Analisis Komponen Utama (AKU) / *Principal Component Analysis* (PCA)

Dhea Dewanti & Nur Khamidah

Pendahuluan

- Pada kasus data berdimensi besar, data tersusun pada format yang kompleks, susah untuk digali informasinya, serta seringkali memiliki korelasi yang tinggi antar peubahnya
- Oleh karenanya, data berdimensi tinggi dapat disederhanakan dengan dilakukan pereduksian dimensi
- Beberapa masalah yang timbul dalam mereduksi dimensi tersebut adalah bagaimana caranya mendapatkan gugus peubah yang lebih kecil namun tetap mampu mempertahankan sebagian besar informasi yang terkandung pada data asal

Analisis Komponen Utama

- Analisis Komponen Utama / *Principal Component Analysis* (PCA) adalah sebuah teknik untuk menyederhanakan suatu data yang berdimensi besar dan saling berkorelasi menjadi dimensi yang lebih kecil dan saling bebas dengan cara mentransformasi linier peubah-peubah yang diamati membentuk beberapa peubah baru yang dikenal dengan komponen utama (*principal component*).
- Komponen utama dibentuk berdasarkan matriks ragam-peragam atau matriks korelasi.
- Hasil komponen utama merupakan penyederhanaan dari data asal tanpa mengurangi informasi atau keragaman data awal secara signifikan
- PCA merupakan analisis antara (menjadi awalan dari analisis berikutnya)

Analisis Komponen Utama

- Misalkan vektor peubah acak $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ memiliki matriks var-cov $\boldsymbol{\Sigma}$ dan akar-akar ciri $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ kemudian dilakukan kombinasi linier berikut:

$$
\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

- Kemudian diperoleh:

$$
\begin{aligned}
\operatorname{Var}(Y_i) &= \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_i \qquad i = 1, 2, \ldots, p \\
\operatorname{Cov}(Y_i, Y_k) &= \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_k \qquad i, k = 1, 2, \ldots, p
\end{aligned}
$$

- Komponen Utama / *Principal Component* (PC) dari **X** adalah kombinasi linier yang saling bebas dari $Y_1, Y_2, \ldots, Y_p$ yang memiliki ragam (dari matriks var-cov sebelumnya) yang bernilai sebesar-besarnya.

Manfaat Analisis Komponen Utama

- **Eksplorasi posisi objek dan penangan masalah kolinear antar peubah**. Eksplorasi posisi objek diperlukan sebagai alat bantu dalam analisis gerombol. Eksplorasi ini dapat dilakukan dengan membuat plot skor komponen utama pertama dengan kedua. Dari plot ini diharapkan dapat terlihat banyaknya kumpulan objek yang terbentuk, yang pada nantinya dapat digunakan sebagai nilai awal dalam penentuan jumlah gerombol yang akan dibangun.

- Komponen utama merupakan salah satu solusi dalam **mengatasi masalah kolinear**. Penerapan ini relevan dengan sifat dari komponen utama yang dibangun yaitu antar komponen utama bersifat saling orthogonal atau saling bebas. Dengan memanfaatkan komponen utama ini masalah kolinear dalam analisis regresi dapat diatasi. Sifat ini juga diperlukan dalam penerapan jarak Euclid, dimana konsep jarak ini juga memerlukan kebebasan antar peubah.

Pembentukan PC dengan Matriks Ragam-Peragam

- Misalkan **Σ** merupakan matriks ragam peragam dari $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ yang memiliki pasangan akar dan vektor ciri $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ dimana $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, komponen utama ke-*i* dibentuk dengan:

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, \ldots, p$$

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \qquad i = 1, 2, \ldots, p \\ \text{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_k = 0 \qquad i \neq k \end{aligned}$$

- Dengan ketentuan bahwa pembentukkan komponen utama berdasarkan matriks ragam peragam dapat dilakukan jika **satuan pengukuran setiap peubah sama**.

Pembentukan PC dengan Matriks Ragam-Peragam

- Total keragaman populasi diperoleh dengan:

$$\sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \text{Var}(Y_i)$$

$$\begin{aligned} \text{Total population variance} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

- Sedangkan proporsi keragaman kontribusi komponen utama ke-*i* diperoleh dengan:

$$\left(\begin{array}{c} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } k\text{th principal} \\ \text{component} \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p$$

Pembentukan PC dengan Matriks Korelasi

- Misalkan **Σ** merupakan matriks ragam peragam dari $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]$ dengan $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$, dan memiliki pasangan akar-vektor ciri $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ dengan $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

  maka komponen utama ke-*i* diberikan:

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, 2, \ldots, p$$

- Dengan ketentuan bahwa pembentukkan komponen utama berdasarkan matriks korelasi dapat dilakukan jika **satuan pengukuran setiap peubah berbeda**. Standardiasi dilakukan dengan:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \qquad Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \qquad \ldots. \qquad Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

Pembentukan PC dengan Matriks Korelasi

- Total keragaman populasi diperoleh dengan:

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p$$

- Sedangkan proporsi keragaman kontribusi komponen utama ke-*i* diperoleh dengan:

$$\begin{pmatrix} \text{Proportion of (standardized)} \\ \text{population variance due} \\ \text{to } k\text{th principal component} \end{pmatrix} = \frac{\lambda_k}{p}, \qquad k = 1, 2, \ldots, p$$

Tahap Pengerjaan Analisis Komponen Utama

1. Sediakan set data X
2. Membuat matriks varian kovarian atau matriks korelasi dari data X
3. Menghitung akar ciri dan vektor ciri dari matriks data yang ditentukan poin 2
4. Menentukan proporsi keragaman menggunakan akar ciri yang diperoleh pada poin 3
5. Menghitung komponen utama dari vektor ciri yang diperoleh di poin 3 sekaligus menentukan banyaknya komponen yang digunakan melalui proporsi keragaman yang telah diperoleh pada poin 4
6. Menghitung skor komponen utama sebanyak komponen utama yang telah ditentukan pada poin 5

Penentuan Banyak Komponen Utama

Metode 1:

- Metode ini didasarkan pada kumulatif proporsi keragaman total yang mampu dijelaskan.
- Metode ini merupakan metode yang paling banyak digunakan, dan dapat diterapkan pada penggunaan matriks korelasi maupun matriks ragam peragam.
- Minimum persentase keragaman 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 buku menyebutkan 70%, 80%, bahkan ada yang 90%.

Penentuan Banyak Komponen Utama

Metode 2:

- Metode ini hanya dapat 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 Banyak Komponen Utama

Metode 3:

- Metode ini menggunakan grafik yang disebut ***plot scree***.
- Cara ini dapat 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.

Contoh Soal

Diketahui matriks kovarian $S = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{bmatrix}$

A. Tentukan akar ciri dan vektor ciri dari matriks S
B. Tentukan Komponen Utama yang terbentuk melalui matriks S
C. Tunjukkan bahwa ragam dari Komponen Utama pertama sama dengan akar ciri terbesar dari matriks kovarian Σ
D. Apa yang dapat Anda ceritakan dari komponen utama yang dihasilkan?

Penyelesaian

a. Akar ciri dan vektor ciri matriks S

Akar ciri matriks S

$$|S - \lambda I| = 0$$

$$\left| \begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \right| = \left| \begin{pmatrix} 2-\lambda & 0 & 0 \\ 0 & 8-\lambda & 0 \\ 0 & 0 & 4-\lambda \end{pmatrix} \right| = 0$$

$$(2 - \lambda)(8 - \lambda)(4 - \lambda) = 0$$

Sehingga akar ciri dari matriks S adalah:

$$\boldsymbol{\lambda_1 = 8 \ , \lambda_2 = 4 \ dan \ \lambda_3 = 2}$$

Penyelesaian

Vektor ciri matriks S

- Untuk $\lambda_1 = 8$:

$$\boldsymbol{Sx} = \lambda \boldsymbol{x}$$

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 8 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\begin{pmatrix} 2x_1 \\ 8x_2 \\ 4x_3 \end{pmatrix} = \begin{pmatrix} 8x_1 \\ 8x_2 \\ 8x_3 \end{pmatrix}$$

Sehingga diperoleh :

$x_1 = 0$

$x_2 = sembarang$ misalnya diambil nilai 1

$x_3 = 0$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\|x\| = \sqrt{0^2 + 1^2 + 0^2} = 1$$

Dengan demikian, maka vektor ciri matriks S untuk $\lambda_1 = 8$ adalah :

$$\boldsymbol{a_1} = \frac{x}{\|x\|} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Penyelesaian

- Untuk $\lambda_2 = 4$:

$$\boldsymbol{Sx} = \lambda \boldsymbol{x}$$

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 4 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\begin{pmatrix} 2x_1 \\ 8x_2 \\ 4x_3 \end{pmatrix} = \begin{pmatrix} 4x_1 \\ 4x_2 \\ 4x_3 \end{pmatrix}$$

Sehingga diperoleh :

$x_1 = 0$

$x_2 = 0$

$x_3 = sembarang$ misalnya diambil nilai 1

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\|x\| = \sqrt{0^2 + 0^2 + 1^2} = 1$$

Dengan demikian, maka vektor ciri matriks S untuk $\lambda_2 = 4$ adalah :

$$\boldsymbol{a_2} = \frac{x}{\|x\|} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Penyelesaian

- Untuk $\lambda_3 = 2$:

$$\boldsymbol{Sx} = \lambda \boldsymbol{x}$$

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 2 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\begin{pmatrix} 2x_1 \\ 8x_2 \\ 4x_3 \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{pmatrix}$$

Sehingga diperoleh :

$x_1 = sembarang$ misalnya diambil nilai 1

$x_2 = 0$

$x_3 = 0$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\|x\| = \sqrt{1^2 + 0^2 + 0^2} = 1$$

Dengan demikian, maka vektor ciri matriks S untuk $\lambda_3 = 2$ adalah

$$\boldsymbol{a_3} = \frac{x}{\|x\|} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Penyelesaian

b. KU yang terbentuk melalui matriks S

$$KU_1 = a_1'x = (0 \quad 1 \quad 0)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_2$$

$$KU_2 = a_2'x = (0 \quad 0 \quad 1)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_3$$

$$KU_3 = a_3'x = (1 \quad 0 \quad 0)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1$$

c. $\text{Var}\,(KU_1) = a_1'Sa_1 = (0 \quad 1 \quad 0)\begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$$\text{Var}\,(KU_1) = (0 \quad 1 \quad 0)\begin{pmatrix} 0 \\ 8 \\ 0 \end{pmatrix} = 8$$

$\text{Var}\,(KU_1) = \lambda_1 = 8$ *(akar ciri paling besar)*

d. KU1 dibentuk/disusun oleh $x2$ karena akar ciri terbesar (keragaman terbesar pertama) berada di $x2$, kemudian KU2 disusun oleh $x3$ (keragaman terbesar kedua), selanjutnya KU3 disusun oleh $x1$ (keragaman terbesar ketiga).

Studi Kasus

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.

Link Data:

https://docs.google.com/spreadsheets/d/1hR7fM82p2x1EWEhRSESoX7y4y3LEKemx/edit?usp=sharing&ouid=112781542983027743433&rtpof=true&sd=true

Penyelesaian

- Untuk melihat performa 55 negara dari 7 cabang lari sulit hanya dilihat dari rata-rata, karena ketujuh cabang tersebut memiliki satuan yang berbeda dan performanya 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

Link R Code:

https://drive.google.com/file/d/1_7zxLAFX5t3TDSYV-juWWcuaGe2nNCzE/view?usp=sharing

Tugas

- Kerjakan dengan format excel. Setiap soal dan jawaban dalam satu sheet. Artinya ada 4 sheet.
- Beri nama file anda dengan nama file: **paralel_nama_nim**
- Upload jawaban ke: **https://ipb.link/tugas-akhir-uts-sta1342**
- Batas maksimal pengumpulan: **14 Oktober 2023 pukul 23.59**
- Link Soal: **https://ipb.link/soal-tpg**