Exploratory Data Anaysis
Session 4

SSD23407

Egi Safitri, S.Mat., M.Si

Gambaran Umum

Pelajaran ini berfokus pada distribusi normal multivariat. Sama seperti distribusi normal univariat yang menjadi distribusi statistik yang paling penting dalam statistika univariat, distribusi normal multivariat adalah distribusi yang paling penting dalam statistika multivariat.
Pertanyaan yang mungkin diajukan adalah, "Mengapa distribusi normal multivariat begitu penting?"Ada tiga alasan mengapa hal ini bisa terjadi:

1. **Kesederhanaan Matematis**: Distribusi ini relatif mudah digunakan, sehingga memudahkan kita dalam mendapatkan metode multivariat berbasis distribusi ini.
2. **Versi Multivariat dari Teorema Limit Sentral**: Dalam statistika multivariat, untuk sampel besar variabel acak, rata-rata sampel vektor akan mendekati distribusi normal multivariat.
3. Banyak fenomena alami yang dapat dimodelkan menggunakan distribusi ini, sama seperti dalam kasus univariat.

Tujuan Pembelajaran

Setelah menyelesaikan pelajaran ini, Anda seharusnya dapat:

- Memahami definisi distribusi normal multivariat.
- Menghitung nilai eigen dan vektor eigen untuk matriks 2x2.
- Menentukan bentuk distribusi normal multivariat dari nilai eigen dan vektor eigen dari matriks kovarians.

Distribusi Normal Univariat

Sebuah variabel acak X didistribusikan secara normal dengan mean μ dan varians σ^2 jika memiliki fungsi kepadatan probabilitas sebagai berikut:

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Varians σ^2 menentukan seberapa besar sebaran distribusi. Notasi singkat yang digunakan adalah:

$$X \sim N(\mu, \sigma^2)$$

Distribusi Normal Multivariat

Jika kita memiliki vektor acak $p \times 1$ $\mathbf{X}$ yang didistribusikan menurut distribusi normal multivariat dengan vektor rata-rata μ dan matriks kovarians Σ, maka fungsi kepadatan gabungannya adalah:

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Notasi singkatnya adalah:

$$\mathbf{X} \sim N(\mu, \Sigma)$$

Jarak Mahalanobis

Istilah di dalam eksponen disebut bentuk kuadratik:

$$(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)$$

Ini dikenal sebagai jarak Mahalanobis kuadrat. Jika variabel-variabel tersebut tidak berkorelasi, maka matriks kovarians akan berupa matriks diagonal:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

Catatan: Simbol Pi dan Kombinasi Linear

Catatan! Simbol Pi (Π) digunakan seperti tanda penjumlahan, namun bedanya kita mengalikan elemen-elemen mulai dari $j = 1$ hingga $j = p$. Di dalam produk ini terdapat distribusi normal univariat di mana variabel-variabel acak disubscript dengan j. Dalam hal ini, elemen-elemen dari vektor acak $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ adalah variabel acak yang saling bebas.

Kombinasi Linear dari Variabel Acak Multivariat

Kita juga dapat mempertimbangkan kombinasi linear dari elemen-elemen variabel acak normal multivariat seperti pada persamaan di bawah ini:

$$Y = \sum_{j=1}^{p} c_j X_j = \mathbf{c}^\top \mathbf{X}$$

Catatan! Untuk mendefinisikan kombinasi linear, variabel acak X_j tidak harus tidak berkorelasi. Koefisien c_j dipilih secara arbitrer atau dipilih berdasarkan masalah tertentu. Misalnya, kita dapat memilih koefisien untuk mendapatkan total asupan vitamin A dan C dari data survei gizi.

Distribusi Normal Multivariat

Misalkan vektor acak $\mathbf{X}$ berdistribusi normal multivariat dengan vektor rata-rata μ dan matriks kovarians Σ:

$$\mathbf{X} \sim N(\mu, \Sigma)$$

Maka Y akan berdistribusi normal dengan mean dan varians sebagai berikut:

$$\text{Mean: } \mathbf{c}^\top \mu = \sum_{j=1}^{p} c_j \mu_j$$

$$\text{Varians: } \mathbf{c}^\top \Sigma \mathbf{c} = \sum_{j=1}^{p} \sum_{k=1}^{p} c_j c_k \sigma_{jk}$$

Jadi, Y mengikuti distribusi normal dengan mean $\mathbf{c}^\top \mu$ dan varians $\mathbf{c}^\top \Sigma \mathbf{c}$:

$$Y \sim N(\mathbf{c}^\top \mu, \mathbf{c}^\top \Sigma \mathbf{c})$$

Fakta-fakta Penting tentang Distribusi Normal Multivariat

Beberapa fakta berguna untuk variabel dengan distribusi normal multivariat:

- Setiap variabel tunggal memiliki distribusi normal univariat.
- Setiap subset variabel juga memiliki distribusi normal multivariat.
- Setiap kombinasi linear dari variabel memiliki distribusi normal univariat.
- Setiap distribusi kondisional dari subset variabel dengan nilai yang diketahui untuk subset lain adalah distribusi multivariat.

Contoh 4-1: Kombinasi Linear Pengukuran Kolesterol

Pengukuran diambil dari pasien serangan jantung pada level kolesterol mereka. Pengukuran diambil pada hari ke-0, hari ke-2, dan hari ke-4 setelah serangan. Vektor mean sampel adalah sebagai berikut:

Variabel	Mean
X_1 = Hari ke-0	259.5
X_2 = Hari ke-2	230.8
X_3 = Hari ke-4	221.5

Matriks kovariansnya adalah:

$$\Sigma = \begin{pmatrix} 2276 & 1508 & 813 \\ 1508 & 2206 & 1349 \\ 813 & 1349 & 1865 \end{pmatrix}$$

Perbedaan antara Hari ke-0 dan Hari ke-2

Misalkan kita tertarik pada perbedaan $X_1 - X_2$, yaitu perbedaan antara pengukuran pada hari ke-0 dan hari ke-2. Kombinasi linear ini dapat ditulis sebagai:

$$\mathbf{a}^\top \mathbf{x} = \begin{pmatrix} 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Nilai rata-rata untuk perbedaan ini adalah:

$$\begin{pmatrix} 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 259.5 \\ 230.8 \\ 221.5 \end{pmatrix} = 28.7$$

Variansnya adalah:

$$\begin{pmatrix} 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 2276 & 1508 & 813 \\ 1508 & 2206 & 1349 \\ 813 & 1349 & 1865 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = 1466$$

Distribusi Normal Bivariat

Untuk lebih memahami distribusi normal multivariat, kita dapat melihat distribusi normal bivariat. Berikut adalah dua variabel X_1 dan X_2 yang berdistribusi normal bivariat dengan komponen vektor mean μ_1 dan μ_2, serta matriks kovarians seperti ditunjukkan di bawah ini:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

Matriks Kovarians dan Determinan

Dalam kasus ini, varians untuk kedua variabel terdapat pada diagonal matriks, dan kovarians antara dua variabel berada di luar diagonal. Kovarians ini setara dengan korelasi dikalikan dengan produk dari dua simpangan baku. Determinan dari matriks kovarians adalah:

$$|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

Invers dari matriks kovarians berbentuk sebagai berikut:

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

Fungsi Kepadatan Probabilitas Gabungan

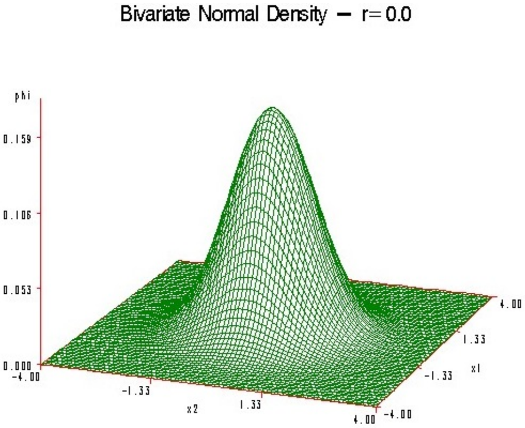
Dengan mengganti ekspresi untuk determinan dan invers dari matriks kovarians, kita mendapatkan fungsi kepadatan probabilitas gabungan dari (X_1, X_2) untuk distribusi normal bivariat sebagai berikut:

$$\phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right) \right]$$

Plot $\rho = 0.0$

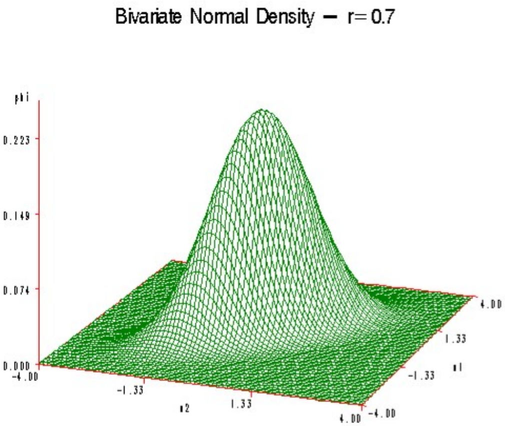
Tiga plot berikut ini adalah plot distribusi bivariat untuk berbagai nilai untuk baris korelasi.

- Plot pertama menunjukkan kasus di mana korelasi ρ sama dengan nol. Kasus khusus ini disebut distribusi normal melingkar. Di sini, kita memiliki kurva berbentuk lonceng yang simetris sempurna dalam tiga dimensi.



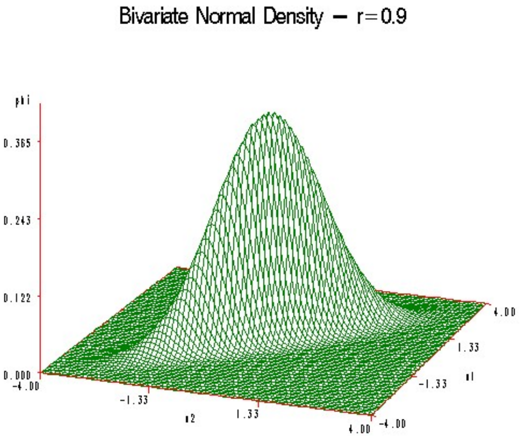
Plot $\rho = 0.7$

- Seiring dengan meningkatnya ρ, kurva berbentuk lonceng tersebut menjadi rata pada garis 45 derajat. Jadi, untuk $\rho = 0.7$, kita dapat melihat bahwa kurva memanjang ke arah minus 4 dan plus 4, dan menjadi rata pada arah tegak lurus.



Plot $\rho = 0.9$

- Meningkat ke 0.9, kurva menjadi lebih luas dan garis 45 derajat dan bahkan lebih datar lagi pada arah tegak lurus.



Eksponen dari Distribusi Normal Multivariat

Ingat kembali fungsi kepadatan distribusi normal multivariat berikut:

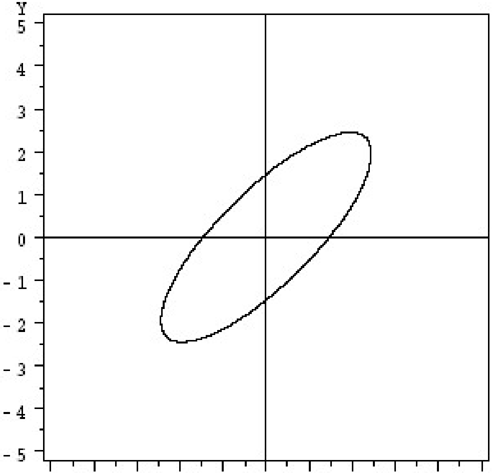
$$\phi(x) = \left(\frac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

Anda akan melihat bahwa fungsi kepadatan ini, $\phi(x)$, hanya bergantung pada x melalui jarak Mahalanobis kuadrat:

$$(x-\mu)^\top \Sigma^{-1}(x-\mu)$$

Ini adalah persamaan untuk hiper-ellips yang berpusat pada μ.

- Untuk normal bivariat, di mana $p = 2$ variabel, kita memiliki elips seperti yang ditunjukkan pada plot di bawah ini:



Fakta-fakta Berguna tentang Komponen Eksponen

- Semua nilai x sehingga $(x-\mu)^\top \Sigma^{-1}(x-\mu) = c$ untuk nilai konstan c memiliki nilai kepadatan yang sama $f(x)$ dan dengan demikian memiliki kemungkinan yang sama.
- Saat nilai $(x-\mu)^\top \Sigma^{-1}(x-\mu)$ meningkat, nilai fungsi kepadatan berkurang.
- Variabel $d^2 = (x-\mu)^\top \Sigma^{-1}(x-\mu)$ memiliki distribusi chi-square dengan p derajat kebebasan.
- Nilai d^2 untuk suatu observasi spesifik $\mathbf{x}_j$ disebut jarak Mahalanobis kuadrat.

Jarak Mahalanobis Kuadrat

Jarak Mahalanobis kuadrat d_j^2 dihitung sebagai berikut:

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^\top \Sigma^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

Jika kita mendefinisikan hiper-ellips tertentu dengan menetapkan jarak Mahalanobis kuadrat sama dengan nilai kritis dari distribusi chi-square dengan p derajat kebebasan, dan mengevaluasi pada α, maka probabilitas bahwa nilai acak X akan jatuh di dalam ellips tersebut adalah:

$$\Pr\left[(x - \mu)^\top \Sigma^{-1} (x - \mu) \leq \chi^2_{p,\alpha}\right] = 1 - \alpha$$

Ellips ini disebut sebagai ellips prediksi tingkat kepercayaan $(1 - \alpha) \times 100\%$ untuk vektor acak normal multivariat dengan vektor rata-rata μ dan matriks kovarians Σ.

Q-Q Plot untuk Evaluasi Normalitas Multivariat dan Outliers

Variabel $d^2 = (x - \mu)^\top \Sigma^{-1}(x - \mu)$ memiliki distribusi chi-square dengan p derajat kebebasan, dan untuk sampel "besar", jarak Mahalanobis yang diamati memiliki distribusi mendekati chi-square. Hasil ini dapat digunakan untuk mengevaluasi apakah titik data mungkin outlier dan apakah data yang diamati mungkin memiliki distribusi normal multivariat.

Q-Q Plot digunakan untuk menggambarkan jarak Mahalanobis untuk sampel. Jarak Mahalanobis terurut diplot melawan kuantil yang diestimasi (persentil) untuk sampel berukuran n dari distribusi chi-square dengan p derajat kebebasan.

menentukan Kuantil

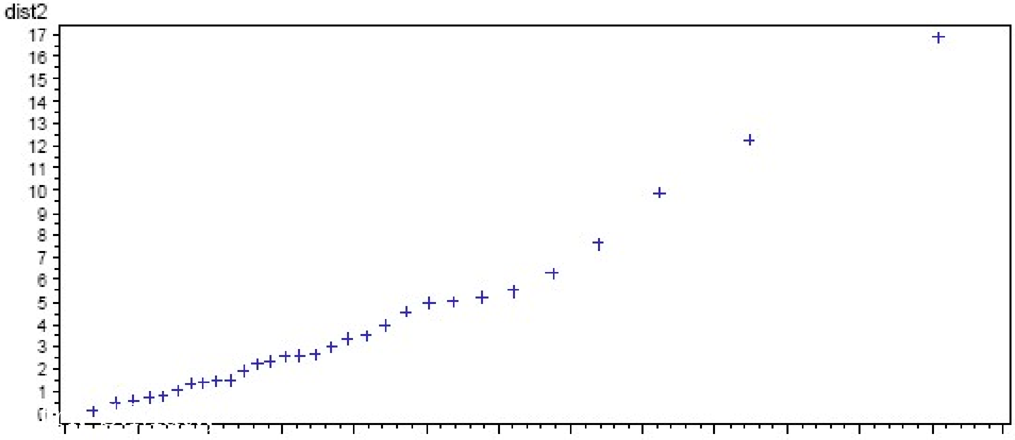
- Kuantil ke-i ditentukan sebagai nilai chi-square (dengan df $= p$) untuk probabilitas kumulatif $\frac{i-0.5}{n}$.
- Untuk menghitung seluruh set kuantil chi-square yang diestimasi, lakukan untuk nilai i dari 1 hingga n.

Contoh 4-2: Q-Q Plot untuk Data Kekakuan Papan

Pada contoh ini, kita memiliki $n = 30$ papan dengan $p = 4$ pengukuran kekakuan papan, menggunakan metode yang berbeda. Jarak Mahalanobis diplot melawan kuantil chi-square. Plot menunjukkan lengkungan ke atas di sisi kanan, yang menandakan outlier potensial dan pelanggaran normalitas multivariat.

Khususnya, titik terakhir memiliki $d^2 \approx 16$ sedangkan kuantil pada sumbu horizontal adalah sekitar 12.5. Titik berikutnya mungkin juga outlier.

SAS Plot of the Mahbalanobis Distance



Nilai Eigen dan Vektor Eigen

Langkah berikutnya adalah mendeskripsikan bentuk ellips ini secara matematis, sehingga kita dapat memahami bagaimana data terdistribusi dalam beberapa dimensi. Untuk ini, kita harus mendefinisikan nilai eigen dan vektor eigen dari sebuah matriks.
Secara khusus, kita akan mempertimbangkan perhitungan nilai eigen dan vektor eigen dari sebuah matriks simetrik **A** sebagai berikut:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{pmatrix}$$

Catatan: Matriks ini disebut simetris jika elemen $a_{ij} = a_{ji}$ untuk setiap i dan j.

Nilai Eigen

Jika kita memiliki matriks $p \times p$ $\mathbf{A}$, kita akan memiliki p nilai eigen $\lambda_1, \lambda_2, \ldots, \lambda_p$. Nilai-nilai ini diperoleh dengan menyelesaikan persamaan:

$$|\mathbf{A} - \lambda \mathbf{I}| = 0$$

Pada sisi kiri, kita memiliki matriks $\mathbf{A}$ dikurangi λ dikalikan dengan matriks identitas. Menyelesaikan determinan dari matriks ini akan menghasilkan polinomial dari orde p. Dengan menyelesaikan polinomial ini, kita akan mendapatkan nilai eigen yang diinginkan.

Vektor Eigen

Vektor eigen yang sesuai $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ diperoleh dengan menyelesaikan persamaan:

$$(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{e}_j = 0$$

Ini tidak selalu menghasilkan solusi unik. Untuk mendapatkan solusi unik, kita sering memerlukan bahwa $\mathbf{e}_j^\top \mathbf{e}_j = 1$, atau bahwa jumlah kuadrat elemen-elemen $\mathbf{e}_j$ sama dengan 1.

Catatan: Vektor eigen juga bersesuaian dengan nilai eigen yang berbeda yang ortogonal.

Contoh 4-3: Pertimbangkan Matriks 2x2

Untuk menggambarkan perhitungan ini, pertimbangkan matriks korelasi **R** seperti yang ditunjukkan di bawah ini:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Kemudian, menggunakan definisi nilai eigen, kita harus menghitung determinan dari $\mathbf{R} - \lambda \mathbf{I}$, di mana **I** adalah matriks identitas.

$$|\mathbf{R} - \lambda \mathbf{I}| = \begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix}$$

Menghitung determinan menghasilkan:

$$(1-\lambda)^2 - \rho^2 = \lambda^2 - 2\lambda + 1 - \rho^2$$

Menyelesaikan untuk Nilai Eigen

Menyelesaikan persamaan kuadrat ini untuk λ:

$$\lambda^2 - 2\lambda + 1 - \rho^2 = 0$$

Menggunakan solusi umum untuk persamaan kuadrat:

$$ay^2 + by + c = 0 \quad y = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Dengan $a = 1$, $b = -2$, dan $c = 1 - \rho^2$, kita memperoleh:

$$\lambda = \frac{2 \pm \sqrt{2^2 - 4(1 - \rho^2)}}{2} = 1 \pm \sqrt{1 - \rho^2}$$

Sehingga, solusi yang kita dapatkan adalah:

$$\lambda_1 = 1 + \rho \quad \text{dan} \quad \lambda_2 = 1 - \rho$$

Menyelesaikan Vektor Eigen

Untuk memperoleh vektor eigen yang sesuai, kita harus menyelesaikan sistem persamaan:

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{e} = 0$$

Untuk masalah ini, sistem persamaannya adalah:

$$\begin{pmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Menyederhanakan menjadi:

$$(1-\lambda)e_1 + \rho e_2 = 0$$
$$\rho e_1 + (1-\lambda)e_2 = 0$$

Solusi Vektor Eigen

Dengan menyelesaikan persamaan di atas, kita memperoleh:

$$e_2 = -\frac{(1-\lambda)}{\rho} e_1$$

Dengan normalisasi $e_1^2 + e_2^2 = 1$, kita dapatkan:

$$e_1 = \frac{1}{\sqrt{2}}, \quad e_2 = \pm\frac{1}{\sqrt{2}}$$

Sehingga, vektor eigen yang diperoleh adalah:

$$\mathbf{e}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{untuk} \quad \lambda_1 = 1 + \rho$$

$$\mathbf{e}_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{untuk} \quad \lambda_2 = 1 - \rho$$

Varians Generalized

Beberapa properti dari nilai eigen matriks kovarians-varians perlu dipertimbangkan. Varians total diberikan oleh jumlah varians:

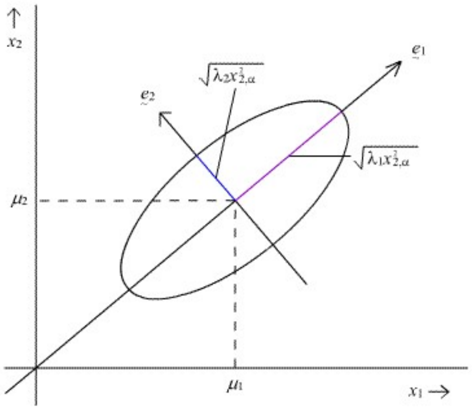
$$\sum_{j=1}^{p} s_j^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Varians generalized setara dengan hasil kali dari nilai eigen:

$$|\Sigma| = \prod_{j=1}^{p} \lambda_j = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_p$$

Geometry of the Multivariate Normal Distribution

Geometri distribusi normal multivariat dapat diselidiki dengan mempertimbangkan orientasi, dan bentuk elips prediksi seperti yang digambarkan dalam diagram berikut:



Ellips Prediksi

Ellips prediksi $(1-\alpha) \times 100\%$ ini terpusat pada rata-rata populasi μ_1 dan μ_2.

Ellips memiliki sumbu yang mengarah ke arah vektor eigen $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$. Pada diagram ini, sumbu terpanjang dari ellips mengarah pada vektor eigen pertama $\mathbf{e}_1$, dan sumbu yang lebih pendek tegak lurus dengan sumbu pertama, mengarah ke arah vektor eigen kedua $\mathbf{e}_2$.

Panjang setengah sumbu l_j diperoleh dari ekspresi berikut:

$$l_j = \sqrt{\lambda_j \chi^2_{p,\alpha}}$$

Diagram di atas menunjukkan panjang sumbu-sumbu ini di dalam ellips.

Volume (Luas) dari Hiper-Ellips

Volume (atau luas) dari hiper-ellips ini setara dengan:

$$\frac{2\pi^{p/2}}{p\Gamma\left(\frac{p}{2}\right)}\left(\chi^2_{p,\alpha}\right)^{p/2}|\Sigma|^{1/2}$$

Catatan: Ini adalah fungsi dari akar kuadrat varians generalized (diberikan oleh akar kuadrat determinan dari matriks varians-kovarians). Sehingga, volume (atau luas) dari ellips prediksi ini sebanding dengan akar kuadrat varians generalized.

Kasus I: p adalah bilangan genap

Dalam ekspresi untuk volume (atau luas) dari hiper-ellips, $\Gamma(x)$ adalah fungsi gamma. Untuk menghitung fungsi gamma, pertimbangkan dua kasus khusus:

Kasus I: p adalah bilangan genap:

$$\Gamma\left(\frac{p}{2}\right) = \left(\frac{p}{2} - 1\right)!$$

Kasus II: p adalah bilangan ganjil

Kasus II: p adalah bilangan ganjil:

$$\Gamma\left(\frac{p}{2}\right) = \frac{1 \times 3 \times 5 \times \cdots \times (p-2) \times \sqrt{\pi}}{2^{(p-1)/2}}$$

Contoh 4-4: Wechsler Adult Intelligence Scale

Pada contoh ini, kita memiliki data untuk $n = 37$ subjek yang mengambil tes Wechsler Adult Intelligence Test. Tes ini dibagi menjadi empat komponen berbeda:

- Information (Info)
- Similarities (Sim)
- Arithmetic (Arith)
- Picture Completion (Pic)

Data disimpan dalam lima kolom yang berbeda. Kolom pertama adalah ID dari subjek, diikuti oleh empat tugas komponen pada empat kolom berikutnya.

download data : Google Drive Link

Analysis

We obtain the following sample means:

Variable	Mean
Information	12.568
Similarities	9.568
Arithmetic	11.486
Picture Completion	7.973

Variance-Covariance Matrix:

$$\mathbf{S} = \begin{pmatrix} 11.474 & 9.086 & 6.383 & 2.071 \\ 9.086 & 12.086 & 5.938 & 0.544 \\ 6.383 & 5.938 & 11.090 & 1.791 \\ 2.071 & 0.544 & 1.791 & 3.694 \end{pmatrix}$$

Analysis

The eigenvalues are given below:

$$\lambda_1 = 26.245, \quad \lambda_2 = 6.255, \quad \lambda_3 = 3.932, \quad \lambda_4 = 1.912$$

Corresponding eigenvectors:

$$\mathbf{e}_1 = \begin{pmatrix} 0.606 \\ 0.605 \\ 0.505 \\ 0.110 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} -0.218 \\ -0.496 \\ 0.795 \\ 0.274 \end{pmatrix}$$

$$\mathbf{e}_3 = \begin{pmatrix} 0.461 \\ -0.320 \\ -0.335 \\ 0.757 \end{pmatrix}, \quad \mathbf{e}_4 = \begin{pmatrix} -0.611 \\ 0.535 \\ -0.035 \\ 0.582 \end{pmatrix}$$

95% Prediction Ellipse

The formula for the half-lengths of the axis of this ellipse is:

$$l_j = \sqrt{\lambda_j \chi^2_{p,\alpha}}$$

For l_1, the largest eigenvalue:

$$l_1 = \sqrt{26.245 \times 9.49} = 15.782$$

For l_2, the second largest eigenvalue:

$$l_2 = \sqrt{6.255 \times 9.49} = 7.705$$

Third and Shortest Axis of Ellipse

Looking at the corresponding eigenvector, $\mathbf{e}_2$, we can see that this axis is pointed in the direction of increasing values for the third variable, Arithmetic, and decreasing values for Similarities, the second variable.

Similar calculations for the third-longest axis:

$$l_3 = \sqrt{\lambda_3 \chi^2_{4,0.05}} = \sqrt{3.931 \times 9.49} = 6.108$$

The shortest axis:

$$l_4 = \sqrt{\lambda_4 \chi^2_{4,0.05}} = \sqrt{1.912 \times 9.49} = 4.260$$

Volume of the Hyper-Ellipse

The overall shape of the ellipse can be determined by comparing the lengths of various axes. It forms a slightly squashed football shape.
The volume of the hyper-ellipse can be obtained using the formula:

$$\frac{2\pi^{p/2}}{p\Gamma\left(\frac{p}{2}\right)}\left(\chi^2_{p,\alpha}\right)^{p/2}|\Sigma|^{1/2}$$

Substituting the values:

$$\frac{2\pi^2}{4\Gamma(2)}(9.49)^2\sqrt{26.245\times 6.255\times 3.932\times 1.912}$$

$$=\frac{444.429}{4}\times\sqrt{1234.17086}=15613.132$$

Thank You!