam penggunaan komponen utama adalah, matriks ragam peragam ataukah matriks korelasi yang harus digunakan untuk mendapatkannya. Perbedaan satuan pada peubah yang ada menyebabkan pemilihan korelasi merupakan ide yang lebih baik. Penggunaan matriks ragam peragam akan menyebabkan dominasi dari catatan di nomor maraton, karena ragamnya paling besar.

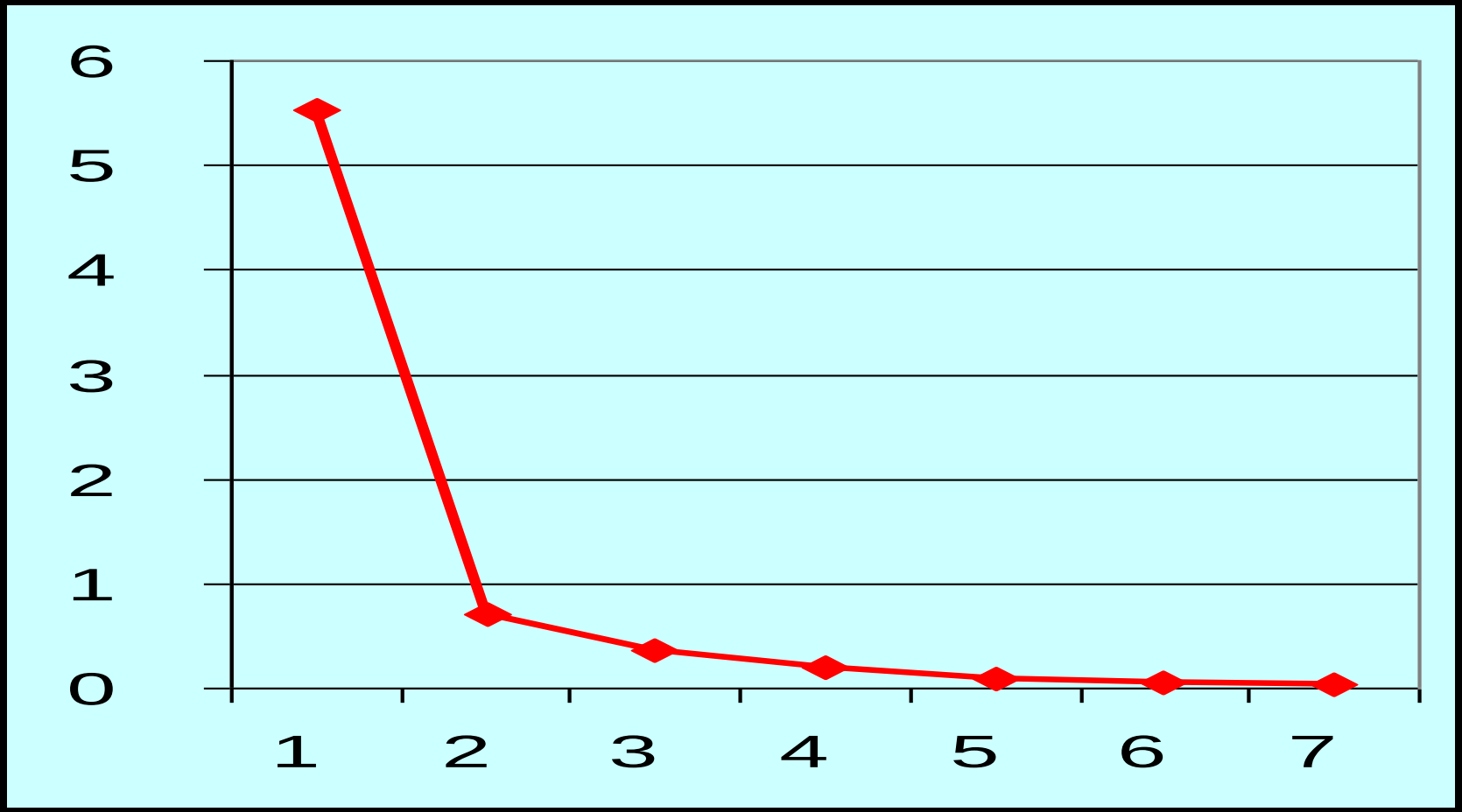
Correlation Matrix

	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

Eigenvalues of the Correlation Matrix

	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

- Metode 1: Menggunakan 2 KU sudah mencapai proporsi keragaman 89.27%
- Metode 2: Hanya 2 KU yang memiliki akarciri lebih besar dari 0.7
- Metode 3: Pada k = 2 terlihat gambar scree plot sangat curam di kiri tapi landai di kanan. Jadi 2 KU yang digunakan sudah mencukupi.

Eigenvectors

	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

Dengan menggunakan matriks korelasi sebagai dasar analisis, diperoleh bahawa akar ciri pertama sebesar 5.53 (yang juga merupakan ragam komponen pertama), dan mampu menerangkan keragaman data sebesar 79.05%. Akar ciri pertama yang berpadanan dengannya adalah

(0.378202, 0.376416, 0.391311, 0.390624, 0.385043, 0.395890, 0.323383)'

memiliki nilai yang semua positif dan hampir sama besar, **bisa diartikan sebagai ukuran performa keseluruhan nomor**.

Perhatikan bahwa karena peubah asalnya adalah catatan waktu di berbagai nomor, maka negara dengan nilai yang lebih kecil merupakan negara yang memiliki pelari lebih cepat.

KU Pertama

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

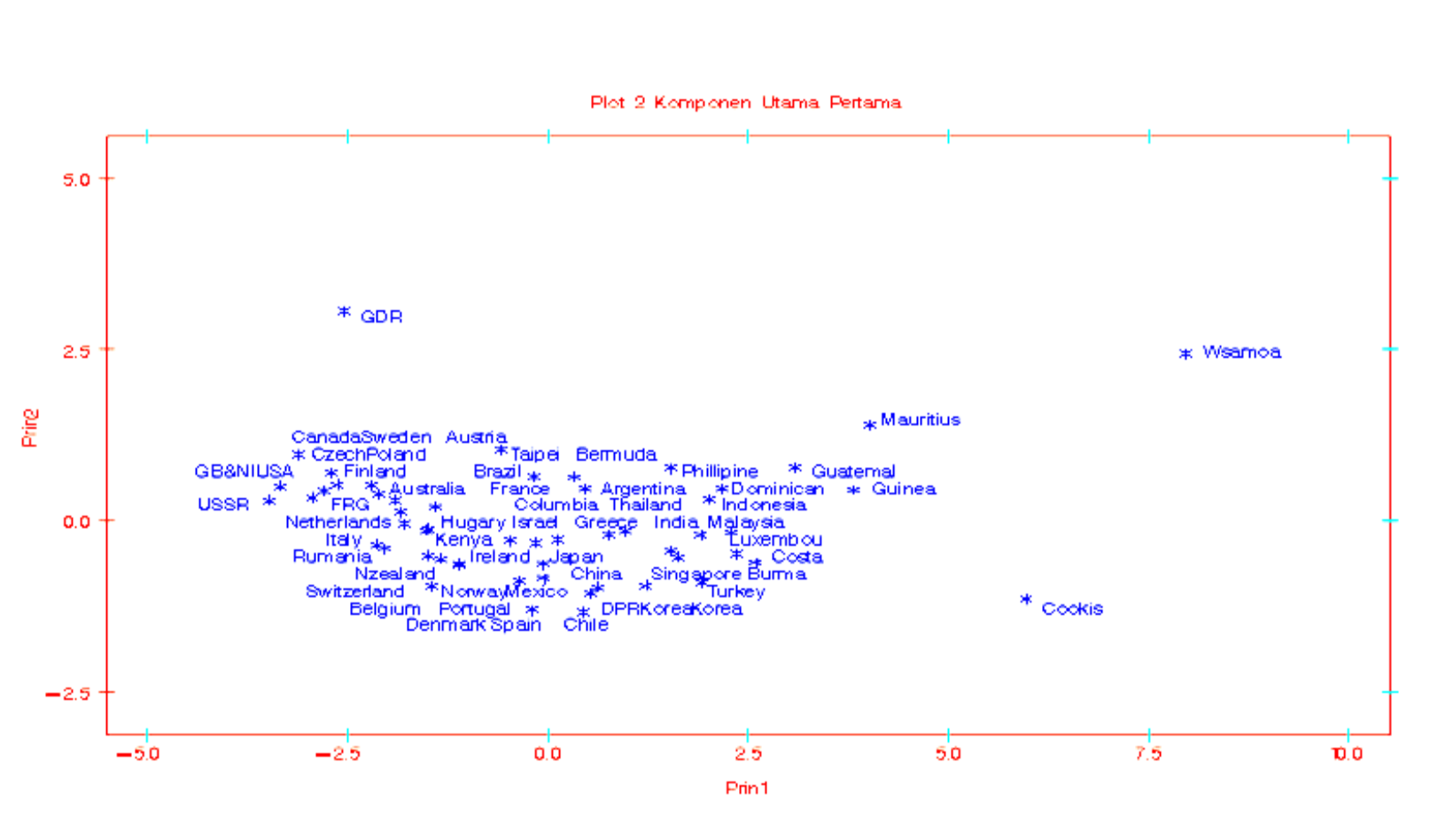
KU Kedua

Komponen utama kedua memiliki ragam sebesar akar ciri terbesar kedua yaitu 0.71 dan memberikan kontribusi keragaman total 10.22%. Sehingga jika digunakan dua komponen utama akan didapatkan keragaman total yang mampu dijelaskan keduanya adalah 89.27%. Akar ciri dari komponen kedua ini adalah

(-.426104, -.452874, -.272232, 0.067673, 0.230072, 0.308242, 0.621855)

Komponen kedua ini bisa diartikan sebagai kontras antara catatan waktu nomor jarak dekat dengan jarak menengah dan jauh. Negara dengan nilai skor komponen utama kedua mendekati nol, berarti memiliki kemampuan yang merata di kedua jenis nomor.

Plot Skor KU


Plot 2 Komponen Utama Pertama
Prin2
Prin1
5.0
2.5
0.0
−2.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
GDR
Wsamoa
Mauritius
CanadaSweden
Austria
CzechPoland
Taipei
Bermuda
GB&NIUSA
Finland
Brazil
Phillipine
Guatemal
Australia
France
Argentina
Dominican
Guinea
USSR
FRG
Columbia
Thailand
Indonesia
Netherlands
Hugary
Israel
Greece
India
Malaysia
Italy
Kenya
Luxembou
Rumania
Ireland
Japan
Costa
Nzealand
China
Singapore
Burma
Switzerland
Norway
Mexico
Turkey
Belgium
Portugal
DPRKorea
Korea
Cookis
Denmark
Spain
Chile

CONTOH APLIKASI REGRESI KOMPONEN UTAMA

REGRESI PENGARUH SIFAT – SIFAT KUANTITATIF PADI SAWAH TERHADAP HASIL

Masalah

Banyak Peubah → Sulit dalam Analisis

Multikolinearitas → Kesimpulan tidak Valid

Langkah-Langkah

- Analisis Hub antar Peubah
- Pemeriksaan Multikolinearitas
- Analisis KU
- Regresi KU dengan Peubah Respon Y
- Transformasi Regresi KU ke Peubah Baku Z
- Transformasi Regresi Z ke Peubah Asal X

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

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

Analisis Komponen Utama

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

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

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