ANALISIS

KOMPONEN UTAMA

Diajukan Untuk Memenuhi Salah Satu Tugas Mata Kuliah Analisis Multivariat

Disusun oleh:

Novitri Simanjuntak (055813)

Dwi Melani P. (055519)

Nurul Kurniawati (041248)

Dena Rahayu (055521)

Naomi Nessyana (055589)

Jurusan Pendidikan Matematika

Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

2009

KATA PENGANTAR

Segala puji bagi Allah SWT yang telah memberikan rahmat, ridho serta kasih sayangnya terhadap umat-Nya sehingga makalah yang berjudul "***ANALISIS KOMPONEN UTAMA***" dapat terselesaikan tepat pada waktunya.

Makalah ini disusun sebagai salah satu tugas untuk mata kuliah Metode Statistika Multivariat. Penulis menyadari betul bahwa masih banyak terdapat kekurangan dalam bentuk penulisan makalah ini. Untuk itu adanya saran dan pendapat serta masukan-masukan yang membangun demi perbaikan makalah ini sangat penulis harapkan.

Pada kesempatan ini penulis menghaturkan terima kasih kepada Bapak Drs. Jarnawi M.kes yang telah membantu dan mendukung dalam pembuatan makalah ini.

Akhir kata, penulis berharap kiranya makalah ini dapat bermanfaat bagi perkembangan Ilmu Pengetahuan Matematika khusunya bidang Statistika sekarang dan pada masa yang akan datang.

Bandung, Juni 2009

Penulis

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pada dasarnya analisis komponen utama bertujuan menerangkan struktur varians-kovarians melalui kombinasi linear dari variabel-variabel. Secara umum analisis komponen utama bertujuan untuk mereduksi data dan menginterpretasikannya. Meskipun dari p buah variabel dasar dapat diturunkan p buah komponen utama untuk menerangkan keragaman total sistem, namun seringkali keragaman total itu dapat diterangkan secara memuaskan oleh sejumlah kecil komponen utama, katakanlah oleh k buah komponen utama, dimana $k < p$. jika demikian halnya, maka kita akan memperoleh bagian terbesar informasi tentang struktur varians-kovarians dari p buah variabel asal itu dalam k buah komponen utama. Dalam hal ini k buah komponen utama dapat mengganti p buah variabel asal serta kumpulan data asli dalam bentuk matriks berukuran n x p dapat direduksi ke dalam matriks berukuran lebih kecil yang mengandung n pengukuran pada k buah komponen utama (matriks berukuran n x k, dimana $k < p$).

Analisis komponen utama sering kali dilakukan tidak saja merupakan akhir dari suatu pekerjaan pengolahan data tetapi juga merupakan tahap (langkah) antara dalam kebanyakan penelitian yang bersifat lebih besar (luas). Analisis komponen utama merupakan tahap antara karena komponen utama dipergunakan sebagai input dalam membangun analisis regresi, demikian pula dalam analisis

gerombol (*cluster analysis*) komponen utama dipergunakan sebagai input untuk melakukan pengelompokan.

1.2 Rumusan Masalah

Untuk memudahkan dalam mengemukakan permasalahan dan mengarahkan pembahasan, maka penulis merumuskan masalahnya sebagai berikut :

1. Bagaimana komponen utama untuk populasi?
2. Bagaimana variasi sampel dengan menggunakan komponen utama?
3. Bagaimana menginterpretasikan komponen utama dalam suatu grafik?
4. Bagaimana analisis komponen utama di dalam sampel ukuran besar?

1.3 Batasan Masalah

Dalam makalah ini, penulis akan membatasi masalah pada analisis komponen utama saja.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini secara umum adalah untuk memperkenalkan dan mengkaji tentang metode Komponen Utama yang di uraikan sebagai berikut:

1. Untuk mengetahui komponen utama pada populasi.
2. Untuk mengetahui nilai variasi sampel dengan menggunakan komponen utama.
3. Untuk mengetahui interpretasi komponen utama dalam suatu grafik.
4. Untuk mengetahui analisis komponen utama dalam sampel ukuran besar.

1.5 Sistematika Penulisan

Sistematika penulisan dalam makalah ini adalah sebagai berikut :

BAB I : Merupakan pendahuluan mencakup latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, serta sistematika penulisan.

BAB II : Mengemukakan

BAB III : Kesimpulan dan saran.

1.6 Daftar Pustaka

Johnson, Richard A. *Applied Multivariate Statistical Analysis*. Prentice Hall.

BAB II

ISI

Novitri Simanjuntak

055813

2.1 Komponen Utama Populasi

Secara aljabar, komponen utama adalah kombinasi linear khusus dari p variabel acak $X_1, X_2, ..., X_p$. Secara geometris, kombinasi linear ini menggambarkan pemilihan dari sistem koordinat yang diperoleh dengan merotasikan sistem awal dengan $X_1, X_2, ..., X_p$ sebagai sumbu koordinat. Seperti yang kita lihat, komponen utama semata-mata bergantung pada matriks kovarians Σ (atau matriks korelasi ρ) dari $X_1, X_2, ..., X_p$. dalam perkembangannya tidak membutuhkan asumsi multivariat normal. Di sisi lain, komponen utama yang berasal dari populasi multivariate normal mempunyai interpretasi yang berguna dalam kepadatan ellipsoid konstan.

Misalkan vektor acak $X' = \left[X_1, X_2, ..., X_p \right]$ memiliki matriks kovarians Σ dengan nilai eigen $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Perhatikan kombinasi linear

$$\begin{aligned} Y_1 &= \ell'_1 X = \ell_{11} X_1 + \ell_{21} X_2 + ... + \ell_{p1} X_p \\ Y_2 &= \ell'_2 X = \ell_{12} X_1 + \ell_{22} X_2 + ... + \ell_{p2} X_p \end{aligned} \qquad \text{(8-1)}$$

. .

. .

. .

$$Y_p = \ell'_p X = \ell_{1p} X_1 + \ell_{2p} X_2 + ... + \ell_{pp} X_p$$

Dengan menggunakan 2-45,

$$Var(Y_1) = \ell'_i \Sigma \ell_i \qquad (8\text{-}2)$$

$$Cov(Y_i, Y_k) = \ell'_i \Sigma \ell_k \qquad (8\text{-}3)$$

komponen utama adalah kombinasi linear $Y_1, Y_2, ..., Y_p$ dimana variansi pada (8-2) sebesar mungkin.

Komponen utama pertama adalah kombinasi linear dengan variansi maksimum. Yang memaksimumkan $Var(Y_1) = \ell'_1 \Sigma \ell_1$. Jelas $Var(Y_1) = \ell'_1 \Sigma \ell_1$ dapat meningkat dengan mengalikan ℓ_1 dengan konstanta. Berdasarkan kenyataan di atas, maka dapat dibuat pernyataan umum yang berkaitan dengan konsep analisis komponen utama, sebagai berikut:

Komponen utama ke-i = kombinasi linear $\ell'_i X$ yang memaksimumkan

$Var(\ell'_i X)$ serta $\ell'_i \ell_i = 1$ dan

$Cov(\ell'_i X, \ell'_k X) = 0$ untuk $k < i$

Result 8.1. Misalkan Σ matriks kovarian yang bersesuaian dengan vektor acak $X' = \left[X_1, X_2, ..., X_p \right]$. Misalkan Σ memiliki pasangan nilai eigen- vektor eigen $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ dimana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Komponen uama ke-I diberikan oleh

$$Y_i = e'_i X = e_{1i} X_1 + e_{2i} X_2 + ... + e_{pi} X_p, \qquad i = 1,2,\ldots,p \qquad (8\text{-}4)$$

Dengan,

$$Var(Y_i) = e'_i \Sigma e_i = \lambda_i \qquad i = 1, 2, ..., p$$

$$Cov(Y_i, Y_k) = e'_i \Sigma e_k = 0 \qquad i \neq k \qquad (8\text{-}5)$$

Jika beberapa λ_i sama, dengan vektor koefisien e_i yang bersesuaian, maka Y_i tdak tunggal.

Bukti. Kita tahu dari (2-51) bahwa $B = \Sigma$,

$$\max_{\ell \neq 0} \frac{\ell' \Sigma \ell}{\ell' \ell} = \lambda_1 \qquad (\text{diperoleh ketika } \ell = e_1)$$

$e'_1 e_1 = 1$ karena vektor eigen dinormalkan. Dengan demikian

$$\max_{\ell \neq 0} \frac{\ell' \Sigma \ell}{\ell' \ell} = \lambda_1 = \frac{e'_1 \Sigma e_1}{e'_1 e_1} = e'_1 \Sigma e_1 = Var(Y_i)$$

Dengan cara yang sama, menggunakan (2-45)

$$\max_{\ell \perp e_1, e_2, ..., e_k} \frac{\ell' \Sigma \ell}{\ell' \ell} = \lambda_{k+1} \qquad k = 1,2,\ldots,p-1$$

Untuk $\ell = e_{k+1}$, dengan $e'_{k+1} e_i = 0$, untuk $i = 1,2,\ldots,k$ *dan* $k = 1,2,..,p-1$,

$$\frac{e'_{k+1} \Sigma e_{k+1}}{e'_{k+1} e_{k+1}} = e'_{k+1} \Sigma e_{k+1} = Var(Y_{k+1})$$

Karena $e'_{k+1}(\Sigma e_{k+1}) = \lambda_{k+1} e'_{k+1} e_{k+1} = \lambda_{k+1}$ maka $Var(Y_{k+1}) = \lambda_{k+1}$.tinggal menunjukkan bahwa e_i tegak lurus terhadap e_k ($e'_i e_k = 0, i \neq k$) memberikan $Cov(Y_i, Y_k) = 0$. Vektor eigen dari Σ orthogonal jika semua nilai eigen $\lambda_1, \lambda_2, ..., \lambda_p$ berbeda.jika nilai eigen tidak berbeda semuanya, maka vektor eigen yang bersesuaian dengan nilai eigen dapat dipilih supaya orthogonal. Dengan demikian, untuk setiap dua vektor eigen e_i dan e_k, $e'_i e_k = 0$,

$i \neq k$. Karena $\Sigma e_k = \lambda_k e_k$, perkalian dengan e'_i memberikan

$Cov(Y_i, Y_k) = e'_i \Sigma e_k = e'_i \lambda_k e_k = \lambda_k e'_i e_k = 0$ untuk setiap $i \neq k$.

$\therefore$ terbukti.

Dari akibat 8.1, komponen utama tidak berkorelasi dan memiliki variansi sama dengan nilai eigen dari Σ.

Result 8.2. Misalkan $X' = \left[X_1, X_2, ..., X_p \right]$ memiliki matriks kovarians Σ, dengan pasangan nilai eigen-vektor eigen $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ dimana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Misalkan $Y_1 = e'_1 X, Y_2 = e'_2 X, ..., Y_p = e'_p X$ adalah komponen utama. Maka

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

Bukti. Dari definisi 2A.28, $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = tr(\Sigma)$. Dari (2-20) dengan $A = \Sigma$, kita dapat menulis $\Sigma = P \Lambda P'$ dimana Λ adalah matriks diagonal dari nilai eigen dan $P = \left[e_1, e_2, ..., e_p \right]$ sedemikian sehingga $PP' = P'P = I$. dengan menggunakan result 2A.12(c), maka diperoleh

$$tr(\Sigma) = tr(P \Lambda P') = tr(\Lambda P' P) = tr(\Lambda) = \lambda_1 + \lambda_2 + ... + \lambda_p$$

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

Result 8.2 mengatakan

Total variansi populasi = $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \lambda_1 + \lambda_2 + ... + \lambda_p$ (8-6)

Dan sebagai akibatnya, proporsi variansi total dari komponen utama ke-*k* adalah

$$\begin{pmatrix} proporsi\ \text{var}\ iansi \\ populasitotaldari \\ komponenutama \\ ke-k \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} \qquad k = 1,2,\ldots,p \qquad (8\text{-}7)$$

Misal apabila p berukuran besar, sedangkan diketahui bahwa sekitar 80% - 90% variansi populasi total telah mampu diterangkan oleh satu, dua, atau tiga komponen utama yang pertama, maka komponen-komponen utama itu telah dapat mengganti p buah varabel asal tanpa mengurangi informasi yang banyak.

Setiap komponen dari vektor koefisien $e'_i = \left[e_{1i}, ..., e_{ki}, ..., e_{pi} \right]$ juga harus diperiksa. Besar e_{ki} diukur dari variabel ke-*k* ke komponen utama ke-i, tanpa memperhatikan variabel yang lain. Secara khusus e_{ki} proporsional terhadap koefisien korelasi antara Y_i dan X_k.

Result 8.3. Misalkan $Y_1 = e'_1 X, Y_2 = e'_2 X, ..., Y_p = e'_p X$ adalah komponen utama yang diperoleh dari matriks kovarian Σ, maka

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2,\ldots, p \qquad (8\text{-}8)$$

adalah koefisien korelasi antara komponen Y_i dan variabel X_k. Disini $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ adalah pasangan nilai eigen – vektor eigen dari Σ.

Bukti. Ambil $\ell'_k = \left[0, ..., 0, 1, 0, ..., 0 \right]$ sedemikian sehingga berdasarkan (2-45) $X_k = \ell'_k X$ dan $Cov(X_k, Y_i) = Cov(\ell'_k X, e'_i X) = \ell'_k \Sigma e_i$. Karena $\Sigma e_i = \lambda_i e_i$, $Cov(X_k, Y_i) = \ell'_k \lambda_i e_i = \lambda_i e_{ki}$.

Maka $Var(Y_i) = \lambda_i$ [lihat (8-5)] dan $Var(X_k) = \sigma_{kk}$ menghasilkan:

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p$$

Contoh 8.1

Misalkan variabel acak X_1, X_2, dan X_3 memiliki matriks kovarian

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Maka didapat pasangan nilai eigen – vektor eigen adalah

$$\lambda_1 = 5,83 \qquad e'_1 = [0,383; -0924; 0]$$

$$\lambda_2 = 2,00 \qquad e'_2 = [0,0,1]$$

$$\lambda_3 = 0,17 \qquad e'_3 = [0,924; 0,383; 0]$$

Sehingga komponen utamanya adalah

$$Y_1 = e'_1 X = 0,383X_1 - 0,924X_2$$
$$Y_2 = e'_2 X = X_3$$
$$Y_3 = e'_3 X = 0,924X_1 + 0,383X_2$$

Variabel X_3 adalah slah satu komponen utama karena tidak berkorelasi dengan dua variabel lainnya.

Persamaan (8-5) dapat ditunjukkan dari komponen utana pertama. Contoh:

$$Var(Y_1) = Var(0,383X_1 - 0,924X_2)$$

$$= (0,383)^2 Var(X_1) + (-0,924)^2 Var(X_2) + 2(0,383)(-0,924)Cov(X_1, X_2)$$

$$= 0,147(1) + 0,854(5) - 0,708(-2)$$

$$= 5,83 = \lambda_1$$

$$Cov(Y_1, Y_2) = Cov(0,383X_1 - 0,924X_2, X_3)$$

$$= 0,383Cov(X_1, X_3) - 0,924Cov(X_2, X_3)$$

$$= 0,383(0) - 0,924(0) = 0$$

Juga dapat ditunjukkan bahwa

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5,83 + 2,00 + 0,17$$

seperti yang ditunjukkan oleh persamaan (8-6). Proporsi variansi total untuk komponen utama pertama adalah $\lambda_1 \big/ (\lambda_1 + \lambda_2 + \lambda_3) = 5,83/8 = 0,73$. Proporsi untuk komponen utama kedua adalah $(5,83 + 2)/8 = 0,98$ dari variansi populasi. Dalam hal ini komponen Y_1 dan Y_2 dapat mengganti ketiga variabel asal tanpa mengurangi informasi yang banyak.

Akhirnya, dengan menggunakan (8-8)

$$\rho_{Y_1, X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0,383\sqrt{5,83}}{\sqrt{1}} = 0,925$$

$$\rho_{Y_1, X_2} = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0,924\sqrt{5,83}}{\sqrt{5}} = -0,998$$

Juga $\rho_{Y_2 X_1} = \rho_{Y_2 X_2} = 0$ dan $\rho_{Y_2 X_3} = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$

Korelasi lainnya dapat diabaikan karena komponen ke-3 tidak dipergunakan.

Misalkan **X** berdistribusi $N_p(\mu, \Sigma)$. Kita tahu dari (4-7) bahwa kepadatan dari X adalah konstanta pada ellipsoid yang berpusat di μ

$$(x - \mu)' \Sigma^{-1} (x - \mu) = c^2$$

dengan sumbu $\pm c\sqrt{\lambda_i e_i}, i = 1, 2, ..., p$, dimana (λ_i, e_i) adalah pasangan nilai eigen-vektor eigen dari Σ. Titik A yang berada pada sumbu ke-*i* dari ellipsoid akan memiliki proporsional koordinat terhadap $e'_i = \left[e_{1i}, ..., e_{ki}, ..., e_{pi} \right]$ dalam sistem koordinat dengan titik asal μ dan sumbu yang sejajar dengan sumbu awal $x_1, x_2, ..., x_p$. Adalah tepat untuk mengambil μ = 0.

Dari bab 2.3 dengan $A = \Sigma^{-1}$, kita dapat menulis

$$c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e'_1 x)^2 + \frac{1}{\lambda_2}(e'_2 x)^2 + ... + \frac{1}{\lambda_p}(e'_p x)^2$$

dimana $e'_1 x, e'_2 x, ..., e'_p x$

adalah komponen utama dari x. Ambil $y_1 = e'_1 x, y_2 = e'_2 x, ..., y_p = e'_p x$, maka didapat

$$c^2 = \frac{1}{\lambda_1} y_1^{\,2} + \frac{1}{\lambda_2} y_2^{\,2} + ... + \frac{1}{\lambda_p} y_p^{\,2}$$

dan persamaan ini didefinisikan oleh sebuah ellipsoid (dengan $\lambda_1, \lambda_2, ..., \lambda_p$ positif) pada sistem koordinat dengan sumbu $y_1, y_2, ..., y_p$ terletak dengan arah $e_1, e_2, ..., e_p$ secara berurutan. Jika λ_1 adalah nilai eigen terbesar, maka sumbu utama terletak pada arah e_1. Sisanya terletak pada arah $e_2, ..., e_p$.

Secara singkat, komponen utama $y_1 = e'_1 x, y_2 = e'_2 x, ..., y_p = e'_p x$ terletak dengan arah sumbu kepadatan ellipsoid konstan. Sehingga, setiap titik pada sumbu ellipsoid ke-*i* proporsional koordinat **x** dengan $e'_i = \left[e_{1i}, e_{2i}, ..., e_{pi} \right]$ dan koordinat komponen utama dengan bentuk $\left[0, ..., 0, y_i, 0, ..., 0 \right]$.

Dwi Melani P.

055519

Komponen Utama yang Diperoleh dari Variabel yang Dibakukan

Komponen utama dapat juga diperoleh untuk variabel yang dibakukan

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

(8-9)

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

Persamaan transformasi Z (persamaan 8-9) dapat dinyatakan secara singkat dalam bentuk matriks,

$$Z = (V^{1/2})^{-1}(X - \mu) \qquad (8\text{-}10)$$

Dimana matriks diagonal simpangan baku $V^{1/2}$ didefinisikan di (2-35) yaitu :

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Dengan jelas $E(Z) = 0$ dan $Cov(Z) = (V^{1/2})^{-1} \sum (V^{1/2})^{-1} = \rho$ oleh (2-37) yaitu :

$V^{1/2} \rho V^{1/2} = \sum$ dan $\rho = \left(V^{1/2}\right)^{-1} \sum \left(V^{1/2}\right)^{-1}$

Komponen utama dari Z mungkin diperoleh dari vektor eigen matriks korelasi ρ dari X. Semua hasil yang sebelumnya berlaku, tapi dengan beberapa

penyederhanaan karena variansi dari setiap Z_i adalah unity(kesatuan). Kita dapat tetap menggunakan notasi Y_i untuk mengacu pada komponen utama ke-*i* dan (λ_i, e_i) untuk pasangan nilai eigen-vektor eigen. Akan tetapi, nilai yang didapat dari $\sum$, secara umum, tidak sama seperti yang didapat dari ρ.

Hasil 8.4. Komponen utama ke-*i* dari variabel baku (variabel asal yang dibakukan satuan pengukurannya) $Z' = [Z_1, Z_2, ..., Z_p]$, dengan $Cov(Z) = \rho$, diberikan oleh

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

Selain itu,

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p \qquad (8\text{-}11)$$

Dan

$$\rho_{Y_i, Z_k} = e_{ki}\sqrt{\lambda_i}, \quad i, k = 1, 2, ..., p$$

Dalam hal ini, $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ adalah sebagai pasangan-pasangan nilai eigen-vektor eigen untuk ρ dengan $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Bukti. Hasil 8.4 mengikuti dari hasil 8.1, 8.2, dan 8.3, dengan $Z_1, Z_2, ..., Z_p$ sebagai pengganti $X_1, X_2, ..., X_p$ dan ρ sebagai pengganti $\sum$.

Kita lihat dari (8-11) bahwa total (variabel baku) variansi populasinya adalah *p*, jumlah elemen-elemen diagonal matriks ρ. Gunakan (8-7) dengan Z sebagai pengganti X, proporsi dari total variansi yang dijelaskan oleh komponen utama ke-*k* dari Z adalah

$$\begin{pmatrix} \text{Proporsi dari (baku)} \\ \text{variansi populasi seharusnya} \\ \text{untuk komponen utama ke-}k \end{pmatrix} = \frac{\lambda_k}{p}, \quad k = 1, 2, ..., p \qquad (8\text{-}12)$$

Dimana λ_k's adalah nilai eigen dari ρ.

Contoh 8.2 (Komponen Utama yang Diperoleh dari Matriks Kovarians dan Korelasi)

Anggaplah matriks kovarians

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

Dan matriks korelasi yang didapat

$$\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

*untuk mencari nilai eigen, digunakan rumus :

$$|\Sigma - \lambda \mathrm{I}| = 0$$

$$\Rightarrow \left| \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\Rightarrow \begin{vmatrix} 1-\lambda & 4 \\ 4 & 100-\lambda \end{vmatrix} = 0$$

$$\Rightarrow \big((1-\lambda)(100-\lambda)\big) - (4)(4) = 0$$

$$\Rightarrow 100 - \lambda - 100\lambda + \lambda^2 - 16 = 0$$

$$\Rightarrow \lambda^2 - 101\lambda + 84 = 0$$

$$\lambda_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\Rightarrow \lambda_{1,2} = \frac{101 \pm \sqrt{(-101)^2 - 4(1)(84)}}{2(1)}$$

$$\Rightarrow \lambda_{1,2} = \frac{101 \pm 99.32270637}{2}$$

$$\lambda_1 = \frac{101 + 99.32270637}{2} = 100.1613532 \approx 100.16$$

dan

$$\lambda_2 = \frac{101 - 99.32270637}{2} = 0.838646815 \approx 0.84$$

*Untuk mencari vektor eigen, digunakan rumus :

Jika $Ax = \lambda x$, maka vektor eigennya adalah $\mathrm{e} = \frac{x}{\sqrt{x'x}}$

$A = \sum = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ dan $\lambda_1 = 100.16, \lambda_2 = 0.84$, maka

$$\sum x = \lambda_1 x \qquad\qquad \sum x = \lambda_2 x$$

$$\begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 100.16\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \ldots(1) \quad \text{dan} \quad \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.84\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \ldots(2)$$

Dari persamaan 1, diperoleh :

$$x_1 + 4x_2 = 100.16x_1$$
$$4x_1 + 100x_2 = 100.16x_2$$

Ambil $x_1 = 1$ (sembarang), maka

$$1 + 4x_2 = 100.16(1)$$
$$4(1) + 100x_2 = 100.16x_2$$

Diperoleh $x_1 = 1$ dan $x_2 = 24.79$, sehingga $x = \begin{bmatrix} 1 \\ 24.79 \end{bmatrix}$.

$$\mathrm{e}_1 = \frac{\begin{bmatrix} 1 \\ 24.79 \end{bmatrix}}{\sqrt{[1, 24.79]\begin{bmatrix} 1 \\ 24.79 \end{bmatrix}}} = \frac{\begin{bmatrix} 1 \\ 24.79 \end{bmatrix}}{\sqrt{(1)(1)+(24.79)(24.79)}} = \frac{\begin{bmatrix} 1 \\ 24.79 \end{bmatrix}}{24.81016122} = \begin{bmatrix} 0.040 \\ 0.999 \end{bmatrix}$$

Dari persamaan 2, diperoleh :

$$x_1 + 4x_2 = 0.84x_1$$
$$4x_1 + 100x_2 = 0.84x_2$$

Ambil $x_1 = 1$ (sembarang), maka

$$1 + 4x_2 = 0.84(1)$$
$$4(1) + 100x_2 = 0.84x_2$$

Diperoleh $x_1 = 1$ dan $x_2 = -0.04$, sehingga $x = \begin{bmatrix} 1 \\ -0.04 \end{bmatrix}$.

$$\mathrm{e}_2 = \frac{\begin{bmatrix} 1 \\ -0.04 \end{bmatrix}}{\sqrt{[1, -0.04]\begin{bmatrix} 1 \\ -0.04 \end{bmatrix}}} = \frac{\begin{bmatrix} 1 \\ -0.04 \end{bmatrix}}{\sqrt{(1)(1)+(-0.04)(-0.04)}} = \frac{\begin{bmatrix} 1 \\ -0.04 \end{bmatrix}}{1.00079968} = \begin{bmatrix} 0.999 \\ -0.040 \end{bmatrix}$$

Pasangan nilai eigen-vektor eigen dari $\sum$ adalah

$$\lambda_1 = 100.16, \quad \mathrm{e}'_1 = [0.040, 0.999]$$
$$\lambda_2 = \quad 0.84, \quad \mathrm{e}'_2 = [0.999, -0.040]$$

Dengan cara yang sama, pasangan nilai eigen-vektor eigen dari ρ adalah

$$\lambda_1 = 1 + \rho = 1.4, \quad \mathrm{e}'_1 = [0.707, 0.707]$$
$$\lambda_2 = 1 - \rho = 0.6, \quad \mathrm{e}'_2 = [0.707, -0.707]$$

Masing-masing komponen utama menjadi

$$\sum: \begin{array}{l} Y_1 = 0.040X_1 + 0.999X_2 \\ Y_2 = 0.999X_1 - 0.040X_2 \end{array}$$

Dan

$$\rho: \quad \begin{aligned} Y_1 &= 0.707Z_1 + 0.707Z_2 = 0.707\left(\frac{X_1-\mu_1}{1}\right) + 0.707\left(\frac{X_2-\mu_2}{10}\right) \\ &= 0.707(X_1-\mu_1) + 0.0707(X_2-\mu_2) \\ Y_1 &= 0.707Z_1 - 0.707Z_2 = 0.707\left(\frac{X_1-\mu_1}{1}\right) - 0.707\left(\frac{X_2-\mu_2}{10}\right) \\ &= 0.707(X_1-\mu_1) - 0.0707(X_2-\mu_2) \end{aligned}$$

Oleh karena variansinya besar, X_2 dengan sepenuhnya mendominasi komponen utama pertama yang ditentukan dari $\sum$. Selain itu, komponen utama pertama menjelaskan proporsi

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{100.16}{101} = 0.992$$

dari total variansi populasi.

Ketika variabel X_1 dan X_2 dibakukan, bagaimanapun, menghasilkan variable yang berkontribusi sama untuk komponen utama yang ditentukan dari ρ. Gunakan hasil 8.4

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = .707\sqrt{1.4} = 0.837$$

Dan

$$\rho_{Y_1,Z_2} = e_{21}\sqrt{\lambda_1} = .707\sqrt{1.4} = 0.837$$

Dalam hal ini, komponen utama pertama menjelaskan proporsi

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$$

Dari total (baku) variansi populasi.

Variabel-variabel mungkin perlu dibakukan jika diukur dalam satuan pengukuran dengan jarak berbeda yang luas atau jika satuan pengukurannya tidak setara/sama. Contohnya, jika X_1 mewakili penjualan tahunan dalam jarak $10,000

sampai \$350,000 dan X_2 adalah rasio/perbandingan (pendapatan tahunan)/(total asset) dalam jarak 0.01 sampai 0.60, maka total variasi akan eksklusif mendekati penjualan dolar. Dalam ini, kita harapkan komponen utama tunggal (penting) dengan menimbang berat X_1. Sebagai kemungkinan lain, jika kedua variable dibakukan, kepentingan yang berikut akan menjadi order yang sama dan X_2 (atau Z_2) akan memainkan peran yang lebih besar dalam konstruksi komponen. Hal ini diperhatikan pada contoh 8.2.

Komponen Utama untuk Matriks Kovarians dengan Struktur Khusus

Ada matriks kovarians dan korelasi berpola tertentu yang komponen utamanya dapat dinyatakan dalam format sederhana. Andaikan $\sum$ adalah matriks diagonal

$$\sum = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \qquad (8\text{-}13)$$

Pilih $\mathbf{e}'_i = [0,\ldots,0,1,0,\ldots,0]$, dengan 1 pada posisi ke-i, kita perhatikan bahwa

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1\sigma_{ii} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{or} \quad \sum \mathbf{e}_i = \sigma_{ii}\mathbf{e}_i$$

Dan kita simpulkan bahwa $(\sigma_{ii}, \mathrm{e}_i)$ adalah pasangan nilai eigen-vektor eigen ke-*i*. Karena kombinasi linear $\mathrm{e}'_i \mathrm{X} = X_i$, kumpulan dari komponen utama hanya kumpulan asli dari variabel-variabel acak yang tidak dikorelasikan.

Untuk matriks kovarians dengan pola pada (8-13), tidak ada apapun yang diperoleh dari mencari komponen utama. Dari segi pandangan lain, jika X berdistribusi $N_p(\mu, \Sigma)$, bentuk dari kepadatan tetap adalah ellipsoid yang sumbu X nya berada pada arah variasi maksimum. Konsekwensinya, tidah usah berputar untuk mengkoordinasi system.

Standardisasi tidak pada hakekatnya mengubah keadaan untuk Σ pada (8-13). Dalam hal ini, $\rho = \mathrm{I}$, matriks identitas p x p. Lebih jelasnya, $\rho \mathrm{e}_i = 1\mathrm{e}_i$, maka nilai eigen 1 mempunyai keragaman p dan $\mathrm{e}'_i = [0,\ldots,0,1,0,\ldots,0]$, $i = 1, 2, \ldots, p$, adalah pilihan tepat untuk vektor eigen itu. Konsekwensinya, komponen utama yang ditentukan dari ρ adalah juga variabel-variabel asli $\mathrm{Z}_1, \ldots, \mathrm{Z}_p$. Selain itu, dalam hal ini nilai eigen sama, elipsoid normal multivariate dari kepadatan tetap adalah spheroids (bentuk bola).

Pola lain matriks kovarians, yang sering menggambarkan korespondensi diantara variabel-variabel yang berhubungan dengan ilmu biologi tertentu seperti ukuran makhluk hidup, mempunyai bentuk umum

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \qquad (8\text{-}14)$$

Menghasilkan matriks korelasi,

$$\rho = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \qquad (8\text{-}15)$$

Adalah juga matriks kovarians dari variabel yang dibakukan. Matriks pada (8-15) menyiratkan bahwa variable $X_1, X_2, \ldots, X_p$ dengan sama dihubungkan.

p nilai eigen dari matriks korelasi (8-15) dapat dibagi menjadi dua grup. Ketika ρ positif, yang paling besar adalah

$$\lambda_1 = 1 + (p-1)\rho \qquad (8\text{-}16)$$

Dengan vektor eigennya

$$\mathrm{e}'_1 = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}} \right] \qquad (8\text{-}17)$$

Sisanya $p-1$ nilai eigen adalah

$$\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$$

Dan satu pilihan untuk vektor eigennya adalah

$$\mathrm{e}'_2 = \left[\frac{1}{\sqrt{1 \text{ x } 2}}, \frac{-1}{\sqrt{1 \text{ x } 2}}, 0, \ldots, 0 \right]$$

$$\mathrm{e}'_3 = \left[\frac{1}{\sqrt{2 \text{ x } 3}}, \frac{1}{\sqrt{2 \text{ x } 3}}, \frac{-2}{\sqrt{2 \text{ x } 3}}, 0, \ldots, 0 \right]$$

$$\vdots$$

$$\mathrm{e}'_i = \left[\frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, \ldots, 0 \right]$$

$$\vdots$$

$$\mathrm{e}'_p = \left[\frac{1}{\sqrt{(p-1)p}}, \ldots, \frac{1}{\sqrt{(p-1)p}}, \frac{-(p-1)}{\sqrt{(p-1)p}} \right]$$

Komponen utama pertama

$$Y_1 = \mathrm{e}'_1 \mathrm{X} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} X_i$$

Sebanding dengan jumlah dari p variable asli. Itu bisa dianggap sebagai "indeks" dengan bobot yang sama. Komponen utama ini menjelaskan proporsi

$$\frac{\lambda_1}{p} = \frac{1+(p+1)\rho}{p} = \rho + \frac{1-\rho}{p} \qquad (8\text{-}18)$$

Dari total variasi populasi. Kita lihat bahwa $\lambda_1 / p = \rho$ untuk ρ dekat dengan 1 atau p besar. Contohnya, jika $\rho = 0.80$ dan $p = 5$, komponen pertama menjelaskan 84% dari total variansi. Ketika ρ dekat 1, $p-1$ komponen terakhir, secara bersama, menyumbang sangat kecil pada total variansi dan sering diabaikan.

Jika variable baku $Z_1, Z_2, \ldots, Z_p$ berdistribusi normal multivariate dengan matriks kovarians yang diberikan oleh (8-15), maka ellipsoid dari kepadatan tetap adalah "cigar-shaped" dengan sumbu utama sebanding dengan komponen utama pertama $Y_1 = \left(1/\sqrt{p}\right)[1,1,\ldots,1]\mathrm{X}$. Komponen utama ini menjadi proyeksi X pada garis equiangular $1' = [1,1,\ldots,1]$. Sumbu tambahan (dan sisa komponen utama) berbentuk bola arah simetris yang tegak lurus dengan sumbu utama (dan komponen utama pertama).

Nurul Kurniawati

041248

Interpretasi dari sampel komponen utama

Sampel komponen utama mempunyai beberapa interpretasi. Pertama kita anggap yang mendasari dari x adalah mendekati $N_p(0,\wedge)$ Maka sampel komponen utama $\widehat{y}_i = \widehat{e}_i(x-\bar{x})$ adalah realisasi dari populasi komponen utama $\widehat{Y}_i = \widehat{e}_i(X-\mu)$ yang berdistribusi $N_p(0,\wedge)$. Matrik diagonal $\wedge$ mempunyai entri-entri $\lambda_1, \lambda_2, ..., \lambda_p$ dan (λ_i, e_i) adalah sepasang nilai eigen-vektor eigen dari Σ juga, dari nilai sampel x_j, kita dapat memperkirakan μ dengan $\bar{x}$ dan $_{\Sigma}$ Σ dengan S. Jika S adalah terdefinisi dan positif. Bentuk garis (contour) terdiri dari semua px1 vektor yang memenuhi $(x-\bar{x})'S^{-1}(x-\bar{x}) = c^2$ (8.24)

Yang memperkirakan kepadatan konstan garis bentuk (contour) $(x-\mu)'\sum^{-1}(x-\mu)$ dengan kepadatan normal garis bentuk kira-kira dapat dilukiskan pada scaterplot dengan mengindikasikan distribusi normal. Scaterplot mungkin aagak menyimpang dari bentuk ellipsoid tapi kita tetap dapat menggali nilai eigen dari S dan memperoleh sampel komponen utama. Secara geometri data meungkin diplot sebagai n titik pada ruang p. Data dapat diekspresikan dalam koordinat baru, yang serupa dengan sumbu garis bentuk dari (8.24) Sekarang (8.24) mendefinisikan sentral hyperlipsoid yang terpusat pada $\bar{x}$ dan sumbu diberikan oleh vektor eigen dari S^{-1} atau sama dengan S . panjang dari

sumbu hyperlipsoid ini adalah sebanding dengan $\sqrt{\lambda_i}$, i= 1,2….,p dimana $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p \geq 0$ adalah nilai eigen dari S. Karena $\widehat{e}_i$ mempunyai panjang 1, nilai mutlak dari komponen utama ke I $\left| \widehat{y}_i = \widehat{e}_i(x - \bar{x}) \right|$ memberikan panjang proyeksi (x-$\bar{x}$) pada arah dari sumbu $\widehat{e}_i$ Konsekuensinya sampel komponen utama dapat dipandang sebagai hasil dari translasi dari system koordinat asli $\bar{x}$ dan koordinat sumbu $\bar{x}$ melewati penyebaran arah dari variansi maksimum. Interpretasi geometri dari sampel komponen utama yang diilustrasikan pada gambar 8.2 untuk p=2. Gambar 8.2(a) menunjukkan sebuah elip dengan jarak konstan, dengan pusat $\bar{x}$ dengan $\hat{\lambda}_1 \geq \hat{\lambda}_2$. Sampel komponen utama ditentukan dengan baik. Mereka terletak sepanjang sumbu x dari ellipsoid pada arah perpotongan dari sampel varians maksimum. Gambar 8.2(b) menunjukkan sebuah jarak ellip dengan pusat $\bar{x}$ dengan $\hat{\lambda}_1 = \hat{\lambda}_2$. Pada kasus ini sumbu dari ellips(lingkaran) jarak konstan ellips(lingkarang adalah tidak unik, dan terletak pada dua arah perpotongan, termasuk perpotongan dari sumbu asli. Ketika garis bentuk dari jarak konstan hampir bundar atau sama dengan ketika nilai eigen dari S hampir sama . Variansi sampel adalah homogen dalam semua arah , maka itu tidak mungkin mewakili data yang baik yang lebik sedikit dari p dimensi.

Jika akhirnya nilai eigen $\hat{\lambda}_i$ cukup kecil sedemikian sehingga varians dalam korespondensi $\hat{e}_i$ dapat diabikan , akhirnya sampel komponen utama dapat

diabaikan dan data menjadi cukup dengan perwakilan dalam ruang dari komponen yang menguasai.

Dena Rahayu

055521

2.2 Variasi Sampel dengan Menggunakan Komponen Utama

Menstandardisasi (membakukan) Sampel Komponen Utama

Sampel komponen utama secara umum, tidak berbeda berkenaan dengan perubahan dalam skala (lihat lat 8.2). Ketika kita menyebutkan perlakuan dalam komponen populasi, satuan pengukuran dari variabel-variabel x_1, x_2, x_3, ..., x_n berbeda, maka satuan varians baku pengukuran itu perlu dibakukan dengan jalan melakukan transformasi variabel x ke dalam variabel baku z. Untuk contoh, standardisasi terpenuhi dengan mengkonstruksi :

$$z_j = D^{-1/2}\left(x_j - \bar{x}\right) = \begin{bmatrix} \frac{x_i - \bar{x}}{\sqrt{s_{11}}} \\ \frac{x_{2j} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{pj} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, ..., n$$

(8-25)

p × n matriks data dari pengamatan yang distandardisasi

$$Z = \left[z_1, z_{2,...,} z_n\right] = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pn} \end{bmatrix} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12} - \bar{x}_1}{\sqrt{s_{11}}} & \cdots & \frac{x_{1n} - \bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{21} - \bar{x}_2}{\sqrt{s_{22}}} & \frac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2n} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{p1} - \bar{x}_p}{\sqrt{s_{pp}}} & \frac{x_{p2} - \bar{x}_p}{\sqrt{s_{pp}}} & \cdots & \frac{x_{pn} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad (8\text{-}26)$$

Akibatnya menghasilkan sampel vektor rata-rata [lihat (3-24)]

$$\bar{z} = \frac{1}{n} Z_1 = \frac{1}{n} \begin{bmatrix} \sum_{j=1}^{n} \frac{x_{1j} - \bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \frac{x_{2j} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{pj} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = 0 \qquad (8\text{-}27)$$

dan matriks sampel kovarians [lihat (3-27)]

$$S_z = \frac{1}{n-1} \left(Z - \frac{1}{n} Z_{11'} \right) \left(Z - \frac{1}{n} Z_{11'} \right) = \frac{1}{n-1} (Z - \bar{z}1')(Z - \bar{z}1')'$$

$$= \frac{1}{n-1} ZZ'$$

$$= \frac{1}{n-1} \begin{bmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{12}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix} = R \qquad (8\text{-}28)$$

Sampel komponen utama dalam pengamatan yang distandardisasi diberikan oleh persamaan (8-20), dengan matriks R sebagai pengganti S. Karena pengamatan telah "dipusatkan" dengan mengkonstruksi, maka tidak usah menulis komponen itu dalam bentuk persamaan (8-21).

Jika $z_1, z_2, \ldots, z_n$ adalah pengamatan yang distandardisasi dengan matriks kovarians R, sampel komponen utama ke-i adalah

$$\hat{y}_i = \widehat{e'}_i z = \hat{e}_{1i} z_1 + \hat{e}_{2i} z_2 + \ldots + \hat{e}_{pi} z_{p,} \qquad i = 1, 2, .., p$$

di mana $(\hat{\lambda}_i, \hat{e}_i)$ adalah pasangan nilai eigen – vektor eigen ke-i dari R dengan $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Juga,

$$\text{varians sampel } (\hat{y}_i) = \hat{\lambda}_i, \qquad i = 1, 2, \ldots, p$$

$$\text{kovarians sampel } (\hat{y}_i, \hat{y}_k) = 0 \qquad i \neq k \qquad (8\text{-}29)$$

Sebagai tambahan, total (yang distandardisasi) varians sampel = tr(R) = p = $\hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p$ dan $r_{\hat{y}_i, z_k} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i}$, i, k = 1, 2, ..., p

Gunakan (8-29), proporsi total varians sampel yang diterangkan oleh sampel komponen utama ke-i adalah

$$\begin{pmatrix} \text{proporsi yang distandardisasi} \\ \text{sampel varians dalam kaitan ke} - \text{i} \\ \text{sampel komponen utama} \end{pmatrix} = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, ..., p \qquad (8\text{-}30)$$

Sebuah aturan menyarankan menahan komponen itu dengan varians, $\hat{\lambda}_i$, adalah lebih besar dari kesatuan atau setara dengan, hanya komponen itu yang secara individu, menjelaskan sedikitnya suatu proporsi 1/p dari total varians. Aturan ini tidak mempunyai banyak pendukung teoritis, bagaimanapun, dan itu harus tidak diterapkan dengan berlebihan.

Contoh 8.5

Tingkat pengembalian mingguan untuk lima bursa/stock (Allied Chemical, du Pont, Union Carbide, Exxon, dan Texaco) yang didaftarkan di pasar bursa New York telah ditentukan untuk periode Januari 1975 sampai Desember 1976. Tingkat pengembalian mingguan digambarkan sebagai (Jumat sekarang yang menutup harga - Jumat sebelumnya yang menutup harga) / (Jumat sebelumnya yang menutup harga) yang disesuaikan untuk saham yang dipecah dan dividen. Data tersebut didaftarkan pada tabel 8.1 dalam latihan. Pengamatan dalam 100 minggu berurutan nampak seperti dengan bebas dibagi-bagikan, tetapi hanyalah tingkat tarip kembalian ke seberang bursa/stock dihubungkan, karena, seperti seseorang harapkan, bursa/stock cenderung untuk pindah bersama-sama sebagai jawaban atas kondisi-kondisi ekonomi umum.

Misalkan $x_1, x_2, \ldots, x_5$ menandakan tingkat pengembalian mingguan yang diamati untuk Allied Chemical, du Pont, Union Carbide, Exxon, dan Texaco secara berurutan. Maka

$$\bar{x}' = [0.0054\ , 0.0048, 0.0057, 0.0063, 0.0037]$$

$$\text{Dan } R = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{bmatrix}$$

Catatan kita bahwa R adalah matriks kovarians dalam pengamatan yang distandardisasi.

$$z_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}}\ , \qquad z_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}}\ , \qquad \ldots, \quad z_5 = \frac{x_5 - \bar{x}_5}{\sqrt{s_{55}}}$$

Nilai eigen dan yang dinormalisir bersesuaian dengan vektors eigen R telah ditentukan oleh suatu komputer dan diberikan di bawah ini.

$\hat{\lambda}_1 = 2.857,\ \ \hat{e}'_1 = [\ 0.464, 0.457, 0.470, 0.421, 0.421]$

$\hat{\lambda}_2 = 0.809,\ \ \hat{e}'_2 = [\ 0.240, 0.509, 0.260, -0.526, -0.582]$

$\hat{\lambda}_3 = 0.540,\ \ \hat{e}'_3 = [\ -0.612, 0.178, 0.335, 0.541, -0.435]$

$\hat{\lambda}_4 = 0.452,\ \ \hat{e}'_4 = [\ 0.387, 0.206, -0.6620, 0.472, -0.382]$

$\hat{\lambda}_5 = 0.343,\ \ \hat{e}'_5 = [\ -0.451, 0.676, -0.400, -0.176, 0.385]$

Penggunaan variabel yang distandardisasi, kita memperoleh dua sampel komponen utama yang pertama.

$$\hat{y}_1 = \hat{e}'_1 z = 0.464z_1 + 0.457z_2 + 0.470z_3 + 0.421z_4 + 0.421z_5$$

$$\hat{y}_2 = \hat{e}'_2 z = 0.240z_1 + 0.509z_2 + 0.260z_3 - 0.526z_4 - 0.582z_5$$

Komponen ini meliputi untuk

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\mathrm{p}}\right) 100\% = \left(\frac{2.857 - 0.809}{5}\right) 100\% = 73\%$$

dari total (yang distandardisasi) sampel variansi, mempunyai penafsiran menarik. Komponen yang pertama adalah (dengan kasar) penjumlahan dengan sama dihargai, atau index, dari lima bursa/stock. Komponen ini boleh jadi disebut suatu *bursa/stock umum - komponen pasar*, atau secara sederhana suatu *komponen pasar*. (Sesungguhnya, lima bursa/stock ini adalah tercakup di Dow Jones Industri Average)

Komponen yang kedua menghadirkan suatu kontras antara bursa/stock kimia (Allied Chemical, du Pont, dan Union Carbide) dan bursa/stock minyak (Exxon dan Texaco). Itu mungkin disebut suatu *komponen industri*. Dengan begitu kita lihat bahwa kebanyakan dari variasi di dalam pengembalian bursa/stock ini adalah dalam kaitan dengan aktivitas pasar dan tidak dihubungkan dengan aktivitas industri. Penafsiran bursa/stock ini menghargai perilaku yang telah pula diusulkan oleh Raja. Komponen yang sisanya tidaklah mudah untuk menginterpretasikannya dan secara bersama, menghadirkan variasi yang mungkin dikhususkan untuk bursa/stock masing-masing. Bagaimanapun juga, mereka tidak menjelaskan sebagian besar total sampel variansi.

Contoh ini menyediakan suatu kasus di mana itu nampak masuk akal untuk mempertahankan suatu komponen ($\hat{y}_2$) berhubungan dengan suatu nilai eigen kurang dari 1.

Contoh 8.6

Ahli genetika sering terkait dengan warisan dalam karakteristik yang dapat diukur beberapa kali selama seumur hidup binatang. Berat badan (dalam gram)

untuk n = 150 tikus-tikus betina telah diperoleh dengan seketika setelah kelahiran mereka yang pertama. Berat lahir tikus betina ditampilkan dari matriks ini dengan sampel vektor rata-rata dan matriks sampel korelasinya adalah

$$\bar{x}' = [39.88\ , 45.08, 48.11, 49.95]$$

$$R = \begin{bmatrix} 1.000 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.000 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.000 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.000 \end{bmatrix}$$

Nilai eigen dari matriks ini adalah

$\hat{\lambda}_1$= 3.058, $\hat{\lambda}_2$ = 0.382, $\hat{\lambda}_3$0.342, dan $\hat{\lambda}_4$= 0.217

Catatan kita bahwa nilai eigen yang pertama mendekati sama dengan 1 + (p – 1) $\bar{r}$ = 1 + (4 – 1)(0.68540 = 3.056, dimana $\bar{r}$ adalah rata-rata aritmatik dari unsur-unsur diagonal-off dalam R. Sisa nilai eigen adalah kecil dan sekitar sama, walaupun $\hat{\lambda}_4$ sedikit banyaknya lebih kecil dibanding $\hat{\lambda}_2$ dan $\hat{\lambda}_3$. Maka ada beberapa bukti dimana bersesuaian dengan populasi matriks korelasi ρ mungkin dalam “korelasi sama” berbentuk seperti dalam (8-15). Dugaan ini diselidiki lebih lanjut dalam contoh 8.9.

Komponen utama yang pertama

$$\hat{y}_i = \widehat{e'}_i z = 0.49z_1 + 0.52z_2 + 0.49z_3 + 0.50z_4$$

meliputi $100\left(\frac{\hat{\lambda}_1}{p}\right)\ \% = 100\left(\frac{3.058}{4}\right)\ \% = 76\%$ dari total variansi. Walaupun berat rata-rata pos kelahiran meningkat dari waktu ke waktu, variasi dalam berat cukup baik diterangkan oleh komponen utama yang pertama dengan koefisien yang hampir sama.

2.3 Grafik komponen utama

Plot dalam komponen utama dapat mengungkapkan kecurigaan pengamatan, seperti halnya menyediakan pemeriksaan pengambil-alihan dalam kenormalan. Karena komponen utama adalah kombinasi linear dalam variabel yang asli, itu tidaklah tidak beralasan untuk mengharapkan plot dalam komponen utama menjadi mendekati normal. Itu sering diperlukan untuk memverifikasi bahwa komponen utama yang awal kira-kira berdistribusi normal ketika plot dalam komponen digunakan sebagai data masukan untuk analisa tambahan.

Komponen utama yang terakhir dapat membantu menunjukkan dengan tepat kecurigaan pengamatan. Masing-masing pengamatan x_j dapat dinyatakan sebagai sebuah kombinasi linear

$$x_j = (x'_j\hat{e}_1)\hat{e}_1 + (x'_j\hat{e}_2)\hat{e}_2 + \ldots + (x'_j\hat{e}_p)\hat{e}_p$$

$$= \hat{y}_{1j}\hat{e}_1 + \hat{y}_{2j}\hat{e}_2 + \ldots + \hat{y}_{pj}\hat{e}_p$$

dari himpunan lengkap vektor eigen $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p$ dalam S. Maka penting dalam menentukan komponen utama yang terakhir seberapa baik kecocokan awal pengamatan. Yaitu :

$\hat{y}_{1j}\hat{e}_1 + \hat{y}_{2j}\hat{e}_2 + \ldots + \hat{y}_{q-1}\hat{e}_{q-1}$ berbeda dengan x_j dari $\hat{y}_{qj}\hat{e}_q + \ldots + \hat{y}_{pj}\hat{e}_p$ yang panjang kuadratnya adalah $\hat{y}^2{}_{qj} + \ldots + \hat{y}^2{}_{pj}$. Mencurigai pengamatan akan sering sedemikian hingga sedikitnya satu dai koordinat $\hat{y}_{qj}, \ldots, \hat{y}_{pj}$ mendukung panjang kuadrat ini akan menjadi besar.

(lihat lampiran 8A untuk hasil perkiraan yang lebih umum).

Pernyataan yang berikut meringkas gagasan ini.

1. Untuk membantu memeriksa asumsi yang normal, konstruksi diagram yang menyebar untuk pasangan komponen utama yang awal. Juga membuat Q-Q plot dari nilai-nilai sampel yang dihasilkan oleh masing-masing komponen utama.

2. Konstruksi diagram yang menyebar dan Q-Q plot untuk awal komponen utama yang terakhir. Bantuan ini mengidentifikasi kecurigaan pengamatan.

Diagnostik menyertakan komponen utama dengan sama kepada pemeriksaan asumsi untuk suatu model regresi berganda multivariat. Sesungguhnya, kita mempunyai beberapa model yang cocok dari metoda penilaian manapun, hal itu bijaksana untuk mempertimbangkan bahwa

$$\text{vektor residual} = (\text{vektor pengamatan}) - \begin{pmatrix}\text{vektor yang diramalkan} \\ \text{nilai} - \text{nilai yang diperkirakan}\end{pmatrix}$$

$$\underset{(p \times 1)}{\hat{e}_j} = \underset{(p \times 1)}{y_j} - \underset{(p \times 1)}{z'_j\hat{\beta}}, \qquad j = 1, 2, ..., n$$

untuk model linier multivariat. Komponen utama, diperoleh dari matriks kovarians yang bersifat sisa, $\frac{\sum_{j=1}^{n}(\hat{e}_j - \bar{\hat{e}}_j)(\hat{e}_j - \bar{\hat{e}}_j)'}{n-p}$ dapat diteliti dengan cara yang sama sebagai yang ditentukan dari suatu sampel acak. Kita harus sadar bahwa ada ketergantungan linier di antara yang bersifat sisa dari suatu analisa regresi linier, sehingga nilai eigen yang terakhir akan menjadi nol di dalam membulatkan kesalahan.

Naomi Nessyana

055589

2.4 Analisis sampel Besar

Nilai eigen dan vektor eigen dari matriks kovarian (korelasi) adalah analisis komponen utama yang penting. Penentuan vektor eigen bertujuan untuk memaksimumkan peubah dan penentuan nilai eigen bertujuan untuk menentukan variansi.

Berkenaan dengan keputusan, kualitas penaksiran komponen utama haruslah berdasarkan pasangan nilai eigen-vektor eigen $(\hat{\lambda}_i, \hat{e}_i)$ yang diambil dari S atau R. Karena variasi penarikan sampel, nilai eigen dan vektor eigen ini akan berbeda dari populasinya.

Sifat-Sifat Sampel Besar

Perhatikan hasil sampel besar dengan interval kepercayaan untuk $\ddot{\lambda}_i$ dan $\hat{e}_i$ diasumsikan dengan mengamati $X_1, X_2, \ldots, X_n$ adalah sampel acak dari populasi normal. Ini juga diasumsikan nilai eigen yang tidak diketahui dari Σ ada dan bernilai positif, sehingga $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. Kecuali, ukuran dimana angka-angka dari nilai eigen diketahui. Biasanya konklusi untuk nilai eigen ada di gunakan kecuali kalau ada alasan yang kuat untuk mempercayai Σ mempunyai matriks yang khusus untuk menghasilkan persamaan nilai eigen. Terkadang asumsi normal dilanggar, interval kepercayaan pada cara ini tersedia untuk beberapa indikasi dari nilai $\ddot{\lambda}_i$ dan $\hat{e}_i$yang belum pasti.

Anderson dan Girshick menentukan teori distribusi sampel-besar dibawah ini untuk nilai eigen $\ddot{\lambda}' = [\ddot{\lambda}_1, \ldots, \ddot{\lambda}_p]$ dan vektor eigen $\hat{e}_1, \ldots, \hat{e}_p$ dari S, yaitu:

1. Misalkan A adalah matriks diagonal dari nilai eigen $\lambda_1, \lambda_2, \ldots, \lambda_p$ dari Σ, maka $\sqrt{n}\,(\tilde{\lambda} - \lambda)$ adalah penaksir $N_p(0, 2\Delta^2)$

2. Misalkan $E_i = \lambda_i \sum_{\substack{k=i \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k \hat{e}_k$ maka $\sqrt{n}(\hat{e}_i - e_i)$ adalah penaksir $N_p(0, E_i)$

3. Setiap $\tilde{\lambda}_i$ berdistribusi bebas dari anggota yang berasosiasi $\hat{e}_i$.

Hasil 1 implikasinya adalah untuk n besar, $\tilde{\lambda}_i$ berdistribusi bebas. Selanjutnya $\ddot{\lambda}_i$ berdistribusi dengan penaksirnya distribusi $N\left(\lambda_i, \frac{2\lambda_i^2}{n}\right)$. Dengan menggunakan distribusi normal $P[(\tilde{\lambda}_i - \lambda_i)] \leq z\left(\frac{\alpha}{2}\right)\lambda_i\sqrt{\frac{2}{n}} = 1 - \alpha$. Untuk sampel besar $100(1-\alpha)\%$ interval kepercayaannya untuk λ_i menjadi

$$\frac{\hat{\lambda}_i}{\left(1 + z\left(\frac{\alpha}{2}\right)\sqrt{2/n}\right)} < \lambda_i \leq \frac{\hat{\lambda}_i}{\left(1 - z\left(\frac{\alpha}{2}\right)\sqrt{2/n}\right)} \qquad (8\text{-}33)$$

dimana $z\left(\frac{\alpha}{2}\right)$ diatas persentil $100\left(\frac{\alpha}{2}\right)$ dari distribusi normal standar. Jenis persamaan simultan Bonterroni interval $100(1-\alpha)\%$ untuk m λ_i diganti $z(\alpha/2)$.

Hasil 2 implikasi bahwa $\hat{e}_i$ adalah distribusi normal yang berkorespondensi $\hat{e}_i$ untuk sampel besar. Elemen-elemen setiap $\hat{e}_i$ berkorelasi dan korelasinya bergantung untuk pemisahan nilai eigen $\lambda_1, \lambda_2, \ldots, \lambda_p$ yang tidak diketahui dan sampel berukuran n penaksiran standar eror untuk koefisien diberikan dengan akar kuadrat dari diagonal-diagonal elemen-elemen dari $(1/n)\ddot{E}_i$ dimana $\ddot{E}_i$ didapatkan dari E_i dengan mensubstitusi $\ddot{\lambda}_i$ untuk λ_i dan $\hat{e}_i$ untuk e_i

Contoh 8.8

Didapatkan interval kepercayaan untuk variansi populasi komponen utama menggunakan persediaan harga pada data tabel 8.1.

Asumsikan persediaan suku dari hasil yang mewakili gambar dari populasi $N_5(\mu,)$ dimana adalah definit positif dengan nilai eigen berbeda dengan $\lambda_1 > \lambda_2 > \cdots > \lambda_5 > 0$. Karena n=100 besar, kita menggunakan 8.33 dengan i=1 untuk mengkontruksi interval kepercayaan untuk λ_i sebesar 95%.

Dari 8.10, $\ddot{\lambda}_i = 0.0036$ dan $z(0.025) = 1.96$ maka dengan taraf nyata 95%

$$\frac{0.036}{\left(1+1.96\sqrt{2/100}\right)} < \lambda_i \leq \frac{0.036}{\left(1-1.96\sqrt{2/100}\right)}$$

$$\Leftrightarrow 0.0028 \leq \lambda_i \leq 0.050$$

Sewaktu-waktu nilai eigen besar, misalkan 100 atau bahkan 1000. Pada umumnya dapat menjadi besar, untuk level kepercayaan masuk akal. Pada umumnya interval kepercayaan memperoleh rata-rata yang sama lebih besar sehingga nilai λ_i membesar.

Pengujian Kesamaan Struktur Korelasi

Struktur korelasi yang khusus $\mathrm{cov}(x_1, x_k) = \sqrt{\tau_{1i}\tau_{kk}\rho}$ atau $\mathrm{corr}(x_1, x_k) = \rho, \forall\, i \neq k$ adalah struktur penting dimana nilai eigen dari Σ tidak berbeda dan hasil sebelumnya tidak digunakan.

Untuk pengujian struktur ini, misalkan

$$H_0 : \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

$$H_1: \rho \neq \rho_0$$

Pengujian H_0 melawan H_1 didasarkan dengan rasio statistik likelihood. Tetapi lawley menunjukkan hal itu ekuivalen dengan prosedur uji yang dapat dikonstruksi dari elemen diagonal dari R.

Prosedur Lawley memerlukan kuantitas

$$\bar{r}_k = \frac{1}{p-1}\sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik} \quad k = 1,2,\ldots,p$$

$$\bar{r} = \frac{2}{p(p-1)}\sum\sum_{i<k} r_{ik} \tag{8-34}$$

$$\hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}$$

Ini jelas bahwa $\bar{r}_k$ adalah rata-rata elemen diagonal di kolom (baris) ke-k dari R dan $\bar{r}$ adalah secara keseluruhan rata-rata dari elemen diagonal.

Penaksiran sampel besar, uji level-α memepunyai bentuk tolak H_0 dan terima H_1 jika

$$T = \frac{(n-1)}{(1-\bar{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(r_{ik}-\bar{r})^2\right] > \chi^2_{(p+1)(p-2)/2}(\alpha) \tag{8-35}$$

dimana $\chi^2_{(p+1)(p-2)/2}(\alpha)$ dibawah persentil ke (100α) dari distribusi chi-kuadrat dengan derajat kebebasannya $(p+1)(p-2)/2$.

Contoh 8-9:

Matriks sampel korelasi dikonstruksi dari berat lahir tikus betina yang dibahas pada contoh 8-6 dan disajikan di bawah ini

$$R = \begin{bmatrix} 1.0 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.0 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.0 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.0 \end{bmatrix}$$

Kita akan menggunakan matriks korelasi untuk menggambarkan pengujian sampel besar

$p = 4$ dan akan ditentukan

$$H_0: \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

$$H_1: \rho \neq \rho_0$$

Dengan menggunakan 8-34 dan 8-35

$$\bar{r}_1 = \frac{1}{3}(0.7501 + 0.6329 + 0.6363) = 0{,}6731$$

$$\bar{r}_2 = \frac{1}{3}(0.7501 + 0.6925 + 0.7386) = 0{,}7271$$

$$\bar{r}_3 = \frac{1}{3}(0.6329 + 0.6925 + 0.6625) = 0{,}6626$$

$$\bar{r}_4 = \frac{1}{3}(0.6363 + 0.7386 + 0.6625) = 0{,}6791$$

$$\bar{r} = \frac{2}{4(3)}(0.7501 + 0.6329 + 0.6363 + 0.6925 + 0.7386 + 0.6625) = 0{,}6855$$

$$\sum\sum_{i<k}(r_{ik} - \bar{r})^2 = (0.7501 - 0.6855)^2 + (0.6329 - 0.6855)^2 + \cdots + (0.6625 - 0.6855)^2 = 0{,}1277$$

$$\sum_{k=1}^{4}(r_k - \bar{r})^2 = (0.6731 - 0.6855)^2 + \cdots + (0.6791 - 0.6855)^2 = 0{,}0245$$

$$\hat{\gamma} = \frac{(4-1)^2[1-(1-0.6855)^2]}{4-(4-2)(1-0.6855)^2} = 2{,}1329$$

dan

$$T = \frac{(150-1)}{(1-0.6855)^2}[0.01277 - (2.1329)(0.00245)] = 11{,}4$$

Karena $(p+1)(p-2)2 = \frac{5(2)}{2} = 5$, dan nilai kritis 5% untuk pengujian pada (8-15)adalah $\chi_5^2(0.05) = 11.07$. nilai pengujian statistik yang ditaksir sama dengan titik kritis 5% sehingga Ho ditolak.

Perhatikan contoh 8-6, nilai eigen terkecil $\hat{\lambda}_2, \hat{\lambda}_3$ dan $\hat{\lambda}_4$ agak berbeda, dengan $\hat{\lambda}_4$ lebih kecil daripada $\hat{\lambda}_2$ dan $\hat{\lambda}_3$. Akibatnya, dengan ukuran sampel besar pada masalah ini, perbedaannya kecil dari struktur sehingga matriks kesamaan korelasinya menunjukkan ssecara statistik berarti.

Penaksir komponen utama sampel dalam bidang Geometri

Kita akan menunjukkan interpretasi untuk penaksiran data yang didasarkan pada r pertama komponen utama sampel. Interpretasi dari sebaran plot dan bidang dimensi-n mewakili kepercayaan hasil aljabar dibawah ini. Perhatikan penaksiran bentuk $\underset{p\times n}{A} = [a_1, a_2, \ldots, a_n]$ berarti pengertian rata-rata matriks data

$$[x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}]$$

Eror dari penaksiran diukur dari jumlah eror kuadrat np

$$\Sigma_{j=1}^{n}(x_j - \bar{x} - a_j)'(x_j - \bar{x} - a_j) = \Sigma_{i=1}^{p}\Sigma_{j=1}^{n}(x_j - \bar{x} - a_j)^2 \qquad \text{(8A-1)}$$

Hasil 8A-1. MIsalkan $\underset{p\times n}{A}$ sembarang matrik dengan rank (A)$\leq$r<min(p,n). eror dari penaksiran jumlah kuadrat (8A-1) diminimumkan oleh

$$\hat{A} = \hat{E}\hat{E}'[x_j - \bar{x}, \ldots, x_j - \bar{x}] = \hat{E}\begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_r' \end{bmatrix}$$

Sehingga kolom ke-j dari $\ddot{A}$ adalah

$$\hat{a}_j = \hat{y}_{1j}\hat{e}_1 + \hat{y}_{2j}\hat{e}_2 + \cdots + \hat{y}_{rj}\hat{e}_r$$

dimana $[\hat{y}_{1j}, \hat{y}_{2j}, \ldots, \hat{y}_{rj}] = [\hat{e}_1'(x_j - \bar{x}), \hat{e}_2'(x_j - \bar{x}), \ldots, \hat{e}_r'(x_j - \bar{x})]$

adalah nilai r pertama komponen utama sampel untuk unit ke-j. Selanjutnya,

$$\sum_{j=1}^{n}(x_j - \bar{x} - a_j)'(x_j - \bar{x} - a_j) = (n-1)(\hat{\lambda}_{r+1} + \cdots + \hat{\lambda}_p)$$

dimana $\ddot{\lambda}_{r+1} \geq \cdots \geq \ddot{\lambda}_p$ adalah nilai eigen terkecil dari S.

Bukti:

Perhatikan sembarang kolom A adalah kombinasi linear dari himpunan dari r vektor yang tegak lurus $l_1, l_2, \ldots l_r$ sehingga $L = [l_1, l_2, \ldots l_r]$ memenuhi $L'L = I$. untuk L tertentu, $x_j - \bar{x}$ merupakan penaksir terbaik dengan proyeksinya terentang oleh $l_1, l_2, \ldots l_r$ atau

$$(x_j - \bar{x})'l_1l_1 + (x_j - \bar{x})'l_2l_2 + \cdots + (x_j - \bar{x})'l_rl_r$$

$$= [l_1, l_2, \ldots, l_r]\begin{bmatrix} l_1'(x_j - \bar{x}) \\ l_2'(x_j - \bar{x}) \\ \vdots \\ l_r'(x_j - \bar{x}) \end{bmatrix} = LL'(x_j - \bar{x}) \qquad (8A\text{-}2)$$

Karenanya, untuk vektor yang berubah-ubah

$$x_j - \bar{x} - Lb_j = x_j - \bar{x} - LL'(x_j - \bar{x}) + LL'(x_j - \bar{x}) - Lb_j$$

$$= (I - LL')(x_j - \bar{x}) + L(L'(x_j - \bar{x}) - Lb_j)$$

Sehingga jumlah kuadrat eror adalah

$$(x_j - \bar{x} - Lb_j)'(x_j - \bar{x} - Lb_j) = (x_j - \bar{x})'(I - LL')(x_j - \bar{x}) + 0$$

$$+LL'\left((x_j - \bar{x}) - Lb_j\right)'\left((x_j - \bar{x}) - Lb_j\right)$$

Dimana hasil kali menghilang karena $(I - LL')L = L - LL'L = L - L = 0$. Hubungan terakhir bernilai positif kecuali jika b_j dipilih sehingga $Lb_j = LL'(x_j - \bar{x})$ proyeksi

Lebih jauh, dengan memilih $a_j = Lb_j = LL'(x_j - \bar{x})$, (8A-1)menjadi

$$\sum_{j=1}^{n}\left(x_j - \bar{x} - LL'(x_j - \bar{x})\right)'\left(x_j - \bar{x} - LL'(x_j - \bar{x})\right) = \sum_{j=1}^{n}(x_j - \bar{x})'(I - LL')(x_j - \bar{x})$$

$$= \sum_{j=i}^{n}(x_j - \bar{x})'(x_j - \bar{x}) - \sum_{j=1}^{n}(x_j - \bar{x})'LL'(x_j - \bar{x}) \quad \text{(8A-3)}$$

Kita memposisikan untuk meminimumkan eror sehingga memilih L dengan memaksimumkan hubungan terakhir 8A-3. Dengan sifat-sifat dari trace

$$\sum_{j=1}^{n}(x_j - \bar{x})'LL'(x_j - \bar{x}) = \sum_{j=1}^{n}\mathrm{tr}\left[(x_j - \bar{x})'LL'(x_j - \bar{x})\right]$$

$$= \sum_{j=1}^{n}\mathrm{tr}\left[LL'(x_j - \bar{x})(x_j - \bar{x})'\right]$$

$$= (n-1)\mathrm{tr}[LL'S] = (n-1)\mathrm{tr}[L'SL] \quad \text{(8A-}$$

Sehingga pilihan terbaik untuk L dengan memaksimumkan jumlah elemen diagonal dari $L'SL$. Dari 8-19 pemilihan $l_1$untuk memaksimumkan $l_1'Sl_1$, elemen diagonal pertama dari $L'SL$ memberikan $l_1 = \hat{e}_1$. Untuk l_2 yang tegak lurus ke $\hat{e}_1$, $l'Sl_2$ dimaksimumkan oleh $\hat{e}_2$. Selanjutnya, kita menentukan $\hat{L} = [\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_r] = \hat{E}$ dan $A = EE'[x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}]$.'

Dengan memilih ini, elemen diagonal ke-I dari $L'SL$ adalah $\hat{e}_i'S\hat{e}_i = \hat{e}_i'(\hat{\lambda}_i\hat{e}_i) = \hat{\lambda}_i$ sehingga $\mathrm{tr}[\hat{L}'S\hat{L}] = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_r$. Juga

$$\sum_{j=1}^{n}(x_j-\bar{x})'(x_j-\bar{x}) = tr\left[\sum_{j=1}^{n}(x_j-\bar{x})'(x_j-\bar{x})\right] = (n-1)tr(S) = (n-1)(\hat{\lambda}_1+\hat{\lambda}_2+\ldots+\hat{\lambda}_r)$$

Interpretasi Bidang Geometri Dimensi p

Interpretasi geometri meliputi penentuan bidang penaksir terbaik ke plot menyebar dimensi p. bidang asal ditentukan oleh $l_1, l_2, \ldots l_p$ yang terdiri dari semua titik x dengan

$$x = b_1 l_1 + b_2 l_2 + \cdots + b_r l_r = Lb_1$$

Bidang ini diartikan melewati a menjadi a+Lb untuk beberapa b

Kita ingin memilih bidang $a + Lb_j$ dimensi r sehingga meminimumkan jumlah kuadrat jarak antara pengamatan x_j dan bidang. Jika x_j ditaksir oleh $a + Lb_j$ dengan $\sum_{j=1}^{n} b_j = 0$.[4]

$$\sum_{j=1}^{n}(x_j-a-Lb_j)'(x_j-a-Lb_j) = \sum_{j=1}^{n}(x_j-\bar{x}-Lb_j+\bar{x}-a)'(x_j-\bar{x}-Lb_j+\bar{x}-a)$$

$$= \sum_{j=1}^{n}(x_j-\bar{x}-Lb_j)'(x_j-\bar{x}-Lb_j) + n(\bar{x}-a)'(\bar{x}-a)$$

$$\geq \sum_{j=1}^{n}\left(x_j-\bar{x}-\hat{E}\hat{E}(x_j-\bar{x})\right)'\left(x_j-\bar{x}-\hat{E}\hat{E}(x_j-\bar{x})\right)$$

oleh hasil 8A-1 $[Lb_1,\ldots,Lb_r] = A$ mempunyai rank(A) $\leq$ r. Batas bawah dijangkau dengan mengambil $a = \bar{x}$ sehingga bidang melewati rata-rata sampel. Bidang ini ditentukan oleh $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_r$. Koefisien dari $\hat{e}_k$ adalah $\hat{e}_k'(x_j-\bar{x}) = \hat{y}_{kj}$, komponen utama sampel ke-k di evaluasi pada pengamatan ke-j.

Sebuah interpretasi alternative diberikan. Peneliti menempatkan bidang sepanjang $\bar{x}$, dan langkah selanjutnya mendapatkan penyebaran terbaik diantara

bayangan dari pengamatan. Dari 8A-2, proyeksi deviasi $x_j - \bar{x}$ dalam bidang Lb adalah $v_j = LL'(x_j - \bar{x})$. $\bar{v} = 0$ dan jumlah kuadrat panjang proyeksi deviasi adalah

$$\sum_{j=1}^{n} v_j' v_j = \sum_{j=1}^{n} (x_j - \bar{x})' LL'(x_j - \bar{x}) = (n-1)\,tr[L'SL]$$

dimaksimumkan oleh $L = \hat{E}$. Karena $\bar{v} = 0$

$$(n-1)S_v = \sum_{j=1}^{n} (v_j - \bar{v})(v_j - \bar{v})' = \sum_{j=1}^{n} v_j v_j'$$

Dan bidang ini juga memaksimumkan variansi total.

$$tr(S_v) = \frac{1}{(n-1)} tr\left[\sum_{j=1}^{n} v_j' v_j\right]$$

Interpretasi Bidang Geometri Dimensi n

Perhatikan penaksiran di 8A.1 baris demi baris. Untuk $r = 1$, baris ke-i $[x_{i1} - \bar{x}_i, x_{i2} - \bar{x}_i, \ldots, x_{1n} - \bar{x}_i]$ ditaksir oleh kelipatan $c_i b'$ ditentukan dari vektor $b = [b_1, b_2, \ldots, b_n]$. Panjang vektor $b = [b_1, b_2, \ldots, b_n]$. Panjang kuadrat eror dari penaksiran panjang kuadrat

$$L_i^2 = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i - c_i b_j)^2$$

Perhatikan $\underset{p \times n}{A} = \{a_{ij}\}$ dengan $a_{ij} = c_i b_j$ sehingga

$$\tilde{A} = [\hat{e}_1 \hat{e}_1 (x_1 - \bar{x}), \hat{e}_1 \hat{e}_1 (x_2 - \bar{x}), \ldots, \hat{e}_1 \hat{e}_1 (x_n - \bar{x})]$$

$$= \hat{e}_1 [\hat{y}_{11}, \hat{y}_{12}, \ldots, \hat{y}_{1n}]$$

meminimumkan jumlah panjang kuadrat $\sum_{i=1}^{p} L_i^2$ sehingga tujuan terbaiknya ditentukan oleh nilai vektor dari komponen utama pertama. Ilustrasi ini pada gambar 8.6 di halaman 388. Vektor deviasi lebih panjang mempunyai pengaruh paling besar untuk meminimumkan $\sum_{i=1}^{p} L_i^2$.

Jika variabel-variabel adalah standardisasi pertama, vektor hasilnya $\left[\frac{(x_{i1}-\bar{x}_i)}{\sqrt{s_{ii}}}, \frac{(x_{i2}-\bar{x}_i)}{\sqrt{s_{ii}}}, \ldots (x_{ij}-\bar{x}_i)/\sqrt{s_{ii}}\right]$ mempunyai panjang 1 untuk setiap variabel dan setiap pengaruh yang sama menggunakan tujuan pilihan.

Pada ukuran lain, vektor b berpindah mengelilingi tempat-n untuk meminimumkan jumlah dari jarak kuadrat antara $[x_{i1}-\bar{x}, x_{i2}-\bar{x}, x_{i3}-\bar{x}]'$ dan proyeksinya pada garis ditentukan oleh b. Komponen utama kedua meminimumkan kuantitas yang sama selama semua vektor tegak lurus pada pilihan pertama.

BAB III

KESIMPULAN

Pada dasarnya analisis komponen utama bertujuan untuk menerangkan struktur varians-kovarians melalui kombinasi linier dari variabel-variabel. Secara umum analisis komponen utama bertujuan untuk mereduksi data dan menginterpretasikannya. k buah komponen utama dapat mengganti p buah variabel asal dalam bentuk matriks berukuran n x p yang direduksi menjadi matriks berukuran lebih kecil yang mengandung n pengukuran pada k buah komponen utama (matriks berukuran n x k, dimana k < p).

Secara aljabar, komponen utama adalah kombinasi linier khusus dari p variabel acak $X_1, X_2, ..., X_p$. Secara geometris, kombinasi linier ini menggambarkan pemilihan dari sistem koordinat yang diperoleh dengan merotasikan sistem awal dengan $X_1, X_2, ..., X_p$ sebagai sumbu koordinat.

Komponen utama populasi bergantung pada matriks kovarians $\sum$ yang memiliki pasangan nilai eigen-vektor eigen $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ dimana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, maka komponen uama ke-*i* diberikan oleh

$$Y_i = e'_i X = e_{1i} X_1 + e_{2i} X_2 + ... + e_{pi} X_p, \qquad i = 1,2,\ldots,p$$

Dengan,

$$Var(Y_i) = e'_i \Sigma e_i = \lambda_i \qquad i = 1, 2, ..., p$$

$$Cov(Y_i, Y_k) = e'_i \Sigma e_k = 0 \qquad i \neq k$$

Dan proporsi total varians dari komponen utama ke-*k* dari X adalah

$$\begin{pmatrix} proporsi\ variansi \\ populasi\ total\ dari \\ komponen\ utama \\ ke-k \end{pmatrix} = \frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \qquad k = 1,2,\ldots,p$$

Komponen utama populasi yang diperoleh dari variabel yang dibakukan $\left(Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}} \right)$ bergantung pada matriks korelasi ρ yang memiliki pasangan nilai eigen-vektor eigen $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ dimana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, maka komponen utama ke-*i* diberikan oleh

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

Dengan,

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

$$\rho_{Y_i, Z_k} = e_{ki}\sqrt{\lambda_i}, \quad i, k = 1, 2, ..., p$$

Dan proporsi total varians dari komponen utama ke-*k* dari Z adalah

$$\begin{pmatrix} \text{Proporsi dari (baku)} \\ \text{variansi populasi seharusnya} \\ \text{untuk komponen utama ke-}k \end{pmatrix} = \frac{\lambda_k}{p}, \quad k = 1, 2, ..., p$$

Komponen utama sampel bergantung pada matriks kovarians sampel S berukuran p x p yang memiliki pasangan nilai eigen-vektor eigen $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), ..., (\hat{\lambda}_p, \hat{e}_p)$ dimana $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$, maka komponen utama sampel ke-*i* diberikan oleh

$$\hat{y}_i = \hat{e}'_i \mathbf{x} = \hat{e}_{1i}x_1 + \hat{e}_{2i}x_2 + ... + \hat{e}_{pi}x_p, \qquad i = 1,2,\ldots,p$$

Dengan,

Varians sampel $(\hat{y}_k) = \hat{\lambda}_k$, $k = 1, 2, \ldots, p$

Kovarians sampel $(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$

Dan total varians sampel $= \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p$

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \qquad i, k = 1, 2, \ldots, p$$

Komponen utama sampel yang diperoleh dari variabel yang dibakukan $\left(z_j = \frac{(x_{pj} - \bar{x}_p}{\sqrt{S_{pp}}}\right)$ bergantung pada matriks kovarians R (jika $z_1, z_2, \ldots, z_n$ adalah pengamatan yang distandardisasi) di mana ($\hat{\lambda}_i, \hat{e}_i$) adalah pasangan nilai eigen – vektor eigen ke-i dari R dengan $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, maka komponen utama sampel ke-i adalah

$$\hat{y}_i = \widehat{e'}_i z = \hat{e}_{1i} z_1 + \hat{e}_{2i} z_2 + \ldots + \hat{e}_{pi} z_p, \quad i = 1, 2, .., p$$

Dengan,

varians sampel $(\hat{y}_i) = \hat{\lambda}_i, \quad i = 1, 2, \ldots, p$

kovarians sampel $(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$

Dan total (yang distandardisasi) varians sampel = tr(R) = p = $\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p$ dan $r_{\hat{y}_i, z_k} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i}$, i, k = 1, 2, ..., p

Proporsi total varians sampel yang diterangkan oleh komponen utama sampel ke-i adalah

$$\begin{pmatrix} \text{proporsi yang distandardisasi} \\ \text{sampel varians dalam kaitan ke} - i \\ \text{sampel komponen utama} \end{pmatrix} = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, \ldots, p$$