BAB II

TINJAUAN PUSTAKA

Dalam bab ini akan dipaparkan beberapa teori pendukung yang digunakan dalam proses analisis klaster pada bab selanjutnya.

2.1 DATA MULTIVARIAT

Data yang diperoleh dengan mengukur lebih dari satu variabel kriteria pada setiap individu anggota sampel, disebut data multivariat. Jika terdapat *n* objek dan *p* variabel, maka observasi X_{ji} dengan $j = 1, 2, \dots, n$ dan $i = 1, 2, \dots, p$, dapat ditulis sebagai berikut:

Observasi	Var 1	Var 2	⋯	Var *i*	⋯	Var *p*
Objek 1	X_{11}	X_{12}	⋯	X_{1i}	⋯	X_{1p}
Objek 2	X_{21}	X_{22}	⋯	X_{2i}	⋯	X_{2p}
⋮	⋮	⋮	⋱	⋮	⋱	⋮
Objek *j*	X_{j1}	X_{j2}	⋯	X_{ji}	⋯	X_{jp}
⋮	⋮	⋮	⋱	⋮	⋱	⋮
Objek *n*	X_{n1}	X_{n1}	⋯	X_{ni}	⋯	X_{np}

Data tersebut dapat ditulis sebagai matriks X, dengan n baris dan p kolom.

$$X_{nxp} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1i} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2i} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{j1} & X_{j2} & \dots & X_{ji} & \dots & X_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n1} & \dots & X_{ni} & \dots & X_{np} \end{bmatrix}. \quad (2.1)$$

2.2 ANALISIS KOMPONEN UTAMA

Analisis komponen utama pertama kali digambarkan oleh Karl Pearson (1901) yang meyakini bahwa metode analisis komponen utama merupakan solusi yang benar untuk beberapa masalah yang menjadi perhatian pengamat biometri pada saat itu. Tetapi, dia tidak mengajukan praktek perhitungan secara langsung untuk lebih dari tiga variabel. Selanjutnya, metode analisis komponen utama dikembangkan oleh Hotteling (1933) (Jackson, 1991:6).

Analisis komponen utama merupakan salah satu metode dalam menganalisa suatu data, di mana variabel-variabel yang dianalisa dari data tersebut saling berkorelasi. Tujuan utama dari analisis ini adalah menjelaskan sebanyak mungkin jumlah varians data asli dengan sedikit mungkin komponen utama yang disebut faktor atau komponen. Jika objek analisisnya adalah data dengan p variabel $Y_1, Y_2, \dots, Y_p$ maka, tujuannya adalah menemukan kombinasi linear $K_1, K_2, \dots, K_q$, q adalah banyaknya komponen utama. Apabila K_i adalah komponen utama ke i, maka diperoleh q persamaan sebagai berikut:

$$K_q = w_{q1}Y_1 + w_{q2}Y_2 + \cdots + w_{qi}Y_i + \cdots + w_{qp}Y_p \quad (2.2)$$

2.3 PENENTUAN BANYAKNYA KOMPONEN UTAMA BERDASARKAN *SCREE PLOT*

Scree plot merupakan suatu plot dari nilai eigen untuk menentukan banyaknya komponen utama. *Scree plot* seperti garis patah-patah. Jumlah komponen utama dilihat pada bentuk *scree plot* dan diambil di mana terjadinya patahan (*break*) (Supranto, 2004:129). Jumlah komponen utama digunakan untuk menentukan di dimensi berapakah variabel-variabel akan diplot pada *scatterplot*. Selanjutnya dari plot pada *scatterplot* dapat diidentifikasi ada tidaknya klaster yang terbentuk. Analisis komponen utama juga digunakan untuk mencari *z-score.*

Z-score mentransformasikan *p* variabel yang berkorelasi yaitu $Y_1, Y_2, \dots, Y_p$ ke dalam *p* variabel baru yang tidak berkorelasi yaitu $Z_1, Z_2, \dots, Z_p$. Sumbu koordinat dari variabel baru digambarkan oleh vektor eigen $\boldsymbol{u}_i$ yang merupakan elemen-elemen dari matriks U yang digunakan dalam transformasi:

$$\boldsymbol{z} = U'[\boldsymbol{x} - \overline{\boldsymbol{x}}]$$

di mana $\boldsymbol{x}$ dan $\overline{\boldsymbol{x}}$ adalah vektor *p* x 1 dari observasi variabel asli dan rata-ratanya. Komponen utama ke *i* adalah

$$\boldsymbol{z}_i = \boldsymbol{u}_i'[\boldsymbol{x} - \overline{\boldsymbol{x}}] \qquad (2.3)$$

Rata-ratanya nol dan variansinya adalah λ_i, vektor eigen ke *i* (Jackson, 1991:11).

2.4 VEKTOR *MEAN*, MATRIKS KOVARIANSI DAN MATRIKS KORELASI

2.4.1 Vektor *Mean*, Matriks Kovariansi dan Matriks Korelasi Populasi

Misalkan matriks random $\boldsymbol{X} = \{\boldsymbol{X_i}\}$ berorde $px1$ untuk setiap $i = 1,2,\dots,p$ merupakan sebuah vektor random. Mean dari vektor random $\boldsymbol{X}$ untuk populasi adalah:

$$E(\boldsymbol{X}) = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}.$$

Kovariansi dari vektor random $\boldsymbol{X}$ adalah

$$\sum = E\,(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})' = E\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_p-\mu_p \end{bmatrix}\begin{bmatrix} X_1-\mu_1, X_2-\mu_2, \cdots, X_p-\mu_p \end{bmatrix}\right)$$

$$= E\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2 \end{bmatrix}$$

$$\sum = \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \cdots & E(X_1-\mu_1)(X_p-\mu_p) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_2-\mu_2)^2 & \cdots & E(X_2-\mu_2)(X_p-\mu_p \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)(X_2-\mu_2) & \cdots & E(X_p-\mu_p)^2 \end{bmatrix}$$

$$\sum = Kov(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}.$$

Oleh karena $\sigma_{ik} = \sigma_{ki}$, untuk setiap $i = 1,2,\dots,p$ dan $k = 1,2,\dots,p$ dengan $i \neq k$ maka berlaku:

$$\sum = Kov(\boldsymbol{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}.$$

Merupakan matriks simetris dengan μ dan Σ berturut-turut adalah *mean* populasi dan varians-kovarians populasi.

Ukuran hubungan linear antara variabel random X_i dan X_k disebut koefisien korelasi. Koefisien korelasi populasi didefinisikan sebagai rasio kovariansi σ_{ik} dengan variansi σ_{ii} dan σ_{kk} sehingga

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}.$$

Matriks koefisien korelasi populasi merupakan matriks simetris $\boldsymbol{\rho}$, berorde pxp, di mana:

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \ddots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}. \quad (2.4)$$

2.4.2 Vektor *Mean*, Matriks Kovariansi dan Matriks Korelasi Sampel

Misalkan $X_{11}, X_{12}, \dots, X_{1n}$ adalah *n* pengukuran pada variabel pertama. Rata-rata pengukuran disebut juga rata-rata (*mean*) sampel ditulis dengan $\bar{X}_1$ adalah

$$\bar{X}_1 = \frac{1}{n}\sum_{j=1}^{n} X_{j1}.$$

Secara umum *mean* sampel untuk variabel ke-*i* bila ada *p* variabel dan *n* banyaknya data (Kartiko, 1988:2.4), adalah:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ji}\,;\ i = 1, 2, \dots, p. \quad (2.5)$$

Sehingga vektor *mean* sampel $\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{\boldsymbol{x}}_1 \\ \bar{\boldsymbol{x}}_2 \\ \bar{\boldsymbol{x}}_3 \\ \bar{\boldsymbol{x}}_4 \\ \vdots \\ \bar{\boldsymbol{x}}_p \end{bmatrix}$.

Variansi sampel untuk variabel ke-*i* adalah

$$S_{ii} = var\,(X_i) = \frac{1}{n-1}\sum_{\mathrm{j=1}}^{\mathrm{n}} \left(X_{ji} - \bar{X}_i\right)^2 ;\ \mathrm{i} = 1, 2, \dots, \mathrm{p}. \quad (2.6)$$

Sedangkan kovariansi sampel untuk variabel ke-*i* dan ke-*k* adalah

$$S_{ik} = cov(X_{i,}X_k) = \frac{1}{n-1}\sum_{\mathrm{j=1}}^{\mathrm{n}}(X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k) \tag{2.7}$$

dan matriks varians dan kovarians sampel

$$S_n = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}. \tag{2.8}$$

Koefisien korelasi sampel merupakan ukuran hubungan linear antara 2 variabel. Koefisien korelasi sampel untuk variabel ke-*i* dan ke-*k* adalah:

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{n}}(X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)}{\sqrt{\sum_{\mathrm{j=1}}^{\mathrm{n}}(X_{ji} - \bar{X}_i)^2}\sqrt{\sum_{\mathrm{j=1}}^{\mathrm{n}}(X_{jk} - \bar{X}_k)^2}}; \tag{2.9}$$

$i = 1, 2, \ldots, \mathrm{p}, k = 1, 2, \ldots, \mathrm{p},\ r_{ik} = r_{ki}$ untuk setiap *i* dan *k*.

Sehingga diperoleh matriks korelasi sampel

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}. \tag{2.10}$$

2.5 VARIANSI, *SUM OF SQUARE* DAN *CROSS PRODUCTS*

Variansi untuk variabel ke-*i* diberikan oleh:

$$s_i^2 = \frac{\sum_{j=1}^{n} x_{ji}^2}{n-1} = \frac{SS(sum\ of\ square)}{df}$$

di mana x_{ji} adalah *mean-corrected* data untuk observasi ke-*j* dan variabel ke-*i* dan n adalah banyaknya observasi. *Mean-corrected* data diperoleh dari $(X_{ji} - \bar{X}_i)$ dengan X_{ji} adalah observasi ke-*j* dan variabel ke-*i* dan $\bar{X}_i$ adalah rata-rata variabel ke-i

Hubungan linear antar dua perbandingan dapat diukur melalui kovarians antar dua variabel yang diberikan oleh:

$$s_{ik} = \frac{\sum_{j=1}^{n} x_{ji}x_{jk}}{n-1} = \frac{SCP(sum\ of\ the\ cross\ products)}{df}$$

di mana s_{ik}adalah kovarians antara variabel ke-i dan variabel ke-*k* dan x_{jk} adalah *mean-corrected* data untuk observasi ke-j dan variabel ke-*k.*

Gabungan dari SS dan SCP disebut *sum of square and cross products* ($SSCP_t$) matriks, dan varians kovarians disebut kovarians (S) matriks.

$$SSCP_t = \begin{bmatrix} SS_1 & SCP_{12} & \cdots & SCP_{1p} \\ SCP_{21} & SS_2 & \cdots & SCP_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ SCP_{p1} & SCP_{p2} & \cdots & SCP_{pp} \end{bmatrix}$$

$$S_t = \frac{SSCP_t}{df}$$

Within-Group Analysis ($SSCP_w$)

$$SSCP_w = SSCP_1 + SSCP_2 + \cdots + SSCP_g$$

di mana g adalah banyaknya klaster.

Between-Group Analysis ($SSCP_b$)

$$SSCP_b = SSCP_t - SSCP_w.$$

2.6 UKURAN JARAK EUCLID

Dalam analisis klaster, pengelompokan data atau permasalahan dibutuhkan suatu ukuran yang dapat menerangkan kedekatan antara data. Keanggotan dalam suatu klaster ditentukan berdasarkan ukuran kesamaan atau similaritas.

Sesuai dengan prinsip dasar klaster yaitu mengelompokkan objek yang mempunyai kesamaan, maka proses pertama adalah mengukur seberapa jauh ada kesamaan antar objek. Konsep ukuran kesamaan yang digunakan adalah ukuran jarak, di mana ukuran jarak yang biasa digunakan adalah ukuran jarak Euclid. Misalkan $d(x, y)$ sebuah fungsi jarak dari pasangan titik di E disebut metriks jika memenuhi kondisi berikut :

i. $d(x, y) \geq 0$.

ii. $d(x, y) = 0$, jika hanya jika $x = y$.

iii. $d(x, y) = d(y, x)$.

iv. $d(x,z) \leq d(x,y) + d(y,z) \quad \forall\, x,y,z \in E.$

Sebagian besar teknik multivariat didasarkan pada konsep sederhana dari jarak. Misalkan titik $P = (x_1, x_2)$ di bidang, maka jarak Euclid dari P ke titik asal $O = (0,0)$ adalah berdasarkan teorema *Phytagoras* $(O,P) = \sqrt{x_1^2 + x_2^2}$. Secara umum jika titik P mempunyai p koordinat ditulis $P = (x_1, x_2, \dots, x_p)$, maka jarak Euclid dari P ke titik asal $O = (0, 0, \dots, 0)$ adalah

$$d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}\ .$$

Persamaan tersebut adalah sebuah kuadrat jarak

$$d^2(O,P) = x_1^2 + x_2^2 + \cdots + x_p^2 = c^2.$$

Jarak Euclid antara dua titik P dan Q dengan koordinat $P = (x_1, x_2, \dots, x_p)$ dan $Q = (y_1, y_2, \dots, y_p)$ yaitu:

$$\mathrm{d(P,Q)} = \sqrt{(\mathrm{x_1} - \mathrm{y_1})^2 + (\mathrm{x_2} - \mathrm{y_2})^2 + \cdots + (\mathrm{x_p} - \mathrm{y_p})^2}.$$

Jarak Euclid antara dua observasi di $\mathrm{R^p}$ yaitu antar observasi $X = [x_1, x_2, \dots, x_p]'$ dan $Y = [y_1, y_2, \dots, y_p]'$ yaitu:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

$$= \sqrt{(x-y)'(x-y)}$$

$$d_{jl} = \sqrt{\sum_{k=1}^{i} \left(x_{lk} - x_{jk}\right)^2} \; i = 1, 2, \ldots, p. \tag{2.11}$$

Keterangan: d_{jl}= Kuadrat jarak Euclid antar objek ke-j dengan objek ke-*l*.

x_{lk}= Nilai dari objek ke-*l* pada variabel ke-*k*.

x_{jk}= Nilai dari objek ke-*j* pada variabel ke-*k*.

Rumus tersebut digunakan untuk menghitung matriks jarak dari data multivariat yang diberikan. Diperoleh:

$$D_{nxn} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}. \tag{2.12}$$

2.7 *OUTLIERS*

Sampel yang digunakan dalam analisis klaster harus representatif artinya sampel yang dipilih harus mewakili populasi yang ingin dijelaskan dan bahwa semua sifat dalam populasi ada pada sampel yang akan diteliti. Pentingnya representatif karena dalam penelitian harus dapat menghasilkan analisis yang baik. Akibat dari sampel yang harus representatif adalah perlu dideteksi ada atau tidaknya *outlier* pada data yang akan diolah. *Outliers* dapat dilihat dengan membandingkan diagonal utama jarak *Mahalanobis* dengan rumus:

$$d_j^2 = (\boldsymbol{x_j} - \bar{\boldsymbol{x}})' S^{-1} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}), j = 1, 2, \dots, n \quad (2.13)$$

dengan $\chi^2_{(\alpha,p)}$, $\alpha < 0{,}001$ dan p= banyaknya variabel.

Untuk menguji *outliers* akan dibuat hipotesis sebagai berikut:

H_0: Data bebas *outliers*

H_1: Data mengandung *outliers*

Statistik uji:

Statistik yang digunakan adalah jarak Mahalanobis dan $\chi^2_{\alpha,p}$, dengan p adalah banyak variabel.

Kriteria pengujian:

Terima H_0 jika jarak Mahalanobis $< \chi^2_{\alpha,p}$ dengan $\alpha < 0{,}001$.

2.8 KOLINEARITAS

Kolinearitas adalah kondisi dalam sekumpulan data yang memiliki dua variabel yang saling berhubungan (berkorelasi). Untuk mendeteksi kolinearitas dapat dilihat pada tabel korelasi dari output S-PLUS 2000.

Menurut Sarwono (2007:35) terdapat kriteria kolerasi sebagai berikut:

- 0 – 0,25 : korelasi sangat lemah (dianggap tidak ada).
- >0,25 – 0,5 : korelasi cukup.
- >0,5 – 0,75 : korelasi kuat.
- >0,75 – 1 : korelasi sangat kuat.

Selanjutnya timbul suatu permasalahan bagaimana mengatasi masalah kolinearitas. Menurut Sembiring, (1995:186) masalah kolinearitas diatasi dengan menggunakan analisis komponen utama. Dalam tugas akhir ini, komponen utama yang digunakan adalah komponen utama *z-score* yaitu mentransformasikan observasi menjadi nxp dengan n banyaknya observasi dan p banyaknya variabel.