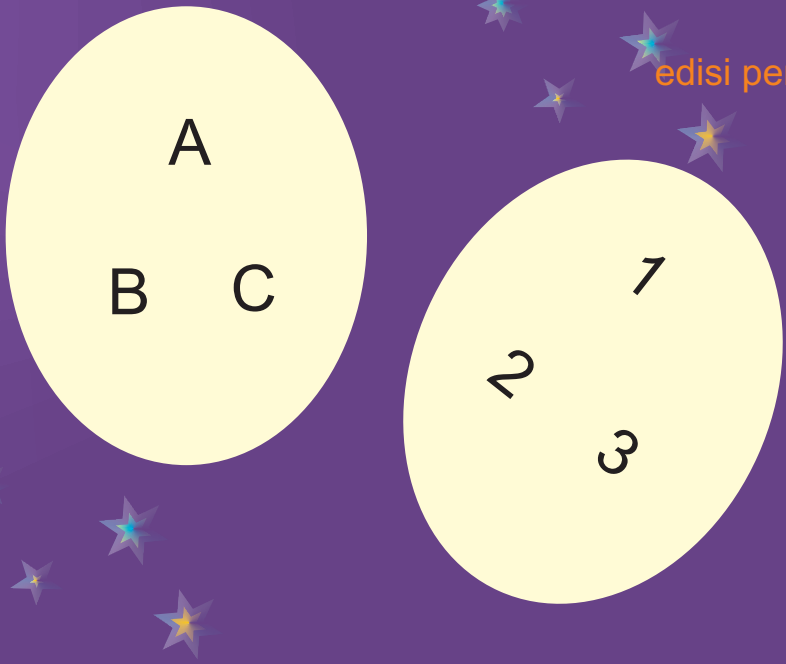
Buku Referensi

Statistika Multivariat Terapan

edisi pertama



Sigit Nugroho, Ph.D.

UNIB Press

STATISTIKA MULTIVARIAT TERAPAN

Sanksi Pelanggaran Pasal 72
Undang-Undang Nomor 19 Tahun 2002 tentang Hak Cipta

1. Barangsiapa dengan sengaja dan tanpa hak melakukan perbuatan sebagaimana dimaksud dalam Pasal 2 ayat (1) atau Pasal 49 ayat (1) dan (2) dipidana dengan pidana penjara masing-masing paling singkat 1 (satu) bulan dan/atau denda paling sedikit Rp. 1.000.000,00 (satu juta rupiah), atau pidana penjara paling lama 7 (tujuh) tahun dan/atau denda paling banyak Rp. 5.000.000.000,00 (lima miliar rupiah)

Barangsiapa dengan sengaja menyiarkan, memamerkan, mengedarkan, atau menjual kepada umum suatu ciptaan atau barang hasil pelanggaran Hak Cipta atau Hak Terkait sebagaimana dimaksud pada ayat (1) dipidana dengan pidana penjara paling lama 5 (lima) tahun dan/atau denda paling banyak Rp. 500.000.000,00 (lima ratus juta rupiah)

Statistika Multivariat Terapan

Ir. Sigit Nugroho, M.Sc., Ph.D.
Universitas Bengkulu

UNIB Press
Bengkulu

STATISTIKA MULTIVARIAT TERAPAN
Sigit Nugroho, Ph.D.
ISBN : 978-979-9431-36-3 132hal.
Cetakan Pertama. Edisi 1. 2008.

Penyeleksi Naskah : Fachri Faisal

Editor : Jose Rizal
Desain Sampul : Ratna Astuti Nugrahaeni

Diterbitkan pertama kali oleh **UNIB Press**, Jalan WR Supratman, Bengkulu.

Kata Pengantar

Statistika merupakan ilmu pengetahuan tentang data. Mulai dari bagaimana cara memperoleh data, menyajikan data, menganalisis data, bahkan sampai menginterpretasikannya merupakan bagian dari tugas ”statistika”.

Untuk data obervasi dari beberapa peubah / variabel yang diamati dari satu obyek, kita kenal dengan data peubah banyak atau multivariat. Bila dalam statistika satu peubah atau univariat kita dapat memandang data berada dalam sebuah garis lurus, maka untuk statistik multivariat, data dengan *k* peubah akan berada dalam ruang berdimensi-*k*.

Berbagai alat analisis dikenalkan dalam buku ini, mulai dari bagaimana mereduksi variabel, mencari faktor penentu dari sekelompok variabel, menentukan posisi sebuah data berdasarkan suatu kriteria, mengelompokkan data berdasarkan sifat kemiripan, dan beberapa pengembangan analisis multivariat lainnya.

Penulis mengucapkan ribuan terima kasih kepada istri *Mucharromah, Ph.D.*, anak-anak *Shofa Ulfiyati Nugrahaeni* dan *Ratna Astuti Nugrahaeni*, serta para kolega yang selama ini telah memberikan dorongan, kritik, dan saran hingga buku ini selesai disusun. Terima kasih khusus ditujukan pada mahasiswa bimbingan skripsi yang telah menyumbangkan tenaga untuk membantu penulisan ini : *Nurul Komariah, Novi Susanti, Nani Sumarni, Priska Julianti, Dian Agustina, Lisa Novianti, Noprita, Rini Handayani, Cici Suhaeni, Rosa Ayu Oktarina, Yulianti,* dan *Trisea Oktaviana*. Kritik dan saran yang membangun lainnya masih penulis harapkan dari siapa saja guna perbaikan buku ini.

Bengkulu, 20 Juli 2008

Sigit Nugroho, Ph.D.

Daftar Isi

Analisis Komponen Utama

Analisis komponen utama (AKU) merupakan analisis statistika peubah ganda yang dapat digunakan untuk mereduksi sejumlah peubah asal menjadi beberapa peubah baru yang bersifat ortogonal dan tetap mempertahankan total keragaman dari peubah asalnya.

Sifat komponen utama:

- Komponen utama yang dihasilkan saling orthogonal, saling bebas (artinya koefisien-koefisiennya bersifat orthogonal dan skor komponennya tidak berkorelasi).
- Sebagian besar keragaman cenderung berkumpul pada komponen utama pertama dan sedikit keragaman dari peubah asal terkumpul pada komponen utama urutan terakhir.

Ragam-Peragam dan Korelasi

Misalkan kita akan menggunakan notasi x_{jk} untuk menandai nilai tertentu dari variabel *ke-k* pada pengamatan *ke-i*. Sebagai konsekwensi, *n* pengukuran pada *p* variabel dapat ditulis sebagai berikut:

	Variabel 1	…	Variabel	…	Varabel *p*
Pengamatan 1:	x_{11}	…	x_{1k}	…	x_{1p}
Pengamatan 2:	x_{21}	…	x_{2k}	…	x_{2p}
…	…		…	…	…
Pengamatan *j*:	x_{j1}	…	x_{jk}	…	x_{jp}
…	…		…	…	…
Pengamatan *n*:	x_{n1}	…	x_{nk}	…	x_{np}

atau dapat ditulis dalam bentuk matriks segi empat, misalkan matriks **X** dengan *n* baris dan *p* kolom.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Rata-rata contoh secara umum dapat dihitung dari *n* pengukuran pada masing-masing *p* variable. Sedemikian sehingga akan ada *p* rata-rata contoh, yaitu:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1,2,...,p$$

Secara umum suatu ukuran tersebar disajikan oleh ragam contoh yang menggambarkan *n* pengukuran pada *p* variabel yaitu:

$$s_k^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2 \quad k = 1,2,...,p$$

dimana $\bar{x}_k$ adalah rata-rata contoh dari x_{jk} dan peragam contoh

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) \quad i = 1,2,...,p,\ k = 1,2,...,p$$

digunakan untuk mengukur kovarian contoh antar variabel *ke-i* dengan variable *ke-k*.

Ukuran yang menyangkut asosiasi linier antar dua variabel tidak tergantung pada bagian pengukuran. Koefisien korelasi contoh (*sample correlation*) untuk variabel *ke-i* dengan variabel *ke-k* digambarkan sebagai berikut:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2}}$$

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

Jika data disajikan dalam bentuk matriks maka dapat ditulis sebagai berikut:

Rata-rata contoh :

$$\underline{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdots \\ \bar{x}_p \end{bmatrix}$$

Matriks ragam-peragam contoh :

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

Matriks korelasi contoh :

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Untuk populasi rata-ratanya adalah

$$E(\underline{\mathbf{X}}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \dots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_p \end{bmatrix} = \underline{\mu}$$

dan matrik ragam-peragamnya adalah

$$\Sigma = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)'$$

$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \dots \\ X_p - \mu_p \end{bmatrix} \begin{bmatrix} X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p \end{bmatrix} \right)$$

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

$$\sum = cov(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{11} & \dots & \sigma_{p2} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

Koefisien korelasi digambarkan dalam kaitannya antara peragam σ_{ik} dengan ragam σ_{ii} dan $\sigma_{kk,}$ yaitu

$$\boldsymbol{\rho}_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

dan korelasi populasi adalah

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

$$= \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{12} & 1 & \dots & \rho_{2p} \\ \dots & \dots & \dots & \dots \\ \rho_{1p} & \rho_{2p} & \dots & 1 \end{bmatrix}$$

Jika diberikan simpangan baku matriks *pxp* adalah

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

maka $\mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2} = \sum$ dan $\boldsymbol{\rho} = (\mathbf{V}^{1/2})^{-1}\sum(\mathbf{V}^{1/2})^{-1}$

Salah satu teknik statistik yang digunakan untuk menjelaskan homogenitas kelompok adalah dengan ragam yang merupakan jumlah kuadrat semua deviasi nilai-nilai individual terhadap rata-rata kelompok. Akar varians disebut standar deviasi atau simpangan baku.

Akar ciri dan Vektor ciri

Jika *A* adalah matriks *n* x *n*, maka vektor taknol x didalam R^n dinamakan *vektor ciri* dari *A* jika $A\underline{x}$ adalah kelipatan skalar dari $\underline{x}$; yakni,

$$A\underline{x} = \lambda\underline{x}$$

untuk suatu skalar λ . Skalar λ dinamakan *akar ciri* (*eigenvalue*) dari A dan $\underline{x}$ dikatakan *vektor ciri* (*eigenvector*) yang bersesuaian dengan λ. Kata "vektor ciri" adalah gabungan bahasa Jerman dan Inggris. Dalam bahasa Jerman "ciri" dapat diterjemahkan sebagai "sebenarnya" atau "karakteristik". Oleh karena itu, akar ciri dapat juga kita namakan nilai sebenarnya atau nilai karakteristik.

Normalitas Data

Suatu data yang berbentuk distribusi normal bila jumlah data di atas dan di bawah rata-rata sama, demikian juga simpangan bakunya. Kurva dikatakan normal umum bila rata-rata dan simpangan bakunya tergantung pada nilai ketika pengumpulan data. Bentuk kurva adalah simetris, sehingga sehingga luas rata-rata ke kiri dan kanan mendekati 50%. Kurva dikatakan normal standar bila nilai rata-ratanya 0 dan simpangan bakunya 1. Kurva normal umum dapat dirubah menjadi kurva normal standar dengan rumus

$$Z = \frac{\left(X_i - \overline{X}\right)}{s}$$

Pengujian normalitas data dapat dilakukan dengan Chi kuadrat (χ^2), dilakukan dengan cara membandingkan kurva normal yang terbentuk dari data yang telah terkumpul dengan kurva kurva normal standar.

Uji normalitas sebenarnya sangat kompleks, karena harus dilakukan pada seluruh variabel secara bersama-sama. Namun uji ini dapat juga dilakukan pada setiap variabel, dengan logika jika secara individual masing-masing variabel memenuhi asumsi normalitas, maka secara bersama-sama (multivariat) variabel-variabel tersebut juga bisa dianggap memenuhi asumsi normalitas.

Komponen Utama dari Populasi

Komponen untuk sebuah populasi dimisalkan dengan kita memiliki sebuah populasi dan vektor acak yang diukur dari setiap individu pada populasi tersebut.

Dengan menggunakan matriks ragam-peragam atau matriks korelasi akan ditentukan komponen utama.

Misalkan $\boldsymbol{\Sigma}$ adalah matrik ragam-peragam dari p buah peubah $x_1, x_2, \ldots, x_p$ Komponen utama pertama dari vektor berukuran px1, $\underline{\mathbf{X}}' = \left(x_1, x_2, \ldots, x_p\right)'$ adalah kombinasi linier

$$y_1 = \underline{a}_1' \mathbf{X} = a_{11}x_1 + \ldots + a_{1p}x_p$$

dimana $\underline{a}_1' = (a_{11}, a_{12}, \ldots, a_{1p})'$ dan $\underline{a}_1'\underline{a}_1 = 1$.

Dengan definisi tersebut maka ragam dari komponen utama pertama itu adalah

$$\sigma^2_{y_1} = \underline{a}_1' \Sigma \underline{a}_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{1i}a_{1j}\sigma_{ij}$$

Vektor a_1 dipilih sedemikian rupa sehingga $\sigma^2_{y_1}$ mencapai maksimum dengan kendala seperti diatas ($\underline{a}_1'\underline{a}_1 = 1$). Dengan menggunakan teknik pemaksimuman berkendala Lagrange diperoleh persamaan

$$f(\underline{a}_1\lambda) = \sigma^2_{y_1} - \lambda(\underline{a}_1'\underline{a}_1 - 1)$$

Jika persamaan diatas diturunkan terhadap vektor a_1 kemudian disamadenganken nol didapatkan

$$\left(\frac{df(\underline{a}_1, \lambda)}{d\underline{a}_1}\right) = 2\Sigma\underline{a}_1 - 2\lambda\underline{a}_1 = 0 \qquad \text{atau} \quad \Sigma\underline{a} = \lambda\underline{a}$$

Jelas bahwa yang memenuhi persamaan diatas adalah a_1 dan λ merupakan pasangan *akar ciri* dan *vektor ciri* matriks $\boldsymbol{\Sigma}$. Perhatikan juga bahwa $\boldsymbol{\Sigma}a_1 = \lambda a_1$, mengakibatkan $a_1'\boldsymbol{\Sigma}a_1 = a_1'\lambda a_1 = \lambda\, a_1'a_1 = \lambda$. Sehingga karena diinginkan ragam y_1, $\sigma^2_{y_1} = a_1'\Sigma\, a_1 = \lambda$ maksimum, maka λ adalah *akar ciri* yang terbesar dari matriks $\boldsymbol{\Sigma}$, dan a_1 adalah *vektor ciri* yang berpadanan dengannya.

Selanjutnya komponen utama kedua adalah

$$y_2 = \underline{a}_2' X = a_{21}x_1 + \ldots + a_{2p}x_p$$

dimana $\underline{a}_2' = (a_{21}, a_{22}, \ldots, a_{2p})'$ dan $\underline{a}_2'\underline{a}_2 = 1$ dipilih sedemikian rupa sehingga $y_2 = a'_2\mathbf{X}$ ini tidak saling berkorelasi dengan $y_1 = a'_1\mathbf{X}$, dan di

antara semua kemungkinan yang bersifat demikian, y_1 memiliki ragam yang paling besar. Perhatikan bahwa ragam dari y_2 adalah

$$\sigma^2_{y_2} = \underline{a}_2' \sum \underline{a}_2$$

ingin dimaksimumkan dengan kendala $a_2'a_2 = 1$ dan

$$cov(y_1, y_2) = cov(\underline{a}_1'\mathbf{X}, \underline{a}_2'\mathbf{X}) = \underline{a}_1' \sum \underline{a}_2 = 0$$

Karena a_1 adalah *vektor ciri* dari $\mathbf{\Sigma}$ yang merupakan matriks simetris, maka

$$a_1' \sum = a_1' \sum' = (\sum a_1)' = (\lambda a_1)' = \lambda a_1'$$

Sehingga kendala

$$\underline{a}_1' \sum \underline{a}_2 = \lambda \underline{a}_1' \underline{a}_2 = 0$$

bisa dituliskan sebagai

$\underline{a}_1'\underline{a}_2 = 0$

Dengan demikian, fungsi Langrange yang dimaksimumkan adalah
$f(\underline{a}_2,\lambda_1,\lambda_2) = \underline{a}_2'\mathbf{\Sigma}\underline{a}_2 - \lambda_1(\underline{a}_2'\underline{a}_2 - 1) - \lambda_2(\underline{a}_1'\mathbf{\Sigma}\underline{a}_2 - 0)$

$= \underline{a}_2'\mathbf{\Sigma}\underline{a}_2 - \lambda_1\underline{a}_2'\underline{a}_2 - \lambda_1 - \lambda_2\underline{a}_1'\mathbf{\Sigma}\underline{a}_2$

dan dengan menurunkan terhadap vektor a_2 didapatkan

$$\frac{df(\underline{a}_2, \lambda_1, \lambda_2)}{d\underline{a}_2} = 2\sum \underline{a}_2 - 2\lambda_1 \underline{a}_2 - \lambda_2 \sum \underline{a}_1 = 0$$

Jika persamaan diatas dikalikan di depan dengan a_1',akan diperoleh

$$2\underline{a}_1' \sum \underline{a}_2 - 2\lambda_1 \underline{a}_1'\underline{a}_2 - \lambda_2 \underline{a}_1' \sum \underline{a}_1 = 0$$

$$2\underline{a}_1' \sum \underline{a}_2 - 2\lambda_1 \underline{a}_1'\underline{a}_2 - \lambda_2 \underline{a}_1' \lambda \underline{a}_1 = 0$$

$$2\underline{a}_1' \sum \underline{a}_2 - 2\lambda_1 \underline{a}_1'\underline{a}_2 - \lambda_2 \lambda \underline{a}_1'\underline{a}_1 = 0$$

$$2\underline{a}_1' \sum \underline{a}_2 - 0 - \lambda_2 \lambda = 0$$

padahal $2\underline{a}_1'\mathbf{\Sigma}\underline{a}_2 = 0$, sehingga jelas bahwa $\lambda_2 = 0$
dengan demikian, persamaan awal setelah f diturunkan terhadap a_2 menjadi

$$\frac{df(\underline{a}_2, \lambda_1, \lambda_2)}{d\underline{a}_2} = 2\sum \underline{a}_2 - 2\lambda_1 \underline{a}_2 = 0$$

$$= \sum \underline{a}_2 - \lambda_1 \underline{a}_2 = 0$$

Sehingga λ_1 dan $\underline{a}_2$ tidak lain adalah pasangan *akar ciri* dan *vektor ciri* dari matriks ragam-peragam $\boldsymbol{\Sigma}$. Seperti halnya penurunan pada pencarian $\underline{a}_1$, akan didapatkan bahwa $\underline{a}_2$ adalah *vektor ciri* yang berpadanan dengan *akar ciri* terbesar kedua dari matriks $\boldsymbol{\Sigma}$**.** Logika yang sama digunakan untuk mendapatkan komponen utama yang lain.

Akar ciri dari matriks ragam-peragam $\boldsymbol{\Sigma}$ adalah ragam dari komponen-komponen utama itu sendiri atau $var(y_1) = \lambda_1$, $var(y_2) = \lambda_2$, ..., $var(y_p) = \lambda_p$. Total keragaman sama dengan jumlah dari seluruh akar cirinya, yaitu

$$tr(\Sigma) = \sum_{i=1}^{p} \lambda_i$$

yang tidak lain adalah p buah komponen utama mampu menjelaskan semua total keragaman. Kontribusi dari setiap komponen utama *ke-j* terhadap total keragaman $\underline{\boldsymbol{X}}$ adalah

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

Seperti halnya *akar ciri*, unsur-unsur dari *vektor ciri* $\underline{a}_1$, $\underline{a}_2$,.., $\underline{a}_p$ juga memiliki interpretasi yang menarik. Perhatikan peragam antara peubah asal *ke-i*, x_i dengan komponen utama *ke-j*, y_j.

$$\begin{aligned} cov(x_i, y_j) &= cov(0x_1 + ... + 1x_i + ... + 0x_p, a_{j1}x_1 + ... + a_{jp}x_p) \\ &= cov(\underline{e}_i' X, \underline{a}_j' X) \\ &= \underline{e}_i' \Sigma \underline{a}_j \\ &= \underline{e}_i' \lambda_j \underline{a}_j \\ &= \lambda_j a_{ji} \end{aligned}$$

dan dengan demikian korelasi antara x_i dan y_j adalah

$$corr(x_i, y_j) = a_{ji} \sqrt{\frac{\lambda_j}{var(x_i)}}$$

Jadi peubah asal dengan koefisien yang lebih besar nilainya pada suatu komponen utama, memiliki kontribusi yang lebih besar pada komponen utama tersebut.

Penggunaan matriks korelasi ($\boldsymbol{\rho}$) bisa dipandang sebagai penggunaan matrik ragam-peragam dengan terlebih dahulu membakukan setiap peubah x_i menjadi x_i* melalui transformasi pembakuan

$$x_i^* = \frac{x_i - \mu_i}{\sigma_{ii}}$$

dimana μ_i dan σ_{ii} adalah rataan dan ragam dari x_i. Perhatikan bahwa jika yang digunakan adalah matriks korelasi, maka total keragamannya adalah:

$$tr(\sum^*) = tr(R) = p = \sum_{i=1}^{p} \lambda_i$$

dan kontribusi komponen utama *ke-j* terhadap total keragaman adalah

$$\frac{\lambda_j}{p}$$

Serta korelasi antara peubah asal x_i dengan komponen utama *ke-j*, y_i adalah

$$corr(x_i, y_j) = a_{ji}\sqrt{\lambda_j} \ .$$

Sudah diungkapkan bahwa *akar ciri* dari matriks ragam-peragam (atau matriks korelasi) mewakili keragaman komponen utama, dan unsur-unsur dari *akar ciri* mewakili korelasi antara komponen utama dengan peubah asal.

Komponen Utama dari Contoh

Misalkan $x_1,...,x_n$ adalah *n* buah vektor data berukuran *px1* yang mewakili *n* buah pengamatan yang masing-masing *p* buah peubah. Misalkan $\underline{\bar{x}}$ adalah vektor rataan contoh, **S** adalah *matriks ragam-peragam contoh* berukuran *pxp*, dan **R** adalah *matriks korelasi contoh*. Komponen utama contoh dari masing masing pengamatan diperoleh dengan cara yang sama ketika membahas komponen populasi kecuali bahwa akar ciri dan vektor ciri untuk mendapatkan komponen utama didapatkan dari matriks **S** (atau **R**) bukan **Σ** (atau **ρ**). Akar-akar ciri matriks **S** (atau **R**) yaitu $\hat{\lambda}_1,\ldots,\hat{\lambda}_p$ menjadi penduga bagi nilai $\lambda_1,...,\lambda_p$. Misalkan $\underline{l}_1,...,\underline{l}_p$ adalah vektor ciri dari matriks **S** yang berpadanan dengan $\hat{\lambda}_1,\ldots,\hat{\lambda}_p$. Skor *p* komponen utama untuk pengamatan *ke-i*, diperoleh dengan cara

$$KU_{1i} = \underline{l}_1'(\underline{x}_i - \underline{\bar{x}})$$
$$KU_{2i} = \underline{l}_2'(\underline{x}_i - \underline{\bar{x}}) \quad \text{untuk } i = 1,..,n.$$
$$\cdots$$
$$KU_{pi} = \underline{l}_p'(\underline{x}_i - \underline{\bar{x}})$$

Besaran KU_1 merupakan skor komponen utama pertama, KU_j adalah skor komponen utama *ke-j*. Jika yang digunakan adalah matriks korelasi **R,** dengan kata lain $l_1,...,l_p$ adalah vektor ciri matriks **R,** maka skor dari *p* buah komponen utamanya bisa didapatkan menggunakan formula

$$KU_{1i} = \underline{l}_1^{'} D^{-1/2}(\underline{x}_i - \underline{\bar{x}})$$
$$KU_{2i} = \underline{l}_2^{'} D^{-1/2}(\underline{x}_i - \underline{\bar{x}})$$
$$\cdots$$
$$KU_{pi} = \underline{l}_p^{'} D^{-1/2}(\underline{x}_i - \underline{\bar{x}})$$

dengan $D^{-1/2}$ adalah matriks

$$\begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

dan s_{ii} adalah unsur diagonal *ke-i* dari matriks **S**. Karena pencarian komponen utama menggunakan data yang dipusatkan maka nilai komponen utama bias, positif, nol, atau negatif. Skor dua komponen pertama umumnya digunakan sebagai bagian dalam eksplorasi data. Misalkan saja *scree plot* dari keduanya digunakan untuk mengidentifikasi pengerombolan objek, atau pola-pola lain.

Matriks ragam-peragam contoh (**S**) atau matriks korelasi contoh (**R**) dengan *akar ciri* dan *vektor ciri*nya adalah $\hat{\lambda}_1,...,\hat{\lambda}_p$ dan $l_1,...,l_p$, maka komponen utama *ke-i* adalah

$$\hat{y}_i = \underline{l}_i^{'} x = l_{i1}x_1 + l_{i2}x_2 + ... + l_{ip}x_p, \quad i = 1,2,...,p$$

dimana $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$.

Varian contoh $(\hat{y}_k) = \hat{\lambda}_k, \quad k = 1,2,...,p$

dan

Covarian contohnya adalah $(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$.

Total varian contohnya adalah

$$\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + ... + \hat{\lambda}_p \qquad \text{dan } r_{\hat{y}_i, x_k} = \frac{l_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, i, k = 1, 2, ..., p$$

Tetapi pada matriks korelasi setiap peubah x_i menjadi x_i* melalui transformasi pembakuan

$$x_i^* = \frac{x_i - \mu_i}{\sigma_{ii}}$$

dimana μ_i dan σ_{ii} adalah rataan dan ragam dari x_i.

Menentukan Banyaknya Komponen Utama

Berdasarkan penjelasan diatas, tahap awal penentuan komponen utama dari vektor peubah **X** adalah mendapatkan *akar ciri* dan *vector ciri* dari matriks ragam-peragam atau matriks korelasi. Jadi titik awal analisis komponen utama bisa menggunakan matriks ragam-peragam maupun matriks korelasi. Sehingga yang sering jadi permasalahan umum adalah memilih matriks mana yang akan digunakan. Karena tidak ada hubungan yang jelas antara *akar ciri* dan *vektor ciri* matriks ragam-peragam dengan matriks korelasi dan komponen utama yang dihasilkan oleh keduanya bisa sangat berbeda. Demikian juga dengan berapa banyak komponen utama yang digunakan. Namun demikian sah juga meskipun hasilnya belum tentu sama, mengganti matriks ragam-peragam dengan matriks korelasi. Dalam banyak literatur sering kali dianjurkan untuk menggunakan matriks korelasi, kecuali ada dukungan yang cukup bahwa peubah-peubah yang ada diukur menggunakan range skala yang sama dan memiliki besaran ragam yang tidak terlalu jauh bedanya.

Perbedaan satuan pengukuran yang umumnya berimplikasi pada perbedaan keragaman peubah, menjadi salah satu pertimbangan utama penggunaan matriks korelasi. Meskipun ada juga pendapat yang mengatakan gunakan selalu matrik korelasi. Penggunaan matriks korelasi memang cukup efektif kecuali pada dua hal. Pertama, secara teori pengujian statistik terhadap akar ciri dan vektor ciri matriks korelasi jauh lebih rumit dibandingkan penggunaan matriks ragam-peragam. Kedua, dengan menggunakan matriks korelasi kita memaksakan setiap peubah memiliki ragam yang sama sehingga seringkali tujuan untuk mendapatkan peubah yang kontribusinya paling besar menjadi tidak tercapai.

Ada tiga metode yang umum digunakan untuk menentukan banyaknya komponen utama. Metode pertama didasarkan pada kumulatif proporsi keragaman total yang mampu dijelaskan. Metode ini merupakan metode yang paling banyak digunakan, dan bisa diterapkan pada penggunaan matriks korelasi maupun matriks ragam-peragam. Minimum persentasi keragaman yang mampu dijelaskan ditentukan terlebih dahulu, dan selanjutnya banyaknya komponen yang paling

kecil hingga batas itu terpenuhi dijadikan sebagai banyaknya komponen utama yang digunakan. Tidak ada patokan baku berapa batas minimum tersebut, sebagian menyebutkan 70%, 80%, bahkan ada yang 90%. Jika $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ adalah akar ciri dari matriks ragam-peragam maka proporsi kumulatif dari k komponen utama pertama adalah

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, k = 1,..., p .$$

Pada kasus penggunaan matriks korelasi maka

$$\sum_{i=1}^{p} \lambda_i = p ,$$

sehingga proporsi kumulatifnya adalah

$$\frac{1}{p} \sum_{i=1}^{k} \lambda_i , k = 1,..., p .$$

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

Metode ketiga adalah penggunaan grafik yang disebut *scree plot*. Cara ini bisa digunakan ketika titik awalnya matrik korelasi maupun ragam-peragam. *Scree plot* merupakan plot antara akar ciri λ_k dengan k. Dengan menggunakan metode ini, banyaknya komponen utama yang dipilih, yaitu k, adalah jika pada titik k tersebut plotnya curam ke kiri tapi tidak curam ke kanan. Ide yang ada dibelakang metode ini adalah bahwa banyaknya komponen utama yang dipilih sedemikian rupa sehingga selisih antara akar ciri yang berurutan sudah tidak besar lagi. Interpretasi terhadap plot ini sangat subjektif.

Analisis Faktor Eksploratori

Analisis faktor merupakan suatu metode statistik untuk menganalisis sejumlah observasi (variabel) dipandang dari segi interkorelasinya. Seperti pada analisis komponen utama, analisis faktor juga merupakan teknik mereduksi dan meringkas data. Analisis faktor berfungsi untuk mendapatkan sejumlah kecil faktor yang memiliki sifat-sifat sebagai berikut :

a. Mampu menerangkan semaksimal mungkin keragaman data
b. Faktor-faktor tersebut saling bebas.
c. Tiap-tiap faktor dapat diinterpretasikan dengan sejelas-jelasnya

Tujuan utama analisis faktor adalah memilih faktor-faktor yang dapat menjelaskan keterkaitan *(interrelationship)* antar variabel asli atau dengan kata lain, analisis faktor bertujuan untuk menjelaskan arti variabel-variabel dalam himpunan data.

Di dalam analisis faktor, setiap variabel dinyatakan sebagai suatu kombinasi linear dari faktor yang mendasari *(underlying factors)*. Jumlah *(amount)* varian yang disumbangkan oleh suatu variabel dengan variabel lainnya yang tercakup dalam analisis disebut *communality*. Kovariansi antar variabel yang diuraikan, dinyatakan dalam suatu *common factors* yang sedikit jumlahnya ditambah dengan faktor yang unik untuk setiap variabel. Faktor-faktor ini tidak secara jelas terlihat *(not overly observed)*.

Model Analisis Faktor Eksploratori :

Suatu vektor peubah acak X yang diamati dengan p komponen dan vektor rataan μ, serta matriks ragam peragam Σ atau matriks korelasi ρ, secara linear bergantung pada sejumlah peubah acak yang tak teramati, yaitu $F_1, F_2, ..., F_k$ yang disebut *common factors* dan p penyimpangan tambahan $\varepsilon_1, \varepsilon_2,..., \varepsilon_p$ yang disebut *specific factors*. Model persamaan analisis faktor dirumuskan sebagai berikut :

$$X_1 - \mu_1 = l_{11}F_1 + l_{11}F_2 + ... + l_{1k}F_k + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2k}F_k + \varepsilon_2$$
$$\vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pk}F_k + \varepsilon_p$$

Atau dalam notasi matriks :

$$X_{p\times 1} - \mu_{p\times 1} = L_{p\times k}F_{k\times 1} + \varepsilon_p$$

$$\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \quad \vdots \\ x_p-\mu_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12}\cdots & l_{1k} \\ l_{21} & l_{22}\cdots & l_{2k} \\ \vdots & \vdots & \vdots \\ l_{p1} & l_{p2} & l_{pk} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

keterangan :

f_j = faktor umum ; j = 1, 2, ..., *k ; k<p*
ε_i = faktor spesifik ; i = 1, 2, ..., *p*
μ_i = rata-rata variabel ke- i
λ_{ij} = loading untuk variabel ke-*i* ada faktor ke- *j*
l_{ij} = koefisien faktor umum = faktor loading

L_{pxk} = Matriks faktor loading = $\begin{pmatrix} l_{11} & l_{13}\cdots & l_{1k} \\ l_{21} & l_{22}\cdots & l_{2k} \\ \vdots & \vdots & \vdots \\ l_{p1} & l_{p2}\cdots & l_{pk} \end{pmatrix}$

dengan asumsi :

E (F) = 0 ; *Cov (F) = E* $(F\,F')$ *= I*

E (ε) = 0 ; *Cov (ε) = E (* $\varepsilon\,\varepsilon'$ *)* $= \psi$; ψ

merupakan matriks diagonal

F dan ε saling bebas, maka *Cov (ε, F) = E (* $\varepsilon' F$*) = 0*

Model $(X-\mu = LF+\varepsilon)$ adalah linier dalam faktor bersama. Bagian dari varian (X_i) yang dapat diterangkan oleh *k* faktor bersama disebut *communality* ke-*i*, sedangkan bagian dari varian (X_i), karena faktor spesifik disebut varian spesifik ke-*i*.

$$\sigma_{ii} = \sigma_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{ik}^2 + \psi_i = h_i^2 + \psi_i$$

keterangan : $h_i^2 = communality$ ke- *i* dan $\psi =$ varians spesifik ke- *i*

Model analisis faktor mensyaratkan bahwa hubungan antar variabel terobservasi harus linear dan nilai koefisien korelasi tak boleh nol, artinya harus benar-benar ada hubungan. Hal ini dapat dilihat dari matriks korelasi. Pembentukan matriks korelasi ρ merupakan langkah awal dalam analisis faktor. Selain matriks korelasi ρ, proses analisis faktor dapat didasarkan pada matriks kovarian $\sum$, tergantung dari kesamaan satuan variabel-variabel yang dianalisis. Matriks kovarian akan digunakan bila seluruh variabel memiliki satuan yang sama, sedangkan matriks korelasi terbebas dari masalah kesamaan satuan pengukuran

dan besarnya nilai variabel – variabel yang digunakan sehingga matriks korelasi lebih banyak digunakan. Pada penelitian ini digunakan matriks korelasi sebagai dasar analisis faktor. Selanjutnya dilakukan pengujian hipotesis matriks korelasi. Ada dua macam pengujian yang akan dilakukan yaitu :

a. Uji Bartlett

Uji Bartlett digunakan untuk menguji apakah matriks korelasi yang dihasilkan adalah matriks identitas, dimana matriks identitas mengindikasikan bahwa di antara peubah tidak terdapat korelasi. Urutan pengujiannya sebagai berikut :

1. Hipotesis yang akan diuji adalah:
   H_0 : matriks korelasi merupakan matriks identitas
   H_1 : matriks korelasi bukan merupakan matriks identitas
2. Statistik uji yang digunakan

$$\lambda^2_{obs} = -\left[(N-1) - \frac{(2p+5)}{6}\right]\ln|R|$$

3. Besaran yang digunakan
   N : jumlah observasi
   $|R|$: determinan matriks korelasi
   p : jumlah variabel
4. Kriteria Pengambilan Keputusan:
   Uji *Bartlett* akan menolak H_0 jika nilai $\lambda^2_{obs} > \lambda^2_{\alpha, p(p-1)/2}$
5. Kesimpulan
   Jika H_0 ditolak berarti matriks korelasi bukan merupakan matriks identitas, tetapi jika H_0 diterima berarti korelasi merupakan matriks identitas.

b. Statistik Kaiser Meyer Olkin (KMO)

Statistik ini digunakan untuk mengetahui apakah data observasi yang ada tersebut layak dianalisis lebih lanjut dengan analisis faktor atau tidak. Syarat untuk dapat melakukan analisis faktor adalah data dari peubah-peubah yang dianalisis harus memiliki nilai statistik KMO minimal sebesar 0,5. Rumusan KMO adalah :

$$KMO = \frac{\sum_i \sum_{\neq j} r_{ij}^2}{\sum_i \sum_{\neq j} r_{ij}^2 + \sum_i \sum_{\neq j} a_{ij}^2}, i = 1,2,...,p\ ;\ j = 1,2,...,p$$

dimana : r_{ij} : koefisien korelasi sederhana antara variabel i dan j
a_{ij} : koefisien korelasi parsial antara variabel i dan j

Penilaian uji KMO dari matriks antar variabel adalah sebagai berikut :

- $0,9 < KMO \leq 1,00$ → unit observasi sangat baik untuk analisis faktor
- $0,8 < KMO \leq 0,9$ → unit observasi baik untuk analisis faktor

- 0,7 < KMO ≤ 0,8 → unit observasi agak baik untuk analisis faktor
- 0,6 < KMO ≤ 0,7 → unit observasi lebih dari cukup untuk analisis faktor
- 0,5 < KMO ≤ 0,6 → unit observasi cukup untuk analisis faktor
- KMO ≤ 0,5 → unit observasi tidak layak untuk analisis faktor

Ekstraksi Faktor

Ekstraksi faktor merupakan langkah inti dari analisis faktor, yaitu mereduksi sejumlah variabel asli (misalkan sebanyak *p* variabel) menjadi sejumlah kecil faktor (misalkan *k* faktor), dimana $p \leq k$. Ekstraksi faktor dilakukan dengan metode komponen utama. Ada beberapa prosedur *heuristic* dan *objective* untuk menentukan *k* faktor yang akan disarikan *(extracted)* di dalam analisis faktor .

1. **Penentuan apriori**
   Jumlah faktor yang diekstrak ditentukan berdasarkan teori, hipotesis maupun penelitian sebelumnya.
2. **Kriteria akar ciri (eigen value)**
   Dalam pendekatan dengan kriteria ini, hanya faktor yang memiliki akar ciri lebih dari satu (λ>1) yang dianggap signifikan dan diikutsertakan di dalam model faktor.
3. **Kriteria persentase keragaman**
   Kriteria ini menentukan banyaknya faktor yang diekstrak berdasarkan kumulatif persentase keragaman yang dijelaskan oleh faktor berurutan mencapai suatu level tertentu yang memuaskan.
4. **Kriteria Uji Scree**
   Kriteria ini digunakan untuk menentukan sejumlah faktor yang optimum, dengan membuat *scree plot* yaitu kurva yang diperoleh dengan membuat plot antara faktor (sebagai sumbu horizontal) dengan akar cirinya (sebagai sumbu vertikal). Kemudian ketajaman kurva dilihat untuk menentukan titik keluaran *(out of points)* yaitu ketika kurva mulai menyerupai garis horizontal.

Metode komponen utama pada analisis faktor merupakan metode yang cukup sederhana. Misalkan *R* adalah matriks korelasi sampel berukuran $p \times p$. Karena matriks *R* adalah simetrik dan definit positif maka dapat dituliskan sebagai : $R = \Gamma \Lambda \Gamma'$ dengan Λ adalah $\text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$, dan $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p > 0$ adalah akar ciri matriks *R*. $\Gamma\Gamma' = \Gamma'\Gamma = I_p$. Γ adalah matriks ortogonal $p \times p$ yang kolom-kolomnya adalah vektor ciri matriks *R*, yaitu $\Gamma_1, \Gamma_2, ..., \Gamma_p$ yang berpadanan dengan vektor ciri $(\lambda_1, \lambda_2, ..., \lambda_p)$. Misalkan *k* adalah banyaknya komponen utama yang dipilih menggunakan kriteria tertentu, maka penduga matriks faktor loading $\hat{L}$ berukuran $p \times k$ adalah :

$$\hat{L}_{p\times k} = \Gamma_{p\times k}\,\Lambda^{1/2}_{k\times k}$$

$$\begin{bmatrix} \hat{l}_{11} & \hat{l}_{12} & \cdots & \hat{l}_{1k} \\ \hat{l}_{21} & \hat{l}_{22} & \cdots & \hat{l}_{2k} \\ \vdots & \vdots & & \vdots \\ \hat{l}_{p1} & \hat{l}_{p2} & \cdots & \hat{l}_{pk} \end{bmatrix} = \begin{bmatrix} \hat{\Gamma}_{11} & \hat{\Gamma}_{12} & \cdots & \hat{\Gamma}_{1k} \\ \hat{\Gamma}_{21} & \hat{\Gamma}_{22} & \cdots & \hat{\Gamma}_{2k} \\ \vdots & \vdots & & \vdots \\ \hat{\Gamma}_{p1} & \hat{\Gamma}_{p2} & \cdots & \hat{\Gamma}_{pk} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_{11}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{\lambda_{kk}} \end{bmatrix}$$

sedangkan matriks diagonal ragam khusus ψ diduga dengan $\hat{\psi}$ yaitu :

$$\hat{\psi} = \begin{bmatrix} 1-h_1^2 & 0 & \cdots & 0 \\ 0 & 1-h_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-h_p^2 \end{bmatrix}$$

$$h_i^2 = \sum_{i=1}^{k} l_{ij}^2,\ i = 1, 2, ..., p$$

Dengan demikian diperoleh model *k-faktor* dengan *L* diduga oleh $\hat{L}$ dan ψ diduga oleh $\hat{\psi}$. Faktor pertama yang terbentuk dari proses ekstraksi menyerap sebagian besar varian dari seluruh variabel, kemudian faktor kedua menyerap sebagian besar varian dari variabel, setelah diperoleh faktor pertama dengan syarat faktor kedua tidak berkorelasi *(orthogonal)* dengan faktor pertama. Begitu seterusnya, hingga faktor yang terbentuk mampu menyerap lebih dari 75% varian dari variabel asli.

Rotasi Faktor

Pada umumnya faktor-faktor yang telah diperoleh masih sulit diinterpretasikan secara langsung. Oleh karena itu dilakukan rotasi terhadap matriks L atau faktor pembobot dengan mengubah faktor penimbang awa1 menjadi faktor penimbang baru untuk meningkatkan daya interpretasi. Ada dua macam rotasi, yaitu rotasi ortogonal dan rotasi *oblique*.

Rotasi ortogonal adalah rotasi yang mempertahankan keortogonalan faktor-faktor (membuat sudut kedua sumbu faktor bersama 90°), sedangkan rotasi *oblique* tidak memperhatikan sifat ortogonal tersebut (sudut kedua sumbu faktor bersama tersebut tidak harus 90°). Rotasi ortogonal ada tiga jenis, yaitu *varimax*, *quartimax*, dan *equamax*. Dari kedua jenis rotasi ini, beberapa ahli menyarankan rotasi ortogonal terutama *varimax (variance of maximum)*, karena rotasi ini lebih mendekati kenyataan dibanding yang lain. Rotasi *varimax* adalah rotasi yang

membuat jumlah varian dari faktor yang memuat *loading* kuadrat dalam masing-masing faktor menjadi maksimum (Johnson dan Dean, 1998). Metode rotasi ini adalah memaksimalkan faktor pembobot dan mengakibatkan variabel asal hanya akan mempunyai korelasi yang tinggi dan kuat dengan faktor tertentu saja (korelasinya mendekati 1) dan memiliki korelasi yang lemah dengan faktor yang lainnya (korelasinya mendekati 0).

Jika : $\hat{L}_{p\times m}$ = matriks faktor pembobot

$T_{m\times m}$ = matriks transformasi

$\hat{L}^*_{p\times q}$ = matriks faktor pembobot yang telah dirotasikan = $\hat{L}\,T$,

maka perotasian faktor pembobot $\hat{L}$ menjadi $\hat{L}^*$ memakai metode tegak lurus *varimax*, dengan cara mengalikan faktor penimbang awal dengan suatu matriks transformasi *T* yang bersifat ortogonal yang menghasilkan matriks loading baru.

$$\hat{L}^*_{p\times m} = \hat{L}_{p\times m}\, T_{m\times m}$$

$$\hat{L}^*\hat{L}^{*\prime} = \hat{L}\,T\,T'\,\hat{L}' = \hat{L}\,I\,\hat{L}' = \hat{L}\,\hat{L}'$$

Meskipun telah dirotasi, matriks dugaan kovarian (korelasi) yang diperoleh tidak berubah karena $\hat{L}\,\hat{L}' + \hat{\Psi} = \hat{L}\,T\,T'\,\hat{L}' + \hat{\Psi} = \hat{L}^*\,\hat{L}^{*\prime} + \hat{\Psi}$, selanjutnya varian spesifik (ψ_i) dan *communality* (h_i^2) juga tidak berubah. Hasil transformasi juga tidak mengubah persentase keragaman kumulatif yang digunakan. Dengan kata lain, persentase keragaman sebelum rotasi = persentase keragaman sesudah rotasi. Untuk $m = 2$ maka $T_{(2\times 2)} = \begin{bmatrix} Cos\,\theta & -Sin\,\theta \\ Sin\,\theta & Cos\,\theta \end{bmatrix}$.

$T_{(2\times 2)}$ adalah rotasi yang berlawanan dengan perputaran arah jarum jam, sedangkan θ adalah sudut rotasi yang dilakukan.

Interpretasi Faktor

Selanjutnya, dilakukan pemberian nama (interpretasi) faktor yang telah terbentuk. Penamaan faktor didasarkan pada peubah-peubah yang mendominasi faktor tersebut, dilihat dari pola pembobot faktor, baik tanda maupun besarnya. Syarat penamaan faktor adalah subyektif, bahkan sering juga ditemukan faktor yang tidak diberi nama karena peubah-peubah yang dominan pada faktor tidak memiliki ciri yang khas.

Analisis Jalur

Menurut Carey (1998), analisis jalur dan regresi berganda saling berhubungan. Analisis jalur merupakan perluasan dari model regresi (Anonim, 1999). Perluasan ini terletak pada kelengkapan penelusuran kausal. Analisis jalur tidak hanya mengetahui berapa besarnya pengaruh namun juga variabel mana yang merupakan pengaruh langsung atau tidak langsung (Anonim, 2005).

Apabila variabel bebas dan variabel tak bebas telah dispesifikasikan berdasarkan teori yang ada, maka hubungan kausal tersebut dapat diselidiki dengan menggunakan analisis jalur. Dalam analisis jalur, variabel-variabel $x_1, x_2, \cdots, x_q$ adalah variabel-variabel *eksogenous* yaitu variabel-variabel yang tidak dipengaruhi oleh variabel-variabel lain di dalam model. Sedangkan $y_1, y_2, \cdots, y_p$ adalah variabel-variabel *endogenous* yaitu variabel-variabel yang dipengaruhi oleh variabel-variabel lain di dalam model. Variabel-variabel *eksogenous* dapat mempengaruhi variabel-variabel *endogenous*, sedangkan variabel-variabel *endogenous* dapat mempengaruhi variabel-variabel *endogenous* lainnya.

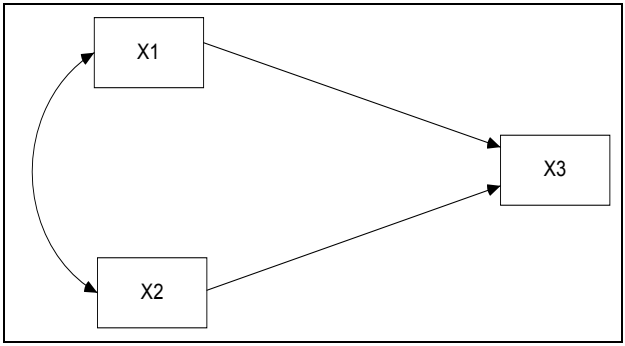
Persyaratan atau asumsi yang harus dipenuhi dalam analisis jalur adalah

1. Hubungan antara variabel bebas dan variabel tak bebas di dalam model adalah linier artinya perubahan yang terjadi pada variabel adalah merupakan fungsi perubahan linier dari variabel lainnya yang bersifat kausal.
2. Variabel yang diamati mempunyai sifat aditif artinya variabel yang mempunyai sifat *multiplikatif* dan eksponensial tidak dapat dipergunakan.
3. Variabel sisa tidak berkorelasi dengan variabel yang sesudahnya (variabel regresi lainnya).
4. Variabel yang diukur berskala interval atau rasio.

2.4.1 Model Jalur (*Path Model*)

Model jalur (*Path Model*) adalah suatu diagram hubungan variabel bebas, perantara, dan variabel tak bebas. Diagram jalur, secara grafis sangat membantu untuk melukiskan pola hubungan kausal antara sejumlah variabel (Sudjana, 2002). Satu anak panah menunjukkan penyebab antara variabel eksogenous atau variabel perantara dengan variabel tak bebas. Anak panah juga menghubungkan error dengan masing-masing variabel endogenousnya. Anak panah ganda menunjukkan korelasi antara variabel eksogenous.

Sebagai contoh gambar diagram jalur di bawah ini :



Gambar 1 Diagram Jalur dengan 2 Variabel Bebas

Keterangan :

1. Garis melengkung berarah panah menunjukkan koefisien korelasi yang bersifat simetris.
2. Garis lurus berarah panah menunjukkan pengaruh antara variabel bebas terhadap variabel tak bebas.

2.4.2 Koefisien Jalur

Koefisien jalur adalah koefisien regresi baku yang menunjukkan pengaruh langsung variabel bebas terhadap variabel tak bebas dalam *path model* (Anonim, 2001). Nilai koefisien jalur diperoleh dari hasil penyelesaian dari seperangkat persamaan. Notasi yang digunakan untuk koefisien jalur adalah P_{ij} yang berarti pengaruh variabel j terhadap variabel *i.*

Wibowo (2005) menggambarkan konsep koefisien jalur sebagai berikut. Misalkan ada m variabel bebas, yakni $X_1, X_2, \cdots, X_m$ dan X_0 sebagai variabel tak bebas. X_u merupakan notasi untuk variabel sisa (*residual*) dan semua variabel saling bekorelasi kecuali variabel sisa. Koefisien C_{0i} $(i = 1, 2, \cdots, m)$ menunjukkan sumbangan nyata X_i secara langsung terhadap X_0 serta semua hubungan adalah linier, maka persamaannya adalah

$$X_0 = C_{01}X_1 + C_{02}X_2 + \cdots + C_{0m}X_m + C_{0u}X_u$$

Jika $X_i = \dfrac{\sigma_i}{\sigma_0} V_i \qquad \forall i,\ i = 0, 1, 2, \cdots m$

Keterangan :

V_i = variabel ke-*i* ($i = 0,1,2,\cdots m$)

σ_i = simpangan baku populasi dari variabel bebas ke-*i*

σ_0 = simpangan baku populasi dari variabel tak bebas

maka

$$V_0 = C_{01}\left(\frac{\sigma_1}{\sigma_0}\right)V_1 + C_{02}\left(\frac{\sigma_2}{\sigma_0}\right)V_2 + \cdots + C_{0m}\left(\frac{\sigma_m}{\sigma_0}\right)V_m + C_{0u}\left(\frac{\sigma_u}{\sigma_0}\right)V_u$$

Apabila $P_{0i} = C_{0i}\left(\frac{\sigma_i}{\sigma_0}\right)$ maka

$$V_0 = P_{01}V_1 + P_{02}V_2 + \cdots + P_{0m}V_m + P_{0u}V_u$$

Dalam bentuk baku, semua koefisien korelasi direduksi dari hasil kali momen yaitu :

$$r_{0i} = \frac{\sum_{i=1}^{m} V_0 V_i}{n}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{m}(P_{01}V_1 + P_{02}V_2 + \cdots + P_{0m}V_m + P_{0u}V_u)V_i\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{m}(P_{01}V_1V_i + P_{02}V_2V_i + \cdots + P_{0m}V_mV_i + P_{0u}V_uV_i)\right)$$

$$= P_{01}\sum_{i=1}^{m}\frac{V_1V_i}{n} + P_{02}\sum_{i=1}^{m}\frac{V_2V_i}{n} + \cdots + P_{0m}\sum_{i=1}^{m}\frac{V_mV_i}{n} + P_{0u}\sum_{i=1}^{m}\frac{V_uV_i}{n}$$

$$= P_{01}r_{1i} + P_{02}r_{2i} + \cdots + P_{0m}r_{mi} + P_{0u}r_{ui}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m}P_{oj}r_{ji}$$

P dengan subskrip ganda menunjukkan hubungan antara variabel tak bebas dan variabel bebas. Oleh karena itu *P* adalah koefisien jalurnya dan

merupakan koefisien yang dipergunakan untuk menghitung arah panah yang tersusun dalam model.

Dari persamaan diatas, maka semua korelasi antara variabel dapat dinyatakan sebagai berikut :

$$r_{01} = P_{01}r_{11} + P_{02}r_{21} + \cdots + P_{0m}r_{m1}$$
$$r_{02} = P_{01}r_{12} + P_{02}r_{22} + \cdots + P_{0m}r_{m2}$$
$$\vdots$$
$$r_{0m} = P_{01}r_{1m} + P_{02}r_{2m} + \cdots + P_{0m}r_{mm}$$

dan bila $i = 0$ maka

$$r_{00} = P_{01}r_{10} + P_{02}r_{20} + \cdots + P_{0m}r_{m0} + P_{0u}r_{u0}$$

Jika $P_{0u} = r_{u0}$ dan karena $r_{00} = 1$ maka

$$1 = P_{01}r_{10} + P_{02}r_{20} + \cdots + P_{0m}r_{m0} + P_{0u}^2$$

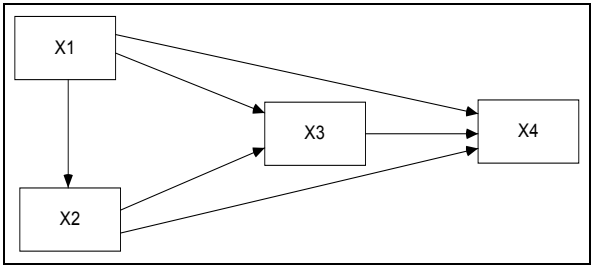
$$P_{0u}^2 = 1 - \left(P_{01}r_{10} + P_{02}r_{20} + \cdots + P_{0m}r_{m0}\right)$$

$$P_{0u}^2 = 1 - \sum_{j=1}^{m} P_{0j}r_{j0}$$

$P_{0u} = \sqrt{1 - \sum_{j=1}^{m} P_{0j}r_{j0}}$, yang merupakan rumus untuk residu atau galat untuk setiap variabel endogenus.

Apabila variabel bebas yang dilibatkan dalam pembahasan hanya satu buah, sedangkan variabel bebas dan variabel sisa tak berkorelasi maka $r_{0i} = p_{0i}$.

Selanjutnya, sebagai contoh untuk diagram jalur seperti dibawah ini :



Gambar 2 Diagram Jalur dengan 4 Variabel

Akan diperoleh persamaan strukturalnya yaitu:

$$X_1 = P_{1u} X_u$$

$$X_2 = P_{21} X_1 + P_{2u} X_u$$

$$X_3 = P_{31} X_1 + P_{32} X_2 + P_{3u} X_u$$

$$X_4 = P_{41} X_1 + P_{42} X_2 + P_{43} X_3 + P_{4u} X_u$$

Koefisien jalur yang mengindikasikan pengaruh variabel 1 terhadap variabel 2 (P_{21}) adalah:

$$r_{12} = \frac{1}{n} \sum X_1 X_2$$

$$r_{12} = \frac{1}{n} \sum X_1 \left(P_{21} X_1 + P_{2u} X_u \right)$$

$$r_{12} = P_{21} \frac{\sum X_1 X_1}{n} + P_{2u} \frac{\sum X_1 X_u}{n}$$

$P_{21} \dfrac{\sum \left(X_1 \right)^2}{n}$ merupakan koefisien jalur dengan varian X_1. Varian X_1 adalah 1.

Sedangkan $\dfrac{\sum X_1 X_u}{n}$ merupakan korelasi antara V_1 dengan error yang bernilai nol. Sehingga diperoleh

$$r_{12} = P_{21}$$

Koefisien jalur yang mengindikasikan pengaruh variabel 1 terhadap variabel 3 (P_{31}) adalah:

$$r_{13} = \frac{1}{n}\sum X_1 X_3$$

$$r_{13} = \frac{1}{n}\sum X_1 \left(P_{31} X_1 + P_{32} X_2 + P_{3u} X_u \right)$$

$$r_{13} = P_{31} \frac{\sum (X_1)^2}{n} + P_{32} \frac{\sum X_1 X_2}{n} + P_{3u} \frac{\sum X_1 X_u}{n}$$

$$r_{13} = P_{31} + P_{32} r_{12}$$

Analog untuk koefisien jalur variabel 2 terhadap variabel 3 (P_{32}) sehingga diperoleh :

$$r_{23} = P_{31} r_{12} + P_{32}$$

Koefisien jalur yang mengindikasikan pengaruh variabel 1 terhadap variabel 4 (P_{41}) adalah:

$$r_{14} = \frac{1}{n}\sum X_1 X_4$$

$$r_{14} = \frac{1}{n}\sum X_1 \left(P_{41} X_1 + P_{42} X_2 + P_{43} X_3 + P_{4u} X_u \right)$$

$$r_{14} = P_{41} \frac{\sum (X_1)^2}{n} + P_{42} \frac{\sum X_1 X_2}{n} + P_{43} \frac{\sum X_1 X_3}{n} + P_{4u} \frac{\sum X_1 X_u}{n}$$

$$r_{14} = P_{41} + P_{42} r_{12} + P_{43} r_{13}$$

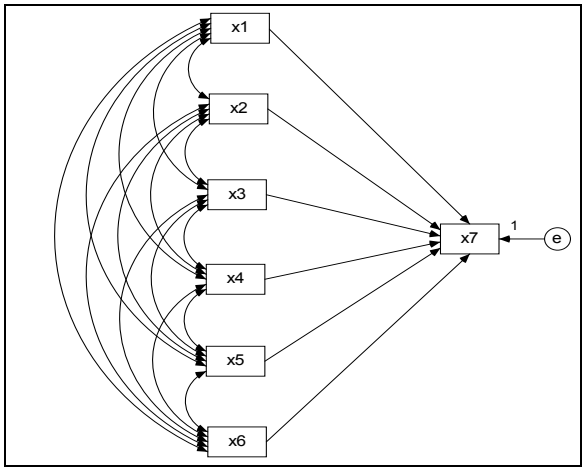
Analog untuk koefisien jalur variabel 2 terhadap variabel 4 (P_{42}) dan koefisien jalur variabel 3 terhadap variabel 4 (P_{43}) sehingga diperoleh :

$$r_{24} = P_{41} r_{12} + P_{42} + P_{43} r_{23}$$

$$r_{34} = P_{41} r_{13} + P_{42} r_{23} + P_{43}$$

Kemudian dihubungkan r_{ij} dengan P_{ij} dalam persamaan, sehingga dapat ditentukan adanya pengaruh langsung dan tidak langsung antarvariabel. Misalnya untuk $r_{14} = P_{41} + P_{42} r_{12} + P_{43} r_{13}$ yang berarti bahwa koefisien korelasi r_{14} antara variabel X_1 dan X_4 terdiri dari dua komponen. Komponen pertama adalah pengaruh langsung variabel X_1 dan X_4. Komponen kedua adalah dua pengaruh tidak langsung yaitu pengaruh tidak langsung variabel X_2 dan X_4 melalui variabel X_1 dan pengaruh tidak langsung variabel X_3 dan X_4 melalui variabel X_1.

Dalam penelitian ini , diagram jalurnya adalah :



Gambar 3 Diagram Jalur dengan 6 Variabel Bebas

Dari diagram jalur diatas, didapat persamaan :

$$X_7 = P_{71}X_1 + P_{72}X_2 + P_{73}X_3 + P_{74}X_4 + P_{75}X_5 + P_{76}V_6 + P_{7u}X_{u1}$$

Koefisien yang mengindikasikan pengaruh variabel 1 terhadap variabel 7 (P_{71}) adalah:

$$r_{17} = \frac{1}{n}\sum X_1 X_7$$

$$r_{17} = \frac{1}{n}\sum X_1 \left(P_{71}X_1 + P_{72}X_2 + P_{73}X_3 + P_{74}X_4 + P_{75}X_5 + P_{76}X_6 + P_{7u}X_u\right)$$

$$r_{17} = P_{71}\frac{\sum (X_1)^2}{n} + P_{72}\frac{\sum X_1X_2}{n} + P_{73}\frac{\sum X_1X_3}{n} + P_{74}\frac{\sum X_1X_4}{n} + P_{75}\frac{\sum X_1X_5}{n} + P_{76}\frac{\sum X_1X_6}{n} + P_{7u}\frac{\sum X_1X_u}{n}$$

$$r_{17} = P_{71} + P_{72}r_{12} + P_{73}r_{13} + P_{74}r_{14} + P_{75}r_{15} + P_{76}r_{16}$$

Analog untuk koefisien jalur variabel 2 terhadap variabel 7 (P_{72}), koefisien jalur variabel 3 terhadap variabel 7 (P_{73}), koefisien jalur variabel 4 terhadap variabel 7 (P_{74}), koefisien jalur variabel 5 terhadap variabel 7 (P_{75}), dan koefisien jalur variabel 6 terhadap variabel 7 (P_{76}).

2.4.3 Dekomposisi Korelasi

Dekomposisi korelasi adalah penguraian korelasi antarvariabel terhadap variabel endogenous. Dekomposisi ini adalah perbedaan mendasar analisis jalur dengan analisis regresi berganda.

Menurut Gaspersz (1992), apabila koefisien jalur telah diperoleh, maka beberapa informasi akan diperoleh berdasarkan metode analisis jalur, antara lain :

1. Pengaruh langsung (*Direct Effect / DE*) variabel bebas terhadap variabel tak bebas.
2. Pengaruh tak langsung (*Indirect Effect / IE*) variabel bebas terhadap variabel tak bebas.
3. Pengaruh galat (error) atau sisaan (residual), yang tidak dapat dijelaskan oleh model analisis jalur (pengaruh-pengaruh yang tidak dapat dijelaskan oleh suatu model dimasukkan sebagai galat atau sisaan).

2.4.4 Pengujian Model

Analisis jalur merupakan alat analisis yang penting untuk menguji suatu teori kausal. Melalui analisis ini dapat ditentukan ada tidaknya korelasi antar variabel yang satu dengan yang lainnya. Jika ada *m* variabel yaitu $x_1, x_2, \cdots, x_m$ dan tiap x_i dan x_j terdapat korelasi berordo *m*m* yang anggotanya adalah koefisien antara x_i dan x_j, perhitungan matriks korelasi ini selalu bisa dilaksanakan terlepas dari bentuk model yang digunakan.

Menurut Wibowo (2005), tahap-tahap untuk menentukan model terbaik adalah sebagai berikut :

1. Menghitung semua koefisien jalur dalam model. Jika koefisien arah β signifikan, maka koefisien jalur tesebut bermakna, koefisien jalur yang tidak bermakna dihilangkan. Dapat juga menggunakan kemaknaan koefisien, yaitu koefisien jalur dianggap tidak bermakna jika lebih kecil dari 0,05.
2. Menghilangkan jalur-jalur yang tidak bermakna atau tidak signifikan sehingga menjadi model yang lebih sederhana dan selanjutnya dapat dibentuk matriks korelasi baru R* dari model baru ini. Apabila matriks R* sama atau mendekati matriks R, kesimpulannya adalah model yang disederhanakan dapat dipertahankan. Bila tidak sama maka model harus diganti dengan model lain. Untuk menentukan kriteria apabila matriks R* sama atau mendekati matriks R, jika perbedaan koefisien korelasi yang sesuai kurang dari 0,05.

Pengujian kecocokan model (*model fit*) dapat digunakan statistik *Chi-Square* yang dianjurkan oleh Specht (1975) dan Pedhazur (1982). Suatu model yang diusulkan dikatakan "cocok" dengan data jika matriks korelasi model teoritis sama dengan matriks korelasi empiris (*reproduced*). Dengan demikian, perumusan hipotesis pada analisis jalur ditulis sebagai berikut :

$$H_0 : R = R(\theta)$$

$$H_1 : R \neq R(\theta)$$

Model dikatakan "cocok" atau *fit* jika hipotesis nol diterima. Untuk menguji hipotesis tersebut dapat digunakan statistik *Chi-Square* yang diusulkan Phedazur (1982) yaitu :

$$W = -(n-d)\ln(Q)$$

di mana n menunjukkan ukuran sampel dan d menunjukkan banyaknya koefisien jalur yang sama dengan nol atau koefisien jalur yang *nonsignificant*, dan *Q* adalah :

$$Q = \frac{1-R_m^2}{1-M}$$

di mana R_m^2 adalah koefisien determinan multipel untuk model yang diusulkan dan M adalah koefisien determinan multipel untuk model setelah terdapat koefisien jalur yang *nonsignificant*. Koefisien determinasi tersebut adalah :

$$M = R_m^2 = 1-\left(1-R_1^2\right)\left(1-R_2^2\right)\cdots\left(1-R_p^2\right)$$

Stasistik *W* mendekati distribusi *Chi-Square* dengan derajat bebas sebesar *d*. Jika nilai *W* sangat kecil atau mendekati nilai nol, maka hipotesis nol diterima. Dengan kata lain bahwa model yang diusulkan "cocok" dengan data.

Selain menggunakan statistik *Chi-Square*, pengujian model juga dapat dilakukan dengan menggunakan indeks kesesuaian model. Berikut ini beberapa indeks kesesuaian model dan *cut of value*-nya yang dapat digunakan untuk menguji sebuah model dapat diterima atau ditolak.

1. RMSEA (*The Root Mean Square Error of Approximation*)
   RMSEA adalah sebuah indek yang dapat digunakan untuk mengkompensasi *Chi-Square Statistics* dalam sampel yang besar (Baumggartner & Homburg, 1996). Nilai RMSEA yang lebih kecil atau sama dengan 0,08 merupakan indek untuk dapat diterimanya model yang menunjukkan sebuah *close fit* dari model berdasarkan *degrees of freedom* (Browne & Cudeck, 1993).

2. GFI (*Goodness of Fit Index*)
   GFI adalah sebuah ukuran *non-statistical* yang mempunyai rentang nilai antara 0 (*poor fit*) sampai dengan 1,0 (*perfect fit*). Nilai yang tinggi dalam indek ini menunjukkan sebuah '*better fit*'.

3. AGFI (*Adjusted Goodnes of Fit Index*)
   Tanaka & Huba (1989) menyatakan bahwa GFI adalah analog dengan R^2 dalam regresi berganda. Nilai sebesar 0,95 dapat diinterpretasikan sebagai tingkatan yang baik (*good overall model fit*) sedangkan besaran

nilai antara 0,90-0,95 menunjukkan tingkatan cukup (*adequate fit*) (Hulland *et al*., 1996).

4. CMIN (*The Minimum Sample Discrepancy Function*) / DF (*Degree of Freedom*)
   CMIN dibagi dengan *Degree of Freedom*-nya akan menghasilkan indek CMIN/DF, yang umumnya dilaporkan oleh para peneliti sebagai salah satu indikator untuk mengukur tingkat *fit*-nya sebuah model. Dalam hal ini CMIN/DF adalah satistik *Chi-Square,* χ^2 dibagi DF-nya sendiri sehingga disebut χ^2 - relatif . Nilai χ^2 - relatif kurang dari 2,0 atau bahkan kurang dari 3,0 adalah indikasi dari *acceptable fit* antara model dan data.

5. TLI (*Tucker Lewis Index*)
   TLI adalah sebuah alternatif *incremental fit index* yang membandingkan sebuah model yang diuji terhadap sebuah *baseline model* (Baumggartner & Homburg, 1996). Nilai yang direkomendasikan sebagai acuan untuk diterimanya sebuah model model adalah $\geq 0,95$ (Hair dkk, 1995), dan nilai yang sangat mendekati 1 menunjukkan *a very good fit* (Arbuckle, 1997).

6. CFI (*Comparative Fit Index*)
   Besaran indeks ini adalah pada rentang nilai sebesar 0-1. Semakin mendekati 1 mengindikasikan tingkat *fit* yang paling tinggi (Arbuckle, 1997). Sedangkan nilai yang direkomendasikan adalah CFI $\geq 0,95$.
   Keunggulan dari indek ini adalah bahwa indek ini besarnya tidak dipengaruhi oleh ukuran sampel karena iu sangat baik untuk mengukur tingkat penerimaan sebuah model (Hulland *et al*., 1996).

Indeks-indeks yang dapat digunakan untuk menguji kelayakan sebuah model diringkas dalam Tabel berikut ini :

Goodness of Fit Indices

Goodness of Fit	***Cut of Value***
Chi-Square	Diharapkan kecil
Significance Probability	$\geq 0,05$
RMSEA	$\leq 0,08$
GFI	$\geq 0,90$
AGFI	$\geq 0,90$
CMIN/DF	$\leq 2,00$
TLI	$\geq 0,95$
CFI	$\geq 0,95$

Analisis Korelasi Kanonik

Analisis korelasi kanonik pertama kali diperkenalkan oleh Hotelling (1936), sebagai suatu teknik statistika peubah ganda yang menyelediki keeratan hubungan antara dua gugus peubah. Satu gugus diidentifikasikan sebagai gugus peubah ganda (*independent variable*), dan melalui ketergantungan (*dependency*) antar kedua gugus peubah tersebut dapat dijelaskan pengaruh dari satu gugus peubah terhadap gugus peubah lainnya.

Analisis korelasi kanonik digunakan untuk mencari / mengidentifikasi serta mengukur hubungan antara dua himpunan variabel (Johnson dan Wichern, 1998). Analisis korelasi kanonik memfokuskan pada korelasi antara kombinasi linier variabel pada suatu himpunan dan kombinasi linier variabel dalam himpunan lainnya.

Gagasan pertama adalah menentukan pasangan kombinasi linier pertama yang memiliki korelasi terbesar, lalu menentukan pasangan kombinasi linier kedua yang memiliki korelasi terbesar antara semua pasangan yang tidak berkorelasi dengan pemilihan pasangan sebelumnya, begitu seterusnya.

Pada dasarnya, korelasi kanonik merupakan perluasan dari regresi ganda apabila variabel tak bebasnya lebih dari satu. Hal ini menyatakan bahwa, analisis korelasi kanonik merupakan analisis regresi ganda dengan q buah variabel tak bebas dan p buah variabel bebas. yang modelnya adalah sebagai berikut

$$Y_1, Y_2, \ldots, Y_q \quad = \quad X_1, X_2, \ldots, X_p$$

Matriks Matriks

Ciri data untuk korelasi kanonik adalah semua data untuk analisis korelasi kanonik bertipe matriks, yakni data interval atau data rasio.

Dengan demikian data bertipe nominal (seperti jenis kelamin) atau data bertipe ordinal sebaiknya tidak diproses dengan korelasi kanonik. Karena, data bertipe nominal atau ordinal merupakan data *non*matriks.

Apabila variabel tak bebasnya adalah $Y_1, Y_2, \ldots, Y_q$, dan variabel bebasnya adalah $X_1, X_2, \ldots, X_p$, maka data hasil pengamatan untuk keadaan ini adalah seperti pada matriks data berikut :

Matriks data Analisis Korelasi Kanonik

Objek (Responden)	Variabel Bebas (k buah)	Variabel Tak Bebas (m buah)
1	X_{11} X_{21} · · · X_{p1}	Y_{11} Y_{21} · · · Y_{q1}
2	X_{12} X_{22} · · · X_{p2}	Y_{12} Y_{22} · · · Y_{q2}
.	. . . . .	. . . . .
.	. . . . .	. . . . .
.	. . . . .	. . . . .
N	X_{1N} X_{2N} · · · X_{pN}	Y_{1N} Y_{2N} · · · Y_{qN}

Teori dasar tentang korelasi kanonik adalah menggunakan kombinasi linier, yang dibentuk dari variabel independen $X_1, X_2, \ldots, X_p$, dan variabel dependen $Y_1, Y_2, \ldots, Y_q$, kemudian menggunakan metoda kuadrat terkecil dicari koefisien korelasi antara kedua kombinasi linier tersebut. Koefisien korelasi yang diperoleh dengan cara demikian merupakan *koefisien korelasi kanonik* yang dicari. Di samping itu, analisis kanonik juga mampu menguraikan struktur hubungan di dalam gugus peubah bebas maupun di dalam gugus peubah tak bebas.

Dalam analisis korelasi kanonik ada beberapa asumsi yang harus dipenuhi yaitu:

1. Ada hubungan yang bersifat linier (linearitas) antar dua variabel. Seperti jika ada variabel Promosi dan variabel Penjualan, maka seharusnya korelasi antara kedua variabel bersifat linier, dalam artian makin besar pengeluaran promosi, maka makin tinggi penjualan.
2. Perlunya *Multivariate Normality* untuk menguji signifikansi setiap fungsi kanonik. Namun karena pengujian normalitas secara multivariat sulit dilakukan, maka dapat dilakukan uji normalitas untuk setiap variabel. Dengan asumsi, jika secara individu sebuah variabel memenuhi kriteria normalitas, maka secara keseluruhan juga akan memenuhi asumsi normalitas.
3. Tidak ada Multikolinearitas antar anggota kelompok variabel, baik variabel dependen maupun variabel independen.

Jika terjadi penyimpangan asumsi maka penanganan yang dapat dilakukan yaitu dengan melakukan transformasi data.

Menentukan Fungsi Kanonik dan Kesesuaian

Analisis korelasi kanonik berfokus pada korelasi antara kombinasi linier dari gugus peubah bebas dengan kombinasi linier dari peubah tak bebas. Ide utama dari analisis ini adalah mencari pasangan kombinasi linier yang memiliki korelasi terbesar. Selanjutnya, pasangan-pasangan lain diharapkan tidak berkorelasi. Pasangan dari kombinasi linier ini disebut peubah kanonik dan korelasinya disebut korelasi kanonik.

Proporsi Keragaman

Besarnya keragaman sampel yang diterangkan oleh peubah kanonik yang dipilih dapat dihitung dengan menggunakan pendekatan berikut :

Proporsi keragaman X yang diterangkan adalah :

$$R^2_{Z^{(1)}|U_1,U_2,\ldots,U_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{U_i,z_k^{(1)}}}{p}$$

Sedangkan proporsi keragaman Y yang juga diterangkan adalah:

$$R^2_{Z^{(2)}|V_1,V_2,\ldots,V_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{q} r^2_{V_i,z_k^{(2)}}}{q}$$

Besar atau kecilnya nilai proporsi keragaman menunjukkan baik atau tidaknya jumlah kanonik yang dipilih. Semakin besar nilai proporsi keragaman ini menggambarkan semakin baik peubah kanonik yang dipilih menerangkan keragaman asal.

Pendugaan Koefisien Kanonik

Misal, ingin di buat hubungan antara gugus peubah tak bebas $Y_1, Y_2, \ldots, Y_q$ yang dinotasikan dengan vektor peubah acak Y, dengan gugus peubah bebas $X_1, X_2, \ldots, X_p$ yang dinotasikan dengan vektor peubah acak X, dimana $p \le q$.

Misalkan, karakteristik dari vektor peubah acak X dan Y adalah sebagai berikut:

$$E(X^{(1)}) = \mu^{(1)}; \qquad Cov(X^{(1)}) = \sum_{11}$$
$$E(X^{(2)}) = \mu^{(2)}; \qquad Cov(X^{(2)}) = \sum_{22}$$
$$Cov(X^{(1)}, X^{(2)}) = \sum_{12} = \sum'_{21}$$

Kombinasi linear dari kedua gugus peubah tersebut dapat ditulis sebagai berikut:

$$U = \underline{a}'X^{(1)} = a_1 X_1^{(1)} + a_2 X_2^{(1)} + \cdots + a_p X_p^{(1)}$$
$$V = \underline{b}'X^{(2)} = b_1 X_1^{(2)} + b_2 X_2^{(2)} + \cdots + b_q X_q^{(2)}$$

Sehingga,

$$Var\ (U) = \underline{a}'Cov(X^{(1)})\, a = \underline{a}'\sum_{11} a$$
$$Var\ (V) = \underline{b}'Cov(X^{(2)})\, b = \underline{b}'\sum_{22} b$$
$$Cov\ (U,V) = \underline{a}'Cov\ (X^{(1)}, X^{(2)})\, b = \underline{a}'\sum_{12} b$$

Dari sini dicari koefisien vektor a dan b sehingga,

$$Corr\ (U,V) = \frac{\underline{a}'\sum_{12} b}{\sqrt{\underline{a}'\sum_{11} a}\sqrt{\underline{b}'\sum_{22} b}}$$ sebesar mungkin.

Sehingga dapat didefenisikan bahwa pasangan pertama dari peubah kanonik adalah kombinasi linier U_1, V_1 yang memiliki ragam satu dan korelasi terbesar , pasangan kedua dari peubah kanonik adalah kombinasi linier U_2, V_2 yang memiliki ragam satu dan korelasi terbesar kedua serta tidak berkorelasi dengan peubah kanonik pertama, pasangan ke-k dari peubah kanonik adalah kombinasi linier U_k, V_k yang memiliki ragam satu dan korelasi terbesar ke-k serta tidak berkorelasi dengan peubah kanonik *1, 2, ..., k-1*.

Dengan demikian dapat dituliskan :
Peubah kanonik pertama :

$$U_1 = \underline{a}'_1 X^{(1)} \qquad Var = (U_1) = 1$$
$$V_1 = \underline{b}'_1 X^{(2)} \qquad Var = (V_1) = 1$$
$$\text{Maksimum } Corr(U_1, V_{1)} = \rho_1^*$$

Peubah kanonik kedua :

$$U_2 = \underline{a}'_2 X^{(1)} \qquad Var = (U_2) = 1 \qquad Cov(U_1, U_2) = 0$$

$$V_2 = \underline{b}'_2 X^{(2)} \qquad Var = (V_2) = 1 \qquad Cov(V_1, V_2) = 0$$

Maksimum $Corr(U_2, V_2) = \rho_2^*$

Peubah kanonik *ke – k* :

$$U_k = \underline{a}'_k X^{(1)} \qquad Var = (U_k) = 1 \qquad Cov(U_k, U_1) = 0$$

$$V_k = \underline{b}'_k X^{(2)} \qquad Var = (V_k) = 1 \qquad Cov(V_1, V_k) = 0$$

Maksimum $Corr(U_k, V_k) = \rho_k^*$

Teorema Ketaksamaan Cauchy-Schwarz

Misalkan b dan d adalah vektor $p \times 1$, dengan $(b'd)^2 \le (b'b)(d'd)$ jika dan hanya jika $b = c\,d$ (atau $d = c\,b$) untuk beberapa nilai c konstan.

Dengan menggunakan ketaksamaan Cauchy-Schwarz atau metode langrange maka diperoleh $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ adalah nilai eigen dari matriks $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ yang berpadanan dengan vektor eigen $f_1, f_2, \cdots, f_p$. Disamping itu, $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ juga merupakan nilai eigen dari matriks $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$ yang berpadanan dengan vektor eigen $e_1, e_2, \cdots, e_p$.

Diasumsikan bahwa Σ_{11} dan Σ_{22} adalah *non*singuler. Diketahui bahwa $\Sigma_{11}^{-1/2}$ dan $\Sigma_{22}^{-1/2}$ merupakan matriks akar kuadrat simetrik, dengan $\Sigma_{11} = \Sigma_{11}^{1/2}\Sigma_{11}^{1/2}$ dan $\Sigma_{11}^{-1} = \Sigma_{11}^{-1/2}\Sigma_{11}^{-1/2}$.

Misalkan $c = \Sigma_{11}^{1/2}$ dan $d = \Sigma_{22}^{1/2}$,

Selanjutnya dapat diperoleh

$$a = \Sigma_{11}^{-1/2}\, c \text{ dan } b = \Sigma_{22}^{-1/2}\, d$$

maka,

$$Corr\,(\underline{a}' X^{(1)}, \underline{b}' X^{(2)}) = \frac{\underline{a}' \Sigma_{12}\, b}{\sqrt{\underline{a}' \Sigma_{11}\, a}\,\sqrt{\underline{b}' \Sigma_{22}\, b}} = \frac{c' \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}\, d}{\sqrt{c'c}\,\sqrt{d'd}}$$

dari ketaksamaan Cauchy-Schwarz diperoleh

$$c'\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\,d \le \left(c'\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{21}\Sigma_{11}^{-1/2}\,c\right)^{1/2}\left(d'd\right)^{1/2}$$

Karena $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$ adalah matriks simetrik $p \times p$, dengan hasil maksimalnya adalah

$$c'\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\,c \le \lambda_1\, c'c$$

dimana λ_1 adalah nilai eigen terbesar dari $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$, yang diperoleh dalam persamaan diatas. Untuk $c = e_i$, merupakan nilai eigen yang sebanding dengan λ_1. Dari persamaan tersebut juga dapat dilihat jika d sepadan dengan $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}e_1$. Karena itu,

$$\max_{a,b} Corr(\underline{a}'X^{(1)}, \underline{b}'X^{(2)}) = \sqrt{\lambda_1}$$

persamaan berlaku untuk $a = \Sigma_{11}^{-1/2}c = \Sigma_{11}^{-1/2}e_1$ dan dengan b sepadan dengan $\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}e_1$, dimana *b* merupakan lambang untuk korelasi positif. Di ambil $b = \Sigma_{22}^{-1/2}f_1$. Kemudian mengalikan e_1 dikedua sisi, diperoleh $(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2})\; e_1 = \lambda_1\, e_1$

Dari $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$, dihasilkan

$$\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}(\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\,e_1) = \lambda_1\,(\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\,e_1)$$

Karena itu, (λ_1, e_1) adalah pasangan dari nilai vektor-nilai eigen $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$, dan (λ_1, f_1) dengan f_1 membentuk $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\,e_1$ merupakan pasangan dari nilai vektor-nilai eigen untuk $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$. Lambang pada f_1 dipilih untuk memberikan korelasi positif.

Telah ditunjukkan bahwa $U_1 = e'_1\,\Sigma_{11}^{-1/2}X^{(1)}$ dan $V_1 = f'_1\,\Sigma_{22}^{-1/2}X^{(2)}$ merupakan pasangan pertama dari peubah kanonik, yang korelasinya adalah $\rho_1^* = \sqrt{\lambda_1}$, dan $Var(U_1) = e'_1\,\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}\,e_1 = e'_1\,e_1 = 1$, sama halnya dengan $Var(V_1) = 1$.

Selanjutnya diketahui bahwa U_1 merupakan kombinasi linier $\underline{a}'X^{(1)} = c'\Sigma_{11}^{-1/2}X^{(1)}$ yang berubah-ubah dan tidak berkorelasi karena $0 = Cov(U_1, c'\Sigma_{11}^{-1/2}X^{(1)}) = c'\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}\, c = e'_1\, c,$ atau $c \perp e_1$

Pada langkah *ke-k*, ambil $c \perp e_1, e_2, \ldots, e_{k-1}$, menghasilkan $c'\sum_{11}^{-1/2}\sum_{12}\sum_{22}^{-1/2}\sum_{21}\sum_{11}^{-1/2} c \le \lambda_1 c'c$ untuk beberapa $c \perp e_1, \ldots, e_{k-1.}$
Selanjutnya,

$$Corr\,(\underline{a}'X^{(1)}, \underline{b}'X^{(2)}) = \frac{c'\Sigma_{11}^{-1/2}\Sigma_{12}\,\Sigma_{22}^{-1/2}\,d}{\sqrt{c'c}\sqrt{d'd}} \le \sqrt{\lambda_k}$$

untuk persamaan $c = e_k$ atau $a = \Sigma_{11}^{-1/2}e_k$ dan $b = \Sigma_{22}^{-1/2}f_k$. Karena itu, $U_k = e'_k\,\Sigma_{11}^{-1/2}X^{(1)}$ dan $V_k = f'_k\,\Sigma_{22}^{-1/2}X^{(2)}$ adalah pasangan *ke-k* dari peubah kanonik, yang mempunyai korelasi $\sqrt{\lambda_k} = \rho_k^*$.
Sehingga vektor koefisien a dan b diperoleh sebagai berikut :

$$a_1 = e_1\,\Sigma_{11}^{-1/2} \qquad b_1 = f_1\,\Sigma_{22}^{-1/2}$$
$$a_2 = e_2\,\Sigma_{11}^{-1/2} \qquad b_2 = f_2\,\Sigma_{22}^{-1/2}$$
$$\ldots\ldots \qquad\qquad \ldots\ldots$$
$$a_p = e_p\,\Sigma_{11}^{-1/2} \qquad b_p = f_p\,\Sigma_{22}^{-1/2}$$

Hipotesis Dalam Analisis Korelasi Kanonik

Ada beberapa hipotesis yang dapat diuji dalam analisis korelasi kanonik yaitu, ketika $\sum_{12} = 0$, $a'X^{(1)}$ dan $b'X^{(2)}$ mempunyai kovarian $a'\sum_{12} b = 0$ untuk semua vektor a dan b. Sebagai konsekuen, semua korelasi kanonik nilainya adalah nol. Hasil selanjutnya, adalah menguji $\Sigma_{12} = 0$, untuk contoh besar.

Misalkan $X_j = \begin{bmatrix} X_j^{(1)} \\ \ldots\ldots \\ X_j^{(2)} \end{bmatrix}$, $j = 1,2,...,n$

dari contoh acak $N_{p+q}(\mu,\Sigma)$ dengan populasi $\sum=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$

Bentuk hipotesisnya adalah sebagai berikut

$$H_0 : \Sigma_{12} = 0 \qquad vs \qquad H_1 : \Sigma_{12} \neq 0$$

Implikasinya, apabila $\Sigma_{12} = 0$ maka $a'\Sigma_{12}\, b = 0$. Sehingga semua korelasi kanoniknya akan bernilai nol.

Hipotesis nol ditolak jika nilai berikut besar,

$$-2\ \ln\ \Lambda = n\ \ln\left(\frac{|S_{11}|\,|S_{22}|}{|S|}\right) = -n\ \ln\ \prod_{i=1}^{p}(1-\hat{\rho}_i^{*2})$$

dimana, $S=\begin{bmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{bmatrix}$ adalah penduga tak bias dari Σ.

Untuk contoh kasus besar maka statistik uji ini diaproksimasikan menyebar Kai-kuadrat dengan derajat bebas pq.

Bartlett menyarankan mengganti *n* dalam statistik rasio kemungkinan dengan $n-1-\frac{1}{2}(p+q+1)$ untuk mendekati sebaran contoh dari $-2\ \ln\ \Lambda$ dengan sebaran Kai-kuadrat. Sehingga untuk *n* dan *n-(p+q)* besar, $H_0 : \Sigma_{12} = 0$ akan ditolak jika,

$$-n(n-1-\frac{1}{2}(p+q+1))\ \ln\ \prod_{i=1}^{p}(1-\hat{\rho}_i^{*2}) > \chi^2_{pq}(\alpha)$$

Untuk hipotesis kedua, jika hipotesis ditolak maka ada beberapa peubah kanonik yang berkorelasi. Hipotesis dapat ditulis sebagai berikut,

$$H_0^k : \rho_1^* \neq 0, \rho_2^* \neq 0, \cdots, \rho_k^* \neq 0, \rho_{k+1}^* = \cdots = \rho_1^* = 0$$

$$H_1^k : \rho_1^* \neq 0, \text{ untuk beberapa } i \geq k+1.$$

H_0^k ditolak pada taraf α, jika

$$-n(n-1-\frac{1}{2}(p+q+1))\ \ln\ \prod_{i=k+1}^{p}(1-\hat{\rho}_i^{*2}) > \chi^2_{(p-k)(q-k)}\ (\alpha).$$

Tahap-tahap yang akan dilakukan dalam penulisan skripsi ini adalah sebagai berikut

1. Membangkitkan data berdistribusi normal yang terdiri atas variabel bebas X_1 hingga X_5 dan variabel tak bebas Y_1 hingga Y_2, yang terdiri dari 25 dan 50 sampel.
2. Menentukan kombinasi linear dari masing-masing variabel bebas (X_i) dan variabel tak bebas (Y_i).
3. Mengecek asumsi-asumsi dalam analisi korelasi kanonik.
4. Menentukan fungsi kanonik dan kesesuaiannya.
5. Menduga model regresi antara variabel-variabel kanonik dengan menggunakan Metode kuadrat terkecil.

Analisis Diskriminan

Dalam analisis diskriminan ada dua asumsi yang harus dipenuhi, yaitu :

1. Memiliki distribusi multivariat normal
2. Memiliki kesamaan marik varians kovarians antar grup.

Gambaran konsep analisis diskriminan adalah sebagai berikut. Misalkan ada suatu populasi P terdiri dari g grup yang saling asing dan misalkan π_i adalah proporsi P dalam grup $G_i (i = 1,2,...,g; \quad \sum_i \pi_i = 1)$. $f_i(\mathbf{x})$ didefinisikan sebagai peluang atau fungsi kepekatan peluang (fkp) dari $\mathbf{x}$ jika $\mathbf{x} \in G_i$. Dari populasi P ingin didapatkan suatu pembagian yang cocok $\{R_1, R_2, ..., R_g\}$ sedemikian rupa sehingga anggota P dapat digolongkan ke G_i jika $\mathbf{x} \in R_i$, dimana $R_1 \cup R_2 \cup ... \cup R_g = \mathbf{R}$, yaitu seluruh ruang contoh dari P dan $R_1, R_2, ..., R_g$ saling asing. Peluang untuk menggolongkan anggota P ke G_j bila ternyata anggota tersebut berasal dari G_i adalah

$$\mathrm{P}(j/i) = \int_{R_j} f_i(\mathbf{x}) d\mathbf{x}$$

dimana, $\mathrm{P}(j/i)$ = P(pengamatan yang berasal dari G_i dan salah klasifikasi ke G_j)

$$= \mathrm{P}(\mathbf{x} \in R_j / G_i).$$

sehingga, peluang salah mengklasifikasikan anggota G_i adalah

$$\mathrm{P}(i) = \sum_{j=1, j \neq i}^{g} \mathrm{P}(j/i) = 1 - \mathrm{P}(i/i)$$

dimana, $\mathrm{P}(i)$ = peluang salah mengklasifikasikan pengamatan yang berasal dari G_i.

$\mathrm{P}(i/i) = \mathrm{P}$ (pengamatan yang berasal dari G_i dan dengan benar terklasifikasi ke G_i).

Total peluang salah pengklasifikasian adalah

$$P(\mathbf{R},\mathbf{f}) = \sum_{i=1}^{g} P(\text{kesalahan menggolongkan anggota } G_i)$$

$$= \sum_{i=1, i\neq j}^{g} P(\mathbf{x} \text{ digolongkan ke } G_j / \mathbf{x} \in G_i) \,.\, P(\mathbf{x} \in G_i)$$

$$= \sum_{i=1, i\neq j}^{g} P(j/i)\pi_i$$

$$= \sum_{i=1}^{g} P(i)\pi_i$$

$$P(\mathbf{R},\mathbf{f}) = 1 - \sum_{i=1}^{g} P(i/i)\pi_i \,.$$

Jika $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_k \end{pmatrix}$ adalah suatu vektor pengamatan yang memiliki distribusi *multivariate normal* dengan vektor rata-rata $\boldsymbol{\mu}$ dan matrik varian kovarian $\boldsymbol{\Sigma}$, maka fungsi kepekatan peluang (fkp) dari $\mathbf{x}$ adalah

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^k \left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

dimana k adalah jumlah variabel. Ketika $\mathbf{x}$ memiliki fkp seperti diatas, dapat dikatakan bahwa $\mathbf{x}$ berdistribusi $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Jika parameter $\boldsymbol{\mu}$ dan $\boldsymbol{\Sigma}$ tidak diketahui, maka vektor rata-rata dan matrik varians kovarians dapat diganti oleh masing-masing penduganya, yaitu $\boldsymbol{\mu} = \bar{\mathbf{x}}$ dan $\boldsymbol{\Sigma} = \mathbf{S}$.

Jika ada sebuah sampel dengan ukuran *N* yang terdiri dari *k* buah variabel $x_1, x_2, \ldots, x_k$ maka data pengamatan untuk sampel ini dapat disajikan dalam bentuk matriks

Variabel	X_1	X_2	$\cdots$	X_k
Data Pengamatan	X_{11}	X_{21}	$\cdots$	X_{k1}
	X_{12}	X_{22}	$\cdots$	X_{k2}
	$\cdots$		$\cdots$	$\cdots$
	X_{13}	X_{23}	$\cdots$	X_{kN}

Untuk variabel X_j $(j = 1,2,..,k)$ dapat dihitung variansnya, dengan rumus

$$\mathbf{S}_{jj} = \frac{N\sum_{n=1}^{N} X_{jn}^2 - \left(\sum_{n=1}^{N} X_{jn}\right)^2}{N(N-1)}$$

Semuanya ada *k* buah varians, yaitu $\mathbf{S}_{11}, \mathbf{S}_{22}, ..., \mathbf{S}_{kk}$ yang masing-masing merupakan varians untuk variabel $X_1, X_2, ..., X_k$. Antara X_i dan X_j untuk $i \neq j$, terdapat kovarians yang dapat dihitung dengan rumus

$$\mathbf{S}_{ij} = \frac{N\sum_{n=1}^{N} X_{in}X_{jn} - \left(\sum_{n=1}^{N} X_{in}\right)\left(\sum_{n=1}^{N} X_{jn}\right)}{N(N-1)}$$

Semuanya ada $(k^2 - k)$ buah kovarians. Varians dan kovarians ini disusun dalam sebuah matriks, yang dikenal dengan matriks varians kovarians yang bentuknya adalah

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \dots & \mathbf{S}_{1k} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \dots & \mathbf{S}_{2k} \\ .... & & & \\ \mathbf{S}_{k1} & \mathbf{S}_{k2} & \dots & \mathbf{S}_{kk} \end{bmatrix}$$

dengan catatan bahwa untuk $i = j$ maka $\mathbf{S}_{ij}$ menjadi $\mathbf{S}_{jj}$, yaitu varians untuk variabel X_j dan bahwa $\mathbf{S}_{ij} = \mathbf{S}_{ji}$.

Jika ada *g* grup maka akan ada *g* buah matrik varians kovarians. Untuk *g* buah matriks varians kovarians ini bisa dihitung matriks varians kovarians gabungan S, dengan rumus

$$\mathbf{S} = \frac{1}{N-g}\sum_{i=1}^{g}(N_i - 1)\mathbf{S}_i$$

dimana $N = N_1 + N_2 + ... + N_g$.

Kemudian, aturan pengklasifikasiannya adalah: Jika ada $\mathbf{x}$ pengamatan baru yang tidak diketahui asalnya, maka dapat dihitung skor diskriminan linier

$$W_{ij} = \mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) - \tfrac{1}{2}(\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$$

dimana

$\bar{\mathbf{x}}_i$ = vektor rata-rata sampel grup ke-i

$\bar{\mathbf{x}}_j$ = vektor rata-rata sampel grup ke-j

$\mathbf{S}_i$ = matrik varian kovarian sampel grup ke-i

W_{ij} = fungsi diskriminan yang akan menggolongkan individu ke grup *i* atau grup *j*.

Sehingga diperoleh aturan pengklasifikasian sebagai berikut :

golongkan $\mathbf{x}$ ke grup i jika $W_{ij} > 0$ untuk semua $j \neq i$

dengan catatan bahwa $W_{ij} = -W_{ji}$ dan bahwa sembarang $g-1$ fungsi W_{ij} yang bebas linier membentuk suatu basis gugus statistik yang lengkap jika $g-1 \leq k$. Jika $k < g-1$, ruang dari W_{ij} akan mempunyai rank k, dan aturan pengklasifikasian dapat didefinisikan dalam bentuk *k* nilai.

Berikut adalah sebuah contoh sebagai ilustrasi dari teori di atas. Misalkan $g = 3$ dan $k \geq 2$. Fungsi diskriminannya adalah

$$W_{12} = \mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$W_{13} = \mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_3)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3)$$

$$W_{23} = \mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3) - \tfrac{1}{2}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_3)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3)$$

Karena $W_{23} = W_{13} - W_{12}$ maka hanya diperlukan statistik W_{12} dan W_{13}. Aturan pengklasifikasiannya adalah :

Golongkan x ke
- Grup 1 jika $W_{12} > 0$ dan $W_{13} > 0$
- Grup 2 jika $W_{12} < 0$ dan $W_{13} > W_{12}$
- Grup 3 jika $W_{13} < 0$ dan $W_{12} > W_{13}$.

Untuk keperluan kerja analisis diskriminan, salah satu cara untuk menghindari bias adalah dengan membagi sampel menjadi dua bagian, yaitu :

1. *Training sample ,* digunakan untuk membentuk aturan klasifikasi, yaitu dengan mengestimasi Koefisien fungsi diskriminan.
2. *Validation sample,* digunakan untuk mengevaluasi fungsi klasifikasi.

Kelemahan dari cara ini yaitu membutuhkan sampel yang cukup besar.

Proporsi pembagian sampel ini tidak harus sama besar untuk masing-masing bagian. Misalnya 25% dan 75%, 40% dan 60%, atau yang lainnya. Proses validasi terhadap fungsi diskriminan yang terbentuk di *training sample* harus dilakukan

berkali kali, yang jelas tidak cukup hanya sekali. Kadang-kala dibuat proporsi pembagian sampel sebesar 20% dan 80%.

Jika sampel tidak dibagi menjadi dua bagian, maka prosedur yang digunakan disebut sebagai prosedur *hold out*. Dalam prosedur ini, seluruh sampel difungsikan sebagai *training sample* sekaligus v*alidation sample*. Setelah sampel digunakan untuk membentuk fungsi klasifikasi, kemudian sampel tersebut digunakan lagi untuk mengevaluasi fungsi klasifikasi yang telah terbentuk. Prosedur *Hold out* ini disebut juga sebagai *leaving-one-out method* atau *cross validation*.

Untuk mengestimasi koefisien fungsi diskriminan, ada dua pendekatan yang bisa dilakukan, yaitu *direct method* dan *stepwise discriminant analysis*. *Direct method* meliputi estimasi koefisien fungsi diskriminan dimana seluruh variabel bebas dimasukkan dalam analisis secara simultan bersama-sama. Semua variabel diikutsertakan dalam analisis tanpa memperhatikan *discriminating power*. Sementara *stepwise discriminant analysis*, variabel bebas diikutsertakan secara berurutan (*sequentially*), didasarkan pada kemampuannya untuk mendiskriminasi antar kelompok.

Dalam penerapannya, kedua asumsi dalam analisis diskriminan tidak selamanya dapat dipenuhi. Pelanggaran asumsi *multivariate normal* pada analisis ini biasanya menghasilkan tingkat ketepatan klasifikasi yang rendah. Namun demikian, ada peneliti yang tetap menganjurkan penggunaan analisis diskriminan meskipun ada pelanggaran asumsi, dengan catatan tidak ada data yang outlier.

Analisis Klaster

Analisis klaster merupakan salah satu analisis multivariat yang termasuk dalam metode interdependensi yaitu variabel bebas x atau faktor penyebab tidak dibedakan dengan variabel terikat y atau respon.

Analisis klaster adalah suatu koleksi metode statistik yang mengidentifikasi kelompok sampel berdasarkan karakteristik serupa. Analisis klaster mengelompokkan elemen mirip sebagai obyek penelitian yang mempunyai tingkat homegenitas yang tinggi antar obyek menjadi klaster yang berbeda dengan tingkat heterogenitas obyek yang tinggi antar klaster. Pengklasteran ini didasarkan pada gugus variabel yang dipertimbangkan untuk diteliti.

Hasil pengklasteran diharapkan akan menyediakan beberapa pengertian yang mendalam untuk masing-masing provinsi di Indonesia. Pengklasteran mempunyai efek mengurangi banyak data dengan mengurangi banyaknya obyek. Analisis klaster harus memenuhi asumsi berikut:

a. Sampel yang diambil harus benar-benar bisa mewakili populasi.
b. Multikolinieritas yaitu korelasi antar obyek. Sebaiknya tidak ada, bila ada maka besar multikolinieritas tidaklah tinggi (< 0,5).

Sebelum melakukan analisis klaster, data yang digunakan juga perlu diperhatikan. Apakah terdapat perbedaan nilai yang besar antar variabel. Misalnya, ada yang dalam satuan juta dan ada yang satuan puluhan atau bahkan lebih kecil. Perbedaan data yang besar akan menyebabkan perhitungan jarak menjadi tidak valid sehingga data harus ditransformasi. Transformasi dapat dilakukan dengan uji *z-score*.

Tujuan analisis klaster adalah mereduksi jumlah obyek dengan mengklasifikasikan obyek (kasus atau elemen) ke dalam klaster yang relatif homogen. Obyek-obyek di dalam satu klaster lebih mirip dibandingkan antar obyek pada klaster lain.

Untuk menentukan kedua obyek dikatakan mirip, perlu didefinisikan ukuran kemiripan antar dua obyek. Hal ini dilakukan untuk memperoleh matrik *proximity* yaitu matrik persegi dan simetri dengan jumlah obyek yang sama pada baris dan kolom. Matrik ini menunjukkan kemiripan atau ketakmiripan antar obyek.

Adapun metode yang dapat digunakan untuk mengukur kesamaan antar obyek yaitu:

1. Mengukur jarak antar dua obyek. Metode ini berbentuk matrik simetri *n x n* yang berisi kemiripan atau ketakmiripan antar obyek sehingga jarak antar dua obyek bisa langsung diukur.
2. Mengukur korelasi antar sepasang obyek pada beberapa variabel. Pada metode ini datanya berbentuk matrik. Kesamaan antar obyek didapat dengan transformasi satu-satu sehingga indeks ketakmiripan bisa dikonversi menjadi indeks kemiripan. Salah satu yang jelas bisa

menjadi ukuran ketakmiripan adalah fungsi jarak antara obyek *a* dan *b* di tulis $d(a,b)$. Fungsi ini harus memenuhi:

$$d(a,b) \geq 0$$

$$d(a,a) = 0$$

$$d(a,b) = d(b,a)$$

(a,b) *meningkat seiring tidak miripnya a dan b*

$$d(a,c) \leq d(a,b) + d(b,c)$$

Pengukuran jarak ada bermacam-macam namun yang paling sering digunakan adalah jarak *euclid* yaitu jarak geometris di ruang multidimensional. Beberapa cara dalam mengukur jarak yaitu

a. Menggunakan jarak *euclid* yaitu akar jumlah kuadrat perbedaan nilai untuk tiap variabel (Wichern, 2002).
   Jika $\underline{x}' = (x_1, x_2, \ldots, x_p)$ dan $\underline{y}' = (y_1, y_2, \ldots, y_p)$ maka

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

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

b. Mengunakan jarak kuadrat *euclid*
c. *The City Block or Manhattan Distance* antara dua obyek merupakan jumlah nilai perbedaan mutlak untuk tiap variabel. Jarak ini juga disebut jarak Minkowski.
   Jika $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$; *p* adalah variabel

   maka $\mathbf{x_i}' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ adalah kumpulan variabel pada obyek ke *i*

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - y_{jk} \right|^r \right]^{\frac{1}{r}}$$

d_{ij} adalah jarak antara obyek ke *i* dan obyek ke *j*.

d. The Chebyshev Distance antar dua obyek ialah nilai perbedaan mutlak yang maksimum pada tiap variabel.

Penggunaan jarak yang berbeda mungkin menghasilkan pengklasteran yang berbeda. Oleh karena itu dianjurkan untuk menggunakan jarak lain kemudian membandingkan hasilnya. Hal ini juga merupakan salah satu cara untuk mengakses kehandalan dan kesahihan analisis klaster.

3. Mengukur asosiasi antar obyek

Pada metode ini, data berbentuk non matrik (nominal atau ordinal). Jika akan dibandingkan obyek r dan obyek s maka indeks kesamaan ditunjukkan pada tabel dibawah ini.

.

Kesamaan Antara Dua Obyek

	obyek s			
		1	0	Jml
Obyek r	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	Jml	*a+c*	*b+d*	*a+b+c+d=p*

Asumsikan 1 jika karakteristik ada dan 0 jika karakteristik tidak ada. Koefisien kesamaan didefinisikan pada tabel berikut.

Koefisien Kesamaan Antara Dua Obyek

Nama Koefisien	Persamaan
Jaccard	$S_{rs} = \frac{a}{(a+b+c)}$
Dice	$S_{rs} = \frac{2a}{(2a+b+c)}$
Ochiai	$S_{rs} = \frac{a}{[(a+b)(a+c)]^{\frac{1}{2}}}$
Russell-Rao	$S_{rs} = \frac{a}{p}$
Yule	$S_{rs} = \frac{(ad-bc)}{ad+bc}$
Phi	$S_{rs} = \frac{(adbc)}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$

Konversi ketaksamaan untuk analisa $d_{rs} = 1 - S_{rs}$

d_{rs} = Koefisien ketaksamaan antara obyek r dan obyek s
S_{rs} = Koefisien kesamaan antara obyek r dan obyek s

Metode Pengklasteran

Ada dua metode pengklasteran yaitu metode hirarki dan metode non hirarki (pengklasteran *K-means)*. Pengklasteran yang ideal adalah pengklasteran yang tiap obyek hanya masuk atau menjadi anggota dari salah satu klaster sehingga tidak terjadi tumpang tindih atau *overlapping*. Semua metode pada dasarnya menggunakan kesamaan atau ketaksamaan antar obyek.

Metode Hirarki

Kebanyakan metode analisis klaster adalah hirarki yaitu resultan pengklasteran yang mempunyai suatu peningkatan jumlah kelas tersarang. Metode ini lebih populer dibandingkan metode pengklasteran *K-means* namun metode ini membutuhkan waktu lama jika datanya besar. Masing-masing metode pada dasarnya mempunyai kekuatan tersendiri tergantung masalahnya.

Metode hirarki bisa aglomeratif (*agglomeration)* atau devisif *(devisife)*. Pengklasteran aglomeratif dimulai dengan setiap obyek berada dalam *n* klaster yang berbeda. Klaster dibentuk dengan mengabungkan dua klaster yang paling dekat lalu menentukan kembali kedekatan antar $n-1$ klaster yang baru. Dua klaster terdekat di gabung lagi, begitu seterusnya sampai didapat satu klaster yang memuat seluruh obyek.

Metode Aglomeratif terdiri dari :

1. Linkage Method (Metode *Linkage*)

Metode ini dibagi lagi menjadi tiga metode yaitu (Wichern, 2002):

i. *Single linkage*
Jika jarak antara klaster B_r dan B_s adalah $h(B_r,B_s)$ didefinisikan sebagai berikut:

$$h\left(B_r, B_S\right) = \min\left\{ d\left(x_i, x_j\right); x_i \text{ anggota } B_r \text{ dan } x_j \text{ anggota } B_S \right\}$$

Klaster B_r dan B_s akan digabung jika $h(B_r,B_s)$ adalah jarak yang terkecil sehingga metode ini juga disebut aturan tetangga dekat.

ii. *Complete linkage*
Metode ini hampir sama dengan *single linkage* hanya saja pada metode ini menggunakan jarak yang paling jauh antara dua klaster yang berbeda B_r dan B_s yang didefinisikan sebagai berikut:

$$h\left(B_r, B_S\right) = \max\left\{ d\left(x_i, x_j\right); x_i \text{ anggota } B_r \text{ dan } x_j \text{ anggota } B_S \right\}$$

iii. *Average linkage*
Metode ini menggunakan rata-rata jarak antara semua pasangan obyek sebagai jarak antara dua klaster. Metode ini jarang digunakan jika dibandingkan *single linkage* dan *complete linkage* karena metode ini membutuhkan informasi pada semua pasangan jarak. Namun ke tiga metode *linkage* ini seringkali memberikan hasil yang hampir sama.

Jarak antara dua klaster B_r dan B_s didefinisikan sebagai berikut:

$$h\left(B_r, B_s\right) = \frac{1}{n_1 n_2} \sum_{x_i \in B_r} \sum_{x_j \in B_s} d\left(x_i, x_j\right)$$

2. *Ward method* (Metode Ward)

Ward (1963) mengusulkan penggunaan metode yang didasarkan pada hasil informasi yang minimum dari kenaikan pada jumlah kuadrat deviasi rataan klaster. Proses berhenti pada kenaikan yang menyebabkan *error sum of squares* (ESS) dari gabungan tiap klaster yang mungkin. Nilai ESS digunakan sebagai fungsi obyektif dan didefinisikan sebagai berikut:

$$ESS = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij} \right)^2 \right)$$

x_{ij} : Nilai obyek ke *i* pada klaster ke *j*
k : Jumlah klaster tiap *stage*
n_j : Jumlah obyek ke *i* pada klaster ke *j*

Metode ini juga dikenal dengan metode varian minimum dan harus menggunakan jarak kuadrat *euclid* namun sulit untuk menggunakannya tanpa bantuan komputer.

3. *Centroid Method* (Metode *Centroid*)

Jarak antara dua klaster didefinisikan sebagai jarak *euclid* antar kedua rataan (*centroid*) klaster. Jika $\overline{x}_r$ dan $\overline{x}_s$ adalah vektor rataan (*centroid*) klaster B_r dan B_s, maka jarak antar dua klaster didefinisikan sebagai $h\left(B_r, B_s\right) = d\left(\overline{x}_r, \overline{x}_s\right)$. *Centroid* klaster baru yang terbentuk didapat dengan rumus $\frac{n_r \overline{x}_r + n_s \overline{x}_s}{n_r + n_s}$

n_r dan n_s adalah banyaknya anggota klaster B_r dan B_s.

4. *Median Method* (Metode *Median*)

Terkadang terdapat ukuran klaster B_s jauh lebih kecil dari pada klaster B_r yaitu $n_s << n_r$ bila kedua klater digabungkan maka *centroid* dari klaster baru tidak akan jauh berbeda dengan $\overline{x}_r$. Untuk menghindari kontribusi B_s terhadap pembentukan jarak yang baru tidak terlalu besar, Gower menyarankan penggunaan *median* antara klaster yang digabungkan sebagai titik untuk menghitung jarak yang baru.
Jika klaster B_r dan B_s di gabung maka akan di peroleh median baru yang didefinisikan sebagai berikut:

$$m_{baru} = \frac{m_r + m_s}{2}$$

Median ini dihitung sebagai titik tengah pada garis yang menghubungkan median lama dan median baru. Dengan demikan jarak antar klaster didefinisikan sebagai jarak antar median.

Metode berhirarki devisif merupakan kebalikan metode aglomeratif. Metode pengklasteran ini dimulai dengan menganggap bahwa hanya ada satu klaster yang memuat semua obyek. Langkah pertama yaitu membagi *n* obyek menjadi dua kelompok. Ini dapat dilakukan dalam $\left(2^{n-1} - 1 \right)$ cara. Proses dilanjutkan dengan cara yang tidak sama dengan aglomeratif.

Metode devisif lebih menguntungkan dibandingkan aglomeratif. Devisif tidak dilanjutkan ketika didapat *n* klaster yang mempunyai satu obyek dan bila terdapat jumlah variabel yang lebih sedikit dibandingkan obyek maka perhitungan yang dibutuhkan juga sedikit yaitu d^2 bila dibandingkan dengan perhitungan devisif yang didekati oleh $(n\text{-}1)^2$. Namun metode ini jarang digunakan dan tidak semua *software* menyediakan fasilitas metode ini.

Metode Non Hirarki (Pengklasteran *K-means*)

Berbeda dengan metode hirarki, metode ini justru dimulai dengan menentukan terlebih dahulu jumlah klaster yang diinginkan dan *centroid* di tiap klaster. Pada beberapa *software*, *centroid* yang digunakan adalah *k* pengamatan pertama namun ada juga *software* yang menentukan *centroid* secara acak. Kemudian hitung jarak antara tiap obyek dengan tiap *centroid*. Masukkan tiap obyek ke suatu klaster berdasarkan jarak terdekat dengan *centroid* klaster yang berpadanan. Hitung kembali tiap *centroid* yang terbentuk. Begitu seterusnya hingga tidak ada lagi pemindahan obyek antar klaster. Metode ini biasa disebut pengklasteran *K-means* . Pengklasteran *K-means* pertama kali dipopulerkan oleh Hartigan pada tahun 1975.

Penggunaan pengklasteran *K-means* untuk menjelaskan algoritma dalam penentuan suatu obyek ke dalam klaster tertentu berdasarkan rataan terdekat. Asumsikan *n* adalah obyek dan *p* adalah variabel yang dinotasikan dengan $x(i, j)$ *i=1,2….n* dan *j=1,2,..,p* dan dengan menggunakan jarak *euclid* antar obyek. Jika $p(n, k)$ adalah partisi yang merupakan hasil pada tiap obyek dialokasikan untuk salah satu dari klaster ke *1,2,..,k*. Rata-Rata variabel ke *j* pada klaster ke *l* dinotasikan dengan $\bar{x}(i, j)$, dan jumlah obyek pada klaster ke *l* dinotasikan dengan $n(l)$. Maka jarak antara obyek ke *i* dan klaster ke *l* didefinisikan sebagai berikut:

$$d(i,l) = \left(\sum_{j=1}^{p} \left[x(i,j) - \bar{x}(i,j) \right]^2 \right)^{1/2}$$

dengan

$$E[p(n,k)] = \sum_{i=1}^{n} D[i,l(i)]^2$$

adalah *error* partisi. *l(i)* adalah klaster yang memuat obyek ke *i, D[i,l(i)]* adalah jarak *euclid* antara obyek ke *i* dan rata-rata klaster yang memuat obyek.

Pengklasteran *K-means* sangat cocok untuk data dengan ukuran yang besar karena memiliki kecepatan yang lebih tinggi dibandingkan metode hirarki. Namun, pemilihan banyaknya klaster dan *centroid* yang harus ditentukan lebih dahulu menjadi kelemahan metode ini. Hasil pengklasteran mungkin tergantung pada urutan observasi data.

Penggunaan metode hirarki dan non hirarki (pengklasteran *K-means*) secara berdampingan. Suatu pemecahan pengklasteran awal diperoleh dengan metode hirarki misalnya *average linkage* atau *ward method*. Banyaknya klaster dan *centroid* yang diperoleh digunakan sebagai *input* untuk pengklasteran *K-means*. *Output* metode hirarki adalah semacam ringkasan yang digambarkan oleh dendogram. Dendogram merupakan diagram seperti pohon dua dimensi yang mengilustrasikan pemisahan atau penggabungan dengan tingkat yang berjenjang.

Pada metode hirarki, dendogram dapat membantu peneliti untuk menentukan jumlah klaster yang ideal. Posisi garis pada skala menunjukkan jarak. Jarak pada tahapan awal mempunyai nilai yang hampir sama sehingga sukar untuk menentukan urutan beberapa klaster awal dibentuk. Namun, jelas sekali bahwa dua tahap terakhir mempunyai jarak yang besar untuk digabung. Sebenarnya tidak ada aturan baku dalam menentukan banyak klaster tergantung subjektifitas peneliti. Peneliti juga dapat menggunakan pertimbangan teoritis, konseptual dan praktis.

Interpretasi Profil dan Akses Validitas Klaster

Interpretasi profil klaster meliputi pengkajian mengenai *centroid* yaitu rata-rata nilai obyek yang terdapat dalam klaster pada tiap variabel. Nilai *centroid* memungkinkan untuk menguraikan setiap klaster dengan cara memberi suatu label atau nama. Label suatu klaster juga dapat didasarkan pada manfaat yang akan di cari.

Pengecekan mutu hasil pengklasteran dapat dilakukan dengan analisis klaster yang menggunakan ukuran jarak yang berbeda, menggunakan metode pengklasteran yang berbeda dan membandingkan hasilnya.

Model Persamaan Struktural

Structural Equation Modeling (SEM) adalah sekumpulan alat atau teknik-teknik statistika yang memungkinkan tidak hanya mendapatkan model hubungan namun juga pengujian sebuah rangkaian hubungan yang relatif rumit secara simultan.

Hubungan yang rumit itu dapat dibangun antara satu atau beberapa variabel dependen dengan satu atau beberapa variabel independen. Masing-masing variabel dependen dan independen dapat berbentuk faktor (atau konstruk, yang dibangun dari beberapa variabel indikator). Variabel-variabel itu dapat berbentuk sebuah variabel tunggal yang diobservasi atau yang diukur langsung dalam sebuah proses penelitian.

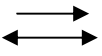
Pemodelan penelitian melalui SEM memungkinkan seorang peneliti dapat menjawab pertanyaan penelitian yang bersifat regresif maupun dimensional (yaitu mengukur apa dimensi-dimensi dari sebuah konsep). Pada saat seorang peneliti menghadapi pertanyaan penelitian berupa identifikasi dimensi-dimensi sebuah konsep atau konstruk (seperti yang lazim dilakukan dalam analisis faktor) dan pada saat yang sama peneliti ingin mengukur pengaruh atau derajat hubungan antar faktor yang telah diidentifikasikan dimensi-dimensinya itu, SEM akan merupakan alternatif jawaban yang layak dipertimbangkan. Itulah sebabnya dapat dikatakan bahwa pada dasarnya SEM adalah kombinasi antara analisis faktor dan analisis regresi berganda.

Konvensi SEM

Beberapa konvensi yang berlaku dalam diagram SEM adalah sebagai berikut:

Variabel terukur (*Measured Variable)*. Variabel ini disebut juga *observed variable*, *indicator variable* atau *manifest variable*, digambarkan dalam bentuk segi empat. Variabel terukur adalah variabel yang datanya harus dicari melalui penelitian lapangan.

Faktor. Faktor adalah sebuah variabel bentukan, yang dibentuk melalui indikator-indikator yang diamati dalam dunia nyata. Karena merupakan variabel bentukan, maka disebut *latent variable*, *construct*, atau *unobserved variable*. Faktor atau konstruk atau variabel laten ini digambarkan dalam bentuk diagram lingkar atau oval atau ellips.



Hubungan antar variabel. Hubungan antar variabel dinyatakan melalui garis. Karena itu bila tidak ada garis berarti tidak ada hubungan langsung yang dihipotesakan.

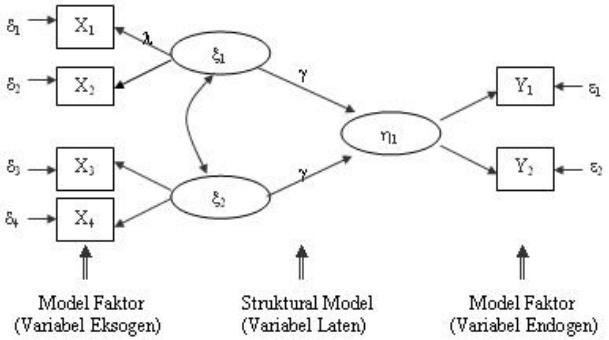
Bentuk-bentuk hubungan antar variabel adalah sebagai berikut:

1. **Garis dengan anak panah satu arah (→)**. Garis ini menunjukkan adanya hubungan yang dihipotesakan antara dua variabel, dimana variabel yang dituju oleh anak panah merupakan variabel dependen. Dalam SEM terdapat dua hipotesis dengan anak panah satu arah, yaitu:

a. Hipotesa mengenai dimensi faktor
Dimensi-dimensi sebuah faktor akan terlihat dalam diagram SEM melalui arah anak panah (→) yang digunakan. Masing-masing indikator sebagai variabel dependen secara bersama-sama dihipotesakan sebagai dimensi dari sebuah konsep atau faktor.

b. Hipotesa mengenai hubungan regresi
Hipotesa mengenai pengaruh satu atau beberapa variabel independen terhadap satu atau beberapa variabel dependen dinyatakan pula dalam anak panah satu arah (→).

2. **Garis dengan anak panah dua arah (↔)**. Anak panah dua arah ini digunakan untuk menggambarkan kovarian atau korelasi antara dua variabel.



Notasi yang digunakan dalam SEM

Gambar di atas menunjukkan bahwa model persamaan struktural merupakan pendekatan terintregasi antara Analisis Faktor Konfirmatori, Model Struktural, dan Analisis Jalur.

Notasi SEM

Notasi-notasi yang dipakai dalam SEM adalah sebagai berikut:
X = indikator variabel eksogen
Y = indikator variabel endogen
ξ = Ksi, variabel laten X atau variabel laten eksogen

η = Eta, variabel laten Y atau variabel laten endogen
δ = Delta, galat pengukuran pada variabel manifes untuk variabel laten X
ε = Epsilon, galat pengukuran pada variabel manifes untuk variabel laten
γ = Gamma, koefisien pengaruh variabel eksogen terhadap variabel endogen
β = Beta, koefisien pengaruh variabel endogen terhadap variabel endogen
λ = Lambda, loading faktor

Dalam SEM dapat dilakukan analisis hubungan antar peubah laten dengan peubah indikator dengan metode Analisis Faktor Konfirmatori, sekaligus mendapatkan model yang bermanfaat untuk prediksi atau prakiraan, dilakukan dengan Model Struktural.

Langkah-langkah SEM

Sebuah pemodelan SEM yang lengkap pada dasarnya terdiri dari *Measurement Model* dan *Structural Model*. Model Pengukuran ditujukan untuk mengkorfimasi sebuah dimensi atau faktor berdasarkan indikator-indikator empirisnya. *Structural Model* adalah model mengenai struktur hubungan yang membentuk atau menjelaskan kausalitas antar faktor.

Untuk membuat pemodelan yang lengkap beberapa langkah berikut ini perlu dilakukan.

1. Pengembangan model berbasis teori.

Langkah pertama dalam pengembangan model SEM adalah pencarian atau pengembangan sebuah model yang mempunyai justifikasi teoritis yang kuat. Tanpa dasar teoritis yang kuat, SEM tidak dapat digunakan, karena SEM tidak digunakan untuk menghasilkan sebuah model, tetapi digunakan untuk mengkonfirmasi model teoritis melalui data empirik. SEM bukanlah untuk menghasilkan kausalitas, tetapi untuk membenarkan adanya kausalitas teoritis melalui uji data empirik. Itulah sebabnya uji hipotesis mengenai perbedaan dengan menggunakan uji chi-square digunakan dalam SEM, dengan

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(\frac{\left(s_{ij}-\sigma_{ij}\right)^2}{\sigma_{ij}}\right)$$

$$db = \tfrac{1}{2}\left[\left(p+q\right)\left(p+q-1\right)\right]-t$$

dimana :

s_{ij} : kovarian sampel baris ke-i dan kolom ke-j

σ_{ij} : kovarian estimasi populasi baris ke-i dan kolom ke-j
r : banyaknya baris
k : banyaknya kolom
db : derajat kebebasan
p : jumlah variabel indikator endogen
q : jumlah variabel indikator eksogen
t : jumlah koefisien parameter yang diestimasi dalam model

Tujuan dari analisis ini adalah menguji apakah sebuah model yang telah mempunyai konsep teori sesuai dengan data empirik yang didapat.

Hipotesis yang digunakan sebagai berikut:

Ho : Tidak ada perbedaan antara matriks kovarian populasi sebuah faktor yang diestimasi dari beberapa variabel dengan matriks kovarian sampelnya.

H_1 : Ada perbedaan antara matriks kovarian populasi sebuah faktor yang diestimasi dari beberapa variabel dengan matriks kovarian sampelnya.

Dalam hal ini yang diharapkan gagal tolak H_0, artinya perbedaan matriks kovarian sampel dan matriks kovarian populasi terestimasi harus kecil dan tidak signifikan sehingga hipotesa nol tidak dapat ditolak. Nilai χ^2 diharapkan sekecil mungkin atau p-value > α (0.05).

2. Pengembangan diagram alur untuk menunjukkan hubungan kausalitas

Setelah didapat model hipotetik, model tersebut kemudian digambarkan dalam sebuah diagram alur (diagram *path*). Diagram alur sangat bermanfaat untuk menunjukkan alur hubungan kausal antar variabel eksogen dan endogen, dimana hubungan kausal yang telah ada justifikasi teori dan konsepnya divisualisasikan kedalam gambar sehingga lebih mudah melihatnya dan lebih menarik. Variabel eksogen, yang dikenal juga sebagai *source variable* atau *independent variable*, tidak diprediksi oleh variabel yang lain dalam model. Secara diagramatis variabel eksogen adalah variabel yang dituju oleh garis dengan satu ujung panah. Sedangkan variabel endogen adalah variabel yang diprediksi oleh satu atau beberap variabel.

3. Konversi diagram alur kedalam serangkaian persamaan struktural dan spesifikasi model pengukuran

Setelah model dikembangkan dan digambarkan dalam sebuah diagram alur, selanjutnya adalah konversi diagram alur kedalam rangkaian persamaan. Persamaan yang akan dibangun akan terdiri dari:

a. Persamaan-persamaan struktural (*structural equation*). Persamaan ini dirumuskan untuk menyatakan hubungan kausalitas antar berbagai konstruk. Persamaan struktural pada umumnya dibangun dengan pedoman:

Variabel Endogen = Variabel Eksogen + Variabel Endogen + Error

b. Persamaan spesifikasi model pengukuran (*measurement model*). Pada spesifikasi ini peneliti mementukan variabel mana mengukur konstruk mana, serta menentukan serangkaian matriks yang menunjukkan korelasi yang dihipotesakan.

4. Pemilihan matriks input dan teknik estimasi atas model yang dibangun.

SEM hanya menggunakan matriks varian/kovarian atau matriks korelasi sebagai data input untuk keseluruhan estimasi yang dilakukannya. Observasi individual digunakan, tetapi input tersebut akan dikonversi kedalam bentuk matrik kovarian atau matrik korelasi sebelum dilakukan estimasi. Matriks kovarian digunakan bila tujuan dari analisis adalah pengujian sebuah model yang telah mempunyai konsep teori. Sedangkan matriks korelasi digunakan bilamana tujuan analisis ingin mendapatkan penjelasan mengenai pola hubungan kausal antar variabel.

Beberapa dasar yang dapat digunakan untuk memilih teknik estimasi yang akan digunakan dapat mengacu pada tabel berikut:

Pertimbangan	Teknik yang dapat dipilih	Keterangan
Bila ukuran sample adalah kecil (100-200) dan asumsi normalitas dipenuhi	Maximum likelihood estimation (ML)	ULS & SLS biasanya tidak menghasilkan uji χ^2
Bila asumsi normalitas dipenuhi dan ukuran sample antara 200-500	ML dan Generalized Least Square estimation (GLS)	Bila ukuran sampel kurang dari 500, GLS cukup baik
Bila asumsi normalitas kurang dipenuhi dan ukuran sampel lebih dari 2500	Asymptotically Distribution-Free estimation (ADF)	ADF kurang cocok bila ukuran sampel kurang dari 2500

Menilai problem identifikasi

Problem identifikasi pada prinsipnya adalah problem mengenai ketidakmampuan dari model yang dikembangkan untuk menghasilkan estimasi yang unik. Problem ini muncul berkenaan dengan pengembangan model, sehingga bilamana setiap penduga parameter muncul problem identifikasi maka harus mendapatkan pertimbangan ulang. Teori dan konsep yang menjadi rujukan pengembangan

model hipotetik harus diteliti ulang sehingga masih dimungkinkan untuk penyempurnaan model, misalkan dengan memperbanyak variabel konstruk.
Problem identifikasi dapat muncul melalui gejala-gejala berikut:

a. Standar error untuk satu atau beberapa koefisien adalah sangat besar
b. Program tidak mampu menghasilkan matrik informasi yang seharusnya disajikan
c. Muncul angka-angka yang aneh, seperti adanya varian error negatif
d. Munculnya korelasi yang sangat tinggi antar koefisien estimasi yang didapat (misalnya lebih dari 0.9)
e. Evaluasi model

Kesesuaian model dievaluasi melalui telaah terhadap berbagai kriteria goodness-of-fit. Untuk itu, yang pertama dilakukan adalah mengevaluasi apakah data yang digunakan dapat memenuhi asumsi-asumsi SEM. Bila asumsi terpenuhi, maka model dapat diuji melalui berbagai cara yang akan diuraikan berikutnya.

Asumsi-asumsi SEM

a. Ukuran Sampel

Ukuran sampel yang harus dipenuhi dalam pemodelan ini adalah minimum berjumlah 100 dan selanjutnya menggunakan perbandingan 5 observasi untuk setiap *estimated parameter*.

b. Normalitas dan Linieritas

Normalitas dapat diuji dengan melihat gambar histogram data atau dapat diuji dengan metode-metode statistik. Uji linieritas dapat dilakukan dengan mengamati scatterplots dari data, yaitu dengan memilih pasangan data dan dilihat pola penyebarannya untuk menduga ada tidaknya linieritas. Sebaran data yang dianalisis harus memenuhi asumsi sebaran Normal, dan hubungan antar *estimated parameter* bersifat linier.

c. Outliers

Outliers adalah observasi yang muncul dengan nilai-nilai ekstrim baik secara univariat maupun multivariat, yaitu yang muncul karena kombinasi karakteristik unik yang dimilikinya dan terlihat sangat jauh berbeda dari observasi-observasi lainnya, dan ini bisa mengganggu pada saat analisis data.

d. Multikolinieritas dan Singularitas

Variabel yang saling berhubungan akan menyebabkan hasil yang bias. Sebaiknya data tidak ada multikolinieritas dan singularitas. Bila hal ini ada sebaiknya data dikeluarkan, atau alternatif lain adalah data tersebut dibuat '*composit variables*', dan variabel komposit ini dapat dianalisis lebih lanjut. Multikolinearitas dapat dideteksi dari determinan matriks kovarian. Nilai determinan matriks kovarian yang sangat kecil (*extremely small*) memberi indikasi adanya problem multikolinieritas.

Uji Kesesuaian dan Uji Statistik

Dalam analisis SEM tidak ada alat uji statistik tunggal untuk mengukur atau menguji hipotesis mengenai model. Evaluasi goodness-of-fit yang dimaksud adalah untuk mengukur kebenaran model yang diajukan. Berikut ini adalah beberapa indeks kesesuaian dan *cut off value*-nya yang digunakan untuk menguji apakah sebuah model dapat diterima atau ditolak.

a. ***Chi-Square Statistics***
Alat uji paling fundamental untuk mengukur *overall fit* adalah chi-square. Chi-square ini bersifat sangat sensitif tehadap besarnya sampel yang digunakan karena itu bila jumlah sampel cukup besar, yaitu lebih dari 2000 sampel, Chi-square harus didampingi oleh alat uji lain. Maka sampel yang disarankan antara range 100 sampai 200 sampel. Semakin kecil nilai χ^2 maka makin kecil kebenaran model tersebut.

b. ***RMSEA (The Root Mean Square Error of Approximation)***
RMSEA adalah sebuah indeks yang dapat digunakan untuk mengkompensasi chi-square statistik dalam sampel besar. Nilai RMSEA lebih kecil atau sama dengan 0.08 merupakan indeks untuk diterimanya model yang menunjukkan sebuah *close fit* dari sebuah model berdasarkan derajat bebas.

c. ***GFI (Goodness of Fit Index)***
Indeks kesesuaian ini akan menghitung proporsi tertimbang dari varian dalam matriks kovarian sampel yang dijelaskan oleh matriks kovarian populasi yang terestimasikan. GFI adalah sebuah ukuran non-statistikal yang mempunyai rentang nilai antara 0 (*poor fit*) sampai dengan 1 (*perfect fit*). Nilai yang tinggi dalam indeks ini menunjukkan sebuah *better fit*.

d. ***AGFI (Adjusted Goodness of Fit Index)***
Tingkat penerimaan yang direkomendasikan adalah bila AGFI mempunyai nilai sama dengan atau lebih besar dari 0.9.

e. ***CMIN/DF***
The Minimum Sample Discrepancy Function (CMIN) dibagi dengan derajat bebas menghasilkan indeks CMIN/DF, yaitu salah satu indikator untuk mengukur tingkat kesesuaian sebuah model. Nilai χ^2 relatif yang diharapkan adalah kurang dari atau sama dengan 2.00. CMIN/DF

f. ***TLI (Tucker Lewis Index)***
TLI adalah sebuah alternatif incremental fit index yang membandingkan sebuah model yang diuji terhadap sebuah baseline model. Nilai yang direkomendasikan sebagai acuan untuk diterimanya sebuah model adalah penerimaan ≥ 0.95.

g. ***CFI (Comperative Fit Index)***
Nilai CFI yang direkomendasikan adalah ≥ 0.95. Semakin mendekati 1, maka model semakin baik. Keunggulan dari indeks ini adalah bahwa besaran ini besarnya tidak dipengaruhi oleh ukuran sampel.

Indeks-indeks yang dapat digunakan untuk menguji kelayakan sebuah model, seperti yang tertera diatas, dapat diringkas dalam tabel berikut:

GOODNESS OF FIT INDEX	CUT-OFF VALUE
χ^2 (Chi-square)	Diharapkan kecil
Significance probability	≥ 0.05
RMSEA	≤ 0.08
GFI	≥ 0.90
AGFI	≥ 0.90
CMIN/DF	≤ 2.00
TLI	≥ 0.95
CFI	≥ 0.95

5. Interpretasi dan Modifikasi model

Bilamana model yang diperoleh cukup baik, maka selanjutnya adalah melakukan interpretasi. Bila model belum baik maka perlu diadakan modifikasi. Modifikasi model hanya dapat dilakukan bila ada justifikasi teoritis atau konsep yang cukup kuat, sebab metode SEM bukan ditujukan untuk menghasilkan teori atau model, tetapi menguji model. Kelemahannya terutama adalah sangat sulitnya diperoleh model yang cocok dengan data (*fitting model*) karena kompleksnya hubungan.

Model Pengukuran (*Measurement Model*)

Model pengukuran (*Measurement Model*) adalah proses pemodelan untuk menyelidiki unidimensionalitas dari indikator-indikator yang menjelaskan sebuah faktor. Pengukuran ini bertujuan untuk mengkonfirmasi faktor-faktor yang telah terbentuk, sehingga analisis ini disebut *Confirmatory Factor Analysis.* Keakuratan model pengukuran adalah dengan pemeriksaan mengenai validitas dan reliabilitas. Hasil yang signifikan dari λ menunjukkan data valid, dan δ maupun ε yang tidak signifikan menunjukkan hasil yang reliabel.

Model Faktor

Model Satu Faktor

Berdasarkan pertimbangan persamaan yang menggunakan p-indikator ditunjukkan oleh model yang terdiri dari satu faktor, yaitu:

$$x_1 = \lambda_1 \xi + \delta_1$$
$$x_2 = \lambda_2 \xi + \delta_2$$
$$\vdots$$
$$x_p = \lambda_p \xi + \delta_p$$

dengan : $x_1, x_2, \ldots, x_p$ adalah indikator dari *common factor* (ξ),

$\lambda_1, \lambda_2, \ldots, \lambda_p$ adalah loading dari *pattern*/ model,

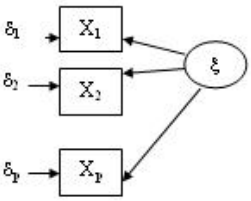
$\delta_1, \delta_2, \ldots, \delta_p$ adalah faktor tunggal (*unique factor*) untuk tiap persamaan error term.

Keterangan:

Indikator adalah variabel yang dapat diamati atau diukur, dalam kasus yang dipilih disebut sebagai variable terukur.

Common factor adalah suatu faktor yang secara bersama-sama dibentuk oleh beberapa indikator, pada kasus yang dipilih, gaya kepemimpinan, motivasi, dan kepuasan kerja adalah bentuk dari *common factor*.

Unique factor, disini dimaksudkan sebagai error, error (galat) adalah suatu observasi yang direkam dengan salah, mungkin karena awalnya direkam secara tak benar atau karena dikopi / diketik secara tak benar pada beberapa tahap selanjutnya.



Gambar Model satu faktor

Bila diasumsikan bahwa p=2, artinya berdasarkan gambar 2 model satu faktor dengan 2 indikator, maka:

$x_1 = \lambda_1 \xi + \delta_1$; $x_2 = \lambda_2 \xi + \delta_2$

Matriks kovarian, Σ antar variabel adalah: $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$

Secara umum diasumsikan:
Mean dari indikator (x), *common factor* (ξ) dan *unique factor* (*error*) adalah nol.

Varian dari indikator dan *common factor* adalah 1, yaitu indikator dan *common factor* yang distandarisasi.

Unique factor (*error*) tidak berkorelasi antar dirinya sendirinya atau dengan *common factor*, yaitu : $E\left(\xi_i, \varepsilon_j\right) = 0$ dan $E\left(\delta_i, \delta_j\right) = 0$

Maka, varian dan kovarian dari indikator adalah:

$\sigma_1^2 = \lambda_1^2 + V\left(\delta_1\right)$; $\sigma_2^2 = \lambda_2^2 + V\left(\delta_2\right)$; $\sigma_{12} = \sigma_{21} = \lambda_1\lambda_2$

pada persamaan λ_1, λ_2, $V(\delta_1)$, dan $V(\delta_2)$, adalah parameter model, vektor θ berisi parameter model $\theta=[\lambda_1, \lambda_2, V(\delta_1), V(\delta_2)]$.

Dengan mensubtitusikan persamaan (6) pada persamaan (5), didapat:

$$\Sigma(\theta) = \begin{bmatrix} \lambda_1^2 + V\left(\delta_1\right) & \lambda_1\lambda_2 \\ \lambda_1\lambda_2 & \lambda_2^2 + V\left(\delta_2\right) \end{bmatrix}$$

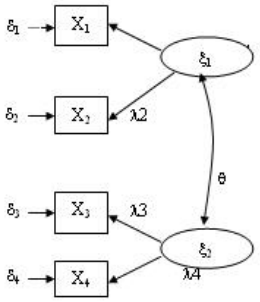
$\Sigma(\theta)$, matrik kovarian yang akan menghasilkan vektor parameter θ, dengan catatan masing-masing vektor parameter akan disimpulkan dalam matrik kovarian *unique*.

Model Dua Faktor

Berdasarkan pertimbangan model dua faktor, persamaan yang dihasilkan :

$$x_1 = \lambda_1\xi_1 + \delta_1 \qquad x_3 = \lambda_3\xi_2 + \delta_3$$
$$x_2 = \lambda_2\xi_1 + \delta_2 \qquad x_4 = \lambda_4\xi_2 + \delta_4$$

x_1 dan x_2 adalah indikator ξ_1, x_3 dan x_4 adalah indikator ξ_2.



Gambar Model dua faktor dengan variabel laten berkorelasi

Mengikuti persamaan hubungan antara model parameter maka elemen matrik kovarian sebagai berikut:

$$\sigma_1^2 = \lambda_1^2 + V(\delta_1) \qquad \sigma_2^2 = \lambda_2^2 + V(\delta_2)$$
$$\sigma_3^2 = \lambda_3^2 + V(\delta_3) \qquad \sigma_4^2 = \lambda_4^2 + V(\delta_4)$$
$$\sigma_{12} = \lambda_1\lambda_2 \qquad \sigma_{13} = \lambda_1\lambda_3\phi \qquad \sigma_{14} = \lambda_1\lambda_4\phi$$
$$\sigma_{23} = \lambda_2\lambda_3\phi \qquad \sigma_{24} = \lambda_2\lambda_4\phi \qquad \sigma_{34} = \lambda_3\lambda_4$$

dimana ϕ adalah kovarian antara dua *construct latent.*

Uji Kesesuaian Model

Untuk menunjukkan model secara keseluruhan layak atau tidak, maka dilakukan pengujian. Statistik uji yang digunakan adalah *chi-square* (χ^2). Perumusan hipotesisnya adalah

H_0 : $\boldsymbol{\Sigma} = \mathbf{S}$
H_1 : $\boldsymbol{\Sigma} \neq \mathbf{S}$

dimana: $\boldsymbol{\Sigma}$ adalah matriks kovarian populasi estimasi
S adalah matriks kovarian sampel

Bila p-value < α, artinya secara statistik model faktor tidak sesuai dengan data.

Analisis Lintas Path

Analisis lintas adalah sebuah metode analisis statistik yang digunakan untuk menentukan variabel mana yang berpengaruh dominan atau jalur mana yang berpengaruh lebih kuat dalam suatu model lintasan (model kausal) umumnya berupa diagram *path* yang diperoleh berdasarkan pertimbangan teoritis dan pengetahuan tertentu.

Model Struktural

Metode pendugaan parameter untuk model struktural dapat dilakukan dengan pendekatan model rekursif, pendekatan kuadrat terkecil tak langsung (*Indirect Least Square*/ ILS) dan pendekatan kuadrat dua tahap (*Two Stage Least Square*/ TSLS). Keakuratan model struktural bisa dilihat melalui koefisien determinasi total, disimbolkan dengan R^2, yang berkisar dari 0 sampai dengan 1. Model dikatakan baik bila koefisien makin besar (makin mendekati 1).

Analisis Korespondensi

Menurut Greenacre *dalam* Sudarsono dan Latra (2005) *Analyses des Correspondances* atau Analisis Korespondensi adalah teknik analisis data yang memperagakan baris dan kolom secara bersamaan dari suatu tabel kontingensi dua arah dalam ruang vektor berdimensi dua. Analisis Korespondensi merupakan bagian dari analisis peubah ganda, secara umum analisis tersebut dibagi menjadi dua bagian yaitu (1) generalisasi analisis peubah tunggal dan (2) pereduksian dimensi data peubah ganda. Topik dalam pereduksian data peubah ganda anatara lain Analisis Komponen Utama (AKU), Biplot, Faktor, Gerombol (*cluster*), dan Analisis Korespondensi. Lebih spesifiknya Analisis Korespondensi dan Biplot merupakan metode yang menyajikan hasil analisis dalam bentuk grafik. Dalam Analisis Korespondensi ada beberapa asumsi yang harus dipenuhi, yaitu

1. Ukuran jarak Kai Kuadrat antar titik-titik (nilai kategori) analogi dengan konsep korelasi antar variabel, maksudnya hubungan keeratan dalam analisis ini dapat dilihat dari jarak Kai Kuadratnya. Semakin kecil jarak Kai Kuadrat antar dua variabel berarti semakin erat hubungan variabel tersebut.
2. Variabel kolom yang tepat di variabel kategori baris diasumsikan homogen, yaitu nilai pengamatan dari variabel pada lajur baris pada setiap kolom yang dijelaskan adalah sama.
3. Analisis Korespondensi hanya digunakan untuk dua atau tiga variabel.
4. Analisis Korespondensi adalah sebuah teknik nonparametrik yang tidak memerlukan pengujian asumsi seperti kenormalan, autokorelasi, multikolinearitas, heteroskedastisitas, linieritas sebelum melakukan analisis selanjutnya.
5. Dimensi yang terbentuk dalam Analisis Korespondensi disebabkan dari kontribusi titik-titik dari dimensi yang terbentuk dan penamaan dari dimensinya subjektif dari kebijakan, pendapat dan *error*.
6. Dalam Analisis Korespondensi variabel yang digunakan yaitu variabel diskrit yang mempunyai banyak kategori.

Tabel Kontingensi Dua Arah

Jika X dan Y adalah dua peubah yang masing-masing mempunyai sebanyak a dan b kategori, maka dapat dibentuk suatu matriks data pengamatan $\mathbf{P}$ dengan ukuran $a \times b$, Dengan $\mathrm{p}_{ij} \geq 0$ menyatakan frekuensi dari sel ke (i, j). Berikut contoh matriks data $\mathbf{P}$

$$\mathbf{P} = \begin{bmatrix} \mathrm{p}_{11} & \mathrm{p}_{12} & \cdots & \mathrm{p}_{1b} \\ \mathrm{p}_{21} & \mathrm{p}_{22} & \cdots & \mathrm{p}_{2b} \\ \cdots & \cdots & \ddots & \vdots \\ \mathrm{p}_{a1} & \mathrm{p}_{a2} & \cdots & \mathrm{p}_{ab} \end{bmatrix}$$

Dari matriks data $\mathbf{P}$ diatas dapat dibentuk tabel kontingensi dua arah sebagai berikut.

Tabel 1. Tabel Kontingensi Dua Arah

	Y_1	⋯	Y_j	⋯	Y_b	Total
X_1	p_{11}	⋯	p_{1j}	⋯	p_{1b}	$p_{1.}$
⋮		⋮		⋮		
X_i	p_{i1}	⋯	p_{ij}	⋯	p_{ib}	$p_{i.}$
⋮		⋮		⋮		
X_a	p_{a1}	⋯	p_{aj}	⋯	p_{ab}	$p_{a.}$
Total	$p_{.1}$	⋯	$p_{.j}$	⋯	$p_{.b}$	$p_{..}$

keterangan :

$p_{i.} = \sum_{j=1}^{b} p_{ij}$, $i = 1, 2, \ldots, a$ peluang marjinal X

$p_{.j} = \sum_{i=1}^{a} p_{ij}$, $j = 1, 2, \ldots, b$ peluang marjinal Y

$p_{..} = \sum_{i} \sum_{j} p_{ij}$, Jumlah total frekuensi dari matriks $\mathbf{P}$

p_{ij} adalah frekuensi pengamatan ke i baris pada j kolom

Dari tabel kontingensi dua arah diatas dapat dibentuk matriks korespondensi sebagai berikut:

$$\mathbf{K} = \begin{bmatrix} k_{11} & \ldots & k_{1j} & \ldots & k_{1b} \\ \vdots & \ddots & \vdots & \ldots & \vdots \\ k_{i1} & \ldots & k_{ij} & \ldots & k_{ib} \\ \vdots & \ldots & \vdots & \ddots & \vdots \\ k_{a1} & \ldots & k_{aj} & \ldots & k_{ab} \end{bmatrix}$$

dengan :

$$k_{ij} = \frac{p_{ij}}{p_{..}} \qquad \begin{matrix} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{matrix}$$

Bila setiap elemen pada suatu baris dijumlahkan maka diperoleh vektor dari jumlah baris matriks $\mathbf{K}$ yaitu $\mathbf{r}' = \mathbf{K}\ \mathbf{I} = \left(k_{1.}, \ldots, k_{a.}\right)'$ Sehingga didapat $\mathbf{D}_r$ =diag(**r**) adalah diagonal matriks baris yaitu:

$$\mathbf{D_r} = \text{diag}(\mathbf{r}) = \begin{bmatrix} k_{1.} & 0 & 0 & 0 \\ 0 & k_{2.} & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & k_{a.} \end{bmatrix}$$

Dengan cara yang sama, akan didapat jumlah setiap kolom dari matriksnya menjadi vektor jumlah kolom dari matriks $\mathbf{K}$ yaitu $\mathbf{c} = \mathbf{K}'\mathbf{I} = \left(k_{.1}, \ldots, k_{.b}\right)'$ sehingga didapat $\mathbf{D_c}$ = diag(**c**) adalah diagonal matriks kolom sebagai berikut:

$$\mathbf{D_c} = \text{diag}(\mathbf{c}) = \begin{bmatrix} k_{.1} & 0 & 0 & 0 \\ 0 & k_{.2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & k_{.b} \end{bmatrix}$$

Profil Baris dan Profil Kolom

Profil adalah proporsi dari setiap baris atau kolom Matriks Korespondensi yaitu setiap frekuensi pengamatan baris ke-i dan kolom ke-j dibagi dengan jumlah setiap total baris dan kolomnya masing-masing.

Matriks diagonal kolom dan baris diatas masing-masing berukuran $b \times b$ dan $a \times a$. Kemudian dapat dibentuk matriks $\mathbf{R}$ yang berukuran $a \times b$ sebagai berikut:

$$\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{K}$$

$$= \begin{bmatrix} \frac{k_{11}}{k_{1.}} & \frac{k_{12}}{k_{1.}} & \cdots & \frac{k_{1b}}{k_{1.}} \\ \frac{k_{21}}{k_{2.}} & \frac{k_{22}}{k_{2.}} & \cdots & \frac{k_{2b}}{k_{2.}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{k_{a1}}{k_{a.}} & \frac{k_{a2}}{k_{a.}} & \cdots & \frac{k_{ab}}{k_{a.}} \end{bmatrix}$$

Matriks $\mathbf{R}$ disebut profil baris (*row profile*) dalam ruang berdimensi b, dengan jumlah unsur-unsur profil dari baris adalah sama dengan satu. Selanjutnya didefinisikan profil baris ke-i sebagai r_i yaitu:

$$\mathbf{r}_i = \left[\frac{k_{i1}}{k_{i.}}, \frac{k_{i2}}{k_{i.}}, \ldots, \frac{k_{ib}}{k_{i.}} \right]$$

sedangkan matriks $\mathbf{C}$ berukuran $b \times a$ adalah:

$$\mathbf{C} = \mathbf{D}_c^{-1} \mathbf{K}'$$

$$= \begin{bmatrix} \frac{k_{11}}{k_{.1}} & \frac{k_{12}}{k_{.1}} & \cdots & \frac{k_{a1}}{k_{.1}} \\ \frac{k_{21}}{k_{.2}} & \frac{k_{22}}{k_{.2}} & \cdots & \frac{k_{a2}}{k_{.2}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{k_{1b}}{k_{.b}} & \frac{k_{2b}}{k_{.b}} & \cdots & \frac{k_{ab}}{k_{.b}} \end{bmatrix}$$

Matriks $\mathbf{C}$ disebut sebagai profil kolom (*column profile*) dalam ruang berdimensi a, dimana jumlah unsur-unsur dari profil kolom sama dengan satu. Sehingga profil kolom ke-j sebagai $\mathbf{c}'_j$ yaitu :

$$\mathbf{c}_j = \left[\frac{k_{1j}}{k_{.j}}, \frac{k_{2j}}{k_{.j}}, \ldots, \frac{k_{aj}}{k_{.j}} \right]$$

Untuk menampilkan profil-profil baris dan profil-profil kolom tersebut kedalam ruang dimensi *Euclid* yang berdimensi dua digunakan pendekatan jarak Kai Kuadrat, yaitu:

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(p_{ij} - m_{ij} \right)^2}{m_{ij}}$$

keterangan :

$m_{ij} = \left(\frac{p_{i.} p_{.j}}{p_{..}} \right)$ Taksiran nilai harapan

$k_{i.} = \frac{p_{i.}}{p_{..}}$ Jumlah setiap baris ke - i dari matriks korespondensi

$k_{.j} = \frac{p_{.j}}{p_{..}}$ Jumlah setiap kolom ke- j dari matriks korespondensi

Dengan menggunakan aljabar, diperoleh persamaan sebagai berikut :

$$\left(p_{ij} - m_{ij} \right)^2 = \left((p_{ij}^2) - 2 p_{ij} m_{ij} + (m_{ij}^2) \right)$$

$$= k_{ij}^2 p_{..}^2 - 2 k_{ij} p_{..} \left(\frac{p_{i.} p_{.j}}{p_{..}} \right) + \left(\frac{p_{i.} p_{.j}}{p_{..}} \right)^2$$

$$= k_{ij}^2 p_{..}^2 - 2 k_{ij} p_{..} \left(\frac{p_{i.} p_{.j}}{p_{..}} \right) + \left(\frac{p_{i.} p_{.j}}{p_{..}} \right)^2$$

$$= k_{ij}^{2} p_{..}^{2} - 2k_{ij} p_{..} \left(\frac{k_{i.} p_{..} k_{.j} p_{..}}{p_{..}} \right) + \left(\frac{k_{i.} p_{..} k_{.j} p_{..}}{p_{..}} \right) \left(\frac{k_{i.} p_{..} k_{.j} p_{..}}{p_{..}} \right)$$

$$= k_{ij}^{2} p_{..}^{2} - 2k_{ij} p_{..}^{2} k_{i.} k_{.j} + k_{i.}^{2} k_{.j}^{2} p_{..}^{2}$$

$$= p_{..}^{2} \left(k_{ij}^{2} - 2k_{ij} k_{i.} k_{.j} + k_{i.}^{2} k_{.j}^{2} \right)$$

$$= p_{..}^{2} \left(k_{ij} - k_{i.} k_{.j} \right)^{2}$$

Dengan mensubstitusikan kembali kedua persamaan terakhir, diperoleh :

$$\chi^{2} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{p_{..}^{2} \left(k_{ij} - k_{i.} k_{.j} \right)^{2}}{\frac{k_{i.} p_{..} k_{.j} p_{..}}{p_{..}}}$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{p_{..}^{2} \left(k_{ij} - k_{i.} k_{.j} \right)^{2}}{p_{..} \left(k_{i.} k_{.j} \right)}$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{b} p_{..} \frac{\left(k_{ij} - k_{i.} k_{.j} \right)^{2}}{k_{i.} k_{.j}}$$

$$= p_{..} \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(k_{ij} - k_{i.} k_{.j} \right)^{2}}{k_{i.} k_{.j}}$$

keterangan :

$m_{ij} = \left(\frac{p_{i.} p_{.j}}{p_{..}} \right)$ Taksiran nilai harapan

$p_{i.} = k_{i.} p_{..}$ Jumlah setiap baris ke-*i* dari matriks korespondensi

$p_{.j} = k_{.j} p_{..}$ Jumlah setiap kolom ke-*j* dari matriks korespondensi

Hasil

Misal diberikan suatu matriks korespondensi dengan $\mathbf{D_r}$ adalah matriks diagonal baris, $\mathbf{D_c}$ adalah matriks diagonal kolom, $\mathbf{r}$ merupakan vektor jumlah baris dan $\mathbf{c}$ adalah vektor jumlah kolom. Maka dapat dibentuk suatu matriks E sedemikian sehingga :

$$E = \mathbf{D_r^{-1}}(\mathbf{K} - \mathbf{r}\ \mathbf{c}')\mathbf{D_c^{-1}}(\mathbf{K} - \mathbf{r}\mathbf{c}')'$$

dengan $tr(E) = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(k_{ij} - k_{i.}k_{.j}\right)^2}{k_{i.}k_{.j}}$

keterangan :

$\mathbf{r}' = \left(k_{1.} \ldots k_{i.} \ldots k_{a.}\right)$ Vektor jumlah baris dari matriks $\mathbf{K}$

$\mathbf{c} = \left(k_{.1} \ldots k_{.j} \ldots k_{.b}\right)'$ Vektor jumlah kolom dari matriks $\mathbf{K}$

Hasil diatas diperoleh dari penurunan rumus sebagai berikut sebagai berikut :
Bila suatu matriks korespondensi dikurang dengan suatu vektor jumlah baris dan jumlah kolom akan didapatkan persamaan sebagai berikut:

$$\left(\mathbf{K} - \mathbf{r}\mathbf{c}'\right) = \begin{bmatrix} k_{11} & \cdots & k_{1j} & \cdots & k_{1b} \\ \vdots & & \vdots & & \vdots \\ k_{i1} & \cdots & k_{ij} & \cdots & k_{ib} \\ \vdots & & \vdots & & \vdots \\ k_{a1} & \cdots & k_{aj} & \cdots & k_{ab} \end{bmatrix} - \begin{bmatrix} k_{1.} \\ \vdots \\ k_{i.} \\ \vdots \\ k_{a.} \end{bmatrix} \begin{bmatrix} k_{.1} & \cdots & k_{.j} & \cdots & k_{.b} \end{bmatrix}$$

$$= \begin{bmatrix} k_{11} & \cdots & k_{1j} & \cdots & k_{1b} \\ \vdots & & \vdots & & \vdots \\ k_{i1} & \cdots & k_{ij} & \cdots & k_{ib} \\ \vdots & & \vdots & & \vdots \\ k_{a1} & \cdots & k_{aj} & \cdots & k_{ab} \end{bmatrix} - \begin{bmatrix} k_{1.}k_{.1} & \cdots & k_{1.}k_{.j} & \cdots & k_{1.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{i.}k_{.1} & \cdots & k_{i.}k_{.j} & \cdots & k_{i.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{a.}k_{.1} & \cdots & k_{a.}k_{.j} & \cdots & k_{a.}k_{.b} \end{bmatrix}$$

$$= \begin{bmatrix} k_{11} - k_{1.}k_{.1} & \cdots & k_{1j} - k_{1.}k_{.j} & \cdots & k_{1b} - k_{1.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{i1} - k_{i.}k_{.1} & \cdots & k_{ij} - k_{i.}k_{.j} & \cdots & k_{ib} - k_{i.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{a1} - k_{a.}k_{.1} & \cdots & k_{aj} - k_{a.}k_{.j} & \cdots & k_{ab} - k_{a.}k_{.b} \end{bmatrix}$$

Dari persamaan diatas akan didapatkan suatu matriks tranpose yaitu :

$$\left(\mathbf{K} - \mathbf{r}\mathbf{c}'\right)' = \begin{bmatrix} k_{11} - k_{1.}k_{.1} & \cdots & k_{i1} - k_{i.}k_{.1} & \cdots & k_{a1} - k_{a.}k_{.1} \\ \vdots & & \vdots & & \vdots \\ k_{1j} - k_{1.}k_{.j} & \cdots & k_{ij} - k_{i.}k_{.j} & \cdots & k_{aj} - k_{a.}k_{.j} \\ \vdots & & \vdots & & \vdots \\ k_{1b} - k_{1.}k_{.b} & \cdots & k_{ib} - k_{i.}k_{.b} & \cdots & k_{ab} - k_{a.}k_{.b} \end{bmatrix}$$

Perhatikan bahwa invers dar $\mathbf{D_r}$ adalah

$$\mathbf{D_r^{-1}} = \begin{bmatrix} \frac{1}{k_{1.}} & & \cdots & \cdots & 0 \\ \vdots & \ddots & & \cdots & 0 \\ 0 & \cdots & \frac{1}{k_{i.}} & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{1}{k_{a.}} \end{bmatrix}$$

sedangkan invers matriks $\mathbf{D_c}$ adalah :

$$\mathbf{D}_{\mathbf{c}}^{-1} = \begin{bmatrix} \frac{1}{k_{.1}} & & \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & 0 \\ 0 & \cdots & \frac{1}{k_{.j}} & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & \cdots & \cdots & \cdots & \frac{1}{k_{.b}} \end{bmatrix}$$

Dengan mengalikan diperoleh hasil :

$$\mathbf{D}_{\mathbf{r}}^{-1}\left(\mathbf{K} - \mathbf{r}\mathbf{c}'\right) = \begin{bmatrix} \frac{1}{k_{1.}} & & \cdots & \cdots & 0 \\ \vdots & \ddots & & \cdots & 0 \\ 0 & \cdots & \frac{1}{k_{i.}} & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{1}{k_{a.}} \end{bmatrix}_{a\times a} \begin{bmatrix} k_{11} - k_{1.}k_{.1} & \cdots & k_{1j} - k_{1.}k_{.j} & \cdots & k_{1b} - k_{1.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{i1} - k_{i.}k_{.1} & \cdots & k_{ij} - k_{i.}k_{.j} & \cdots & k_{ib} - k_{i.}k_{.b} \\ \vdots & & \vdots & & \vdots \\ k_{a1} - k_{a.}k_{.1} & \cdots & k_{aj} - k_{a.}k_{.j} & \cdots & k_{ab} - k_{a.}k_{.b} \end{bmatrix}_{a\times b}$$

Sehingga didapat rumusan secara umum dalam bentuk matriks sebagai berikut :

$$\mathbf{D}_{\mathbf{r}}^{-1}\left(\mathrm{K} - \mathbf{r}\mathbf{c}'\right) = \left[\frac{k_{ij} - k_{i.}k_{.j}}{k_{i.}}\right]_{a\times b}$$

Selanjutnya dapat diperoleh:

$$\mathbf{D_c^{-1}}\left(\mathbf{K}-\mathbf{rc}'\right)' = \begin{bmatrix} \frac{1}{k_{.1}} & & \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & 0 \\ 0 & \cdots & \frac{1}{k_{.j}} & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & \cdots & \cdots & \cdots & \frac{1}{k_{.b}} \end{bmatrix}_{b\times b}$$

$$\begin{bmatrix} k_{11}-k_{1.}k_{.1} & \cdots & k_{i1}-k_{i.}k_{.1} & \cdots & k_{a1}-k_{a.}k_{.1} \\ \vdots & & \vdots & & \vdots \\ k_{1j}-k_{1.}k_{.j} & \cdots & k_{ij}-k_{i.}k_{.j} & \cdots & k_{aj}-k_{a.}k_{.j} \\ \vdots & & \vdots & & \vdots \\ k_{1b}-k_{1.}k_{.b} & \cdots & k_{ib}-k_{i.}k_{.b} & \cdots & k_{ab}-k_{a.}k_{.b} \end{bmatrix}_{b\times a}$$

Sehingga didapat rumusan secara umum dalam bentuk matriks sebagai berikut :

$$\mathbf{D_c^{-1}}\left(\mathbf{K}-\mathbf{rc}'\right)' = \left[\frac{k_{ij}-k_{i.}k_{.j}}{k_{.j}}\right]_{b\times a}$$

dan didapatkan:

$$\mathbf{D_r^{-1}(K-r\ c')D_c^{-1}(K-rc')'} = \left[\sum_{j=1}^{b}\frac{\left(k_{ij}-k_{i.}k_{.j}\right)^2}{k_{i.}k_{.j}}\right]_{a\times a}$$

$$tr(E) = tr\left(\mathbf{D_r^{-1}(K-r\ c')D_c^{-1}(K-rc')'}\right)$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\frac{\left(k_{ij}-k_{i.}k_{.j}\right)^2}{k_{i.}k_{.j}}\right)_{a\times a}$$

Sehingga diperoleh persamaan sebagai berikut :

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(p_{ij} - \frac{p_{i.}p_{.j}}{p_{..}}\right)^2}{\frac{p_{i.}p_{.j}}{p_{..}}}$$

$$= p_{..} \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(k_{ij} - k_{i.}k_{.j}\right)^2}{k_{i.}k_{.j}}$$

$$= p_{..}\, tr(E) = p_{..} \sum_{i}^{m} \lambda_i^2$$

$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_d^2$ adalah nilai *inersia* atau akar ciri tak nol dari E dan $m = \text{rank}\left(E\right) = \text{rank}\left(\mathbf{K} - \mathbf{rc}'\right) = \text{rank}\left(\mathbf{P}\right) = (\min(a, b)-1)$ maka χ^2 dapat juga dituliskan sebagai berikut:

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} np_i \left(\frac{\frac{k_{ij}}{k_{i.}} - k_{.j}}{k_{.j}} \right)$$

$$= \sum_{i=1}^{a} nk_{i.} \left[\left(\mathbf{r_i} - \mathbf{c}\right) \mathbf{D_c^{-1}} \left(\mathbf{r_i} - \mathbf{c}\right) \right]$$

$$= n \sum_{i=1}^{a} k_{i.} d_i^2$$

keterangan:

$n = \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij}$ Jumlah total dari frekuensi matriks $\mathbf{P}$

$$d_i^2 = \left(\mathbf{r_i} - \mathbf{c}\right)' \left(\mathbf{r_i} - \mathbf{c}\right)$$

Besaran d_i^2 mempresentasikan jarak Kai Kuadrat antara profil baris ke-i dan rata-rata profil baris. Jarak ini disebut jarak Kai Kuadrat. Jarak d_i^2 mirip dengan jarak *Euclid* antara vektor $\mathbf{r_i}$ dan $\mathbf{c_j}$, dan besaran $\frac{\chi^2}{n}$ merupakan total *inersia*. Sehingga $\sum_{i=1}^{a} \mathrm{k_{i.}} d_i^2$ menunjukkan total *inersia* yang dinyatakan sebagai rata-rata pembobot jarak Kai Kuadrat antara profil baris dengan rata-ratanya.

Penguraian Nilai Singular (*Singular Value Decomposition*)

Untuk mereduksi dimensi data berdasarkan keragaman data (nilai *eigen*/*inersia*) terbesar dengan mempertahankan informasi optimum, diperlukan penguraian nilai singular. Penguraian nilai singular (*singular value decomposition*) merupakan salah satu konsep aljabar matriks dan konsep *eigen decomposition* yang terdiri dari nilai *eigen* (λ) dan vektor *eigen*. Teorema Dekomposisian Nilai Singular, yaitu misalkan A matriks berukuran $\mathrm{m} \times \mathrm{n}$ maka ada matriks diagonal Σ berukuran $r \times r$ dan $r \leq \min\{\mathrm{m,n}\}$, matriks orthogonal U berukuran $\mathrm{m} \times \mathrm{m}$, matriks orthogonal V berukuran $\mathrm{m} \times \mathrm{n}$, sehingga $A = U \Sigma V^t$ dengan Σ adalah matriks berukuran $\mathrm{m} \times \mathrm{n}$ yang mempunyai bentuk $\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$, $\lambda_1^2 \geq \ldots \geq \lambda_m^2$ adalah nilai inersia dari $U'U$. Berdasarkan Teorema Dekomposisian Nilai Singular tersebut, Matriks yang akan di *singular value decomposition* adalah matriks $U = \mathbf{D_r}^{-1/2}(\mathbf{K} - \mathbf{rc}')\mathbf{D_c}^{-1/2}$ yang hasilnya adalah :

$\mathbf{A}$ adalah matriks berukuran $(a \times \mathrm{m})$

$\mathbf{B}$ adalah matriks berukuran $(b \times \mathrm{m})$

dan $\mathbf{\Lambda}$ merupakan suatu matriks yang elemen-elemennya adalah nilai singular, dimana nilai singular adalah akar dari nilai *inersia*nya.

Penguraian Nilai Singular Umum

Untuk menentukan jumlah anak ruang *Euclid* dan memproyeksikan semua profil baris ke dalam anak ruang *Euclid* digunakan penguraian nilai singular umum atau *Generalized Singular Value Decomposition* (GSVD).

Koordinat dari baris dan kolomnya ditentukan dengan menggunakan GSVD dari matriks $(\mathbf{K}-\mathbf{rc}')$ hasilnya $\mathbf{A}\Lambda\mathbf{B}'$,dengan $\mathbf{A}$ adalah matriks berukuran $a\times \mathrm{m}$, $\mathbf{B}$ adalah matriks berukuran $b\times \mathrm{m}$, Λ adalah matriks diagonal yang mempunyai unsur-unsur diagonalnya nilai singular dari matriks $\mathbf{K}-\mathbf{rc}'$, dimana berlaku $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}=\mathbf{I}_m$ dan $\mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B}=\mathbf{I}_m$.

Tiap himpunan titik dapat dihubungkan dengan sumbu utama dari himpunan titik lainnya yaitu:

Rumusan untuk koordinat

	Rumusan dari koordinat baris	Rumusan untuk koordinat kolom
Analisis profil baris	$F=\mathbf{D_r^{-1}A\Lambda}$	$G=\mathbf{D_c^{-1}B}$
Analisis profil kolom	$F=\mathbf{D_r^{-1}A}$	$G=\mathbf{D_c^{-1}B\Lambda}$
Analisis keduanya (baris dan kolom)	$F=\mathbf{D_r^{-1}A\Lambda}$	$G=\mathbf{D_c^{-1}B\Lambda}$

***Nilai* Inersia**

Untuk mempresentasikan profil-profil baris dan profil-profil kolom ke dalam ruang berdimensi *d* $(\leq \mathrm{m})$. Koordinat dari *i* baris dari matriks yang dibentuk dengan mengambil *d* kolom pertama dari $F=\mathbf{D_r^{-1}A\Lambda}$, dan koordinat *j* profil kolom adalah *j* baris dari matriks yang dibentuk dengan mengambil *k* kolom pertama dari $G=\mathbf{D_c^{-1}B\Lambda}$. Karena total *inersia* yang mempresentasikan semua informasi dalam seluruh ruang adalah $n\,tr(E)=n\sum_i^m \lambda_i^2$, maka pendekatan ruang berdimensi *m* dengan ruang berdimensi *k* dikatakan baik jika $\sum_{i=1}^{d}\lambda_i^2$ mendekati total *inersia* $\sum_{i=1}^{m}\lambda_i^2$ atau $\sum_{i=d+1}^{m}\lambda_i^2$ mendekati nol.

Nilai *inersia* menunjukkan kontribusi dari baris ke-i pada *inersia* total. Sedangkan dimaksud *inersia* total adalah jumlah bobot kuadrat jarak titik-titik ke pusat, massa dan jarak yang didefinisikan :

inersia total baris : $in(a) = \sum_{i=1}^{a} \mathbf{r}_i \left(\mathbf{r}_{i.} - \mathbf{c}\right)' \mathrm{D}_c^{-1} \left(\mathbf{r}_{i.} - \mathbf{c}\right)$

inersia total kolom : $in(b) = \sum_{j=1}^{b} \mathbf{c}_{.j} \left(\mathbf{c}_{.j} - \mathbf{r}\right)' \mathrm{D}_r^{-1} \left(\mathbf{c}_{.j} - \mathbf{r}\right)$

Jumlah bobot kuadrat koordinat titik-titik dalam sumbu utama ke-d pada tiap-tiap himpunan yaitu λ_d^2 yang dinotasikan dengan λ_d. Nilai ini disebut sebagai *inersia* utama ke-d. Persamaan *inersia* utama baris dan kolom serta pusatnya dapat dinyatakan sebagai berikut:

inersia utama baris adalah $F'\mathbf{D}_r F = \mathbf{\Lambda}$

bukti $F'\mathbf{D}_r F = \mathbf{\Lambda}$, akan ditunjukkan sebagai berikut :

$$F'\mathbf{D}_r F = \left(\mathbf{D}_r^{-1}\mathbf{A}\mathbf{\Lambda}\right)' \mathbf{D}_r (\mathbf{D}_r^{-1}\mathbf{A})$$

$$= \mathbf{\Lambda}'\mathbf{A}' \left(\mathbf{D}_r^{-1}\right)^{-1} \mathrm{I}\mathbf{A}$$

$= \mathbf{\Lambda}'\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}$, dengan menggunakan persamaan $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{I}_m$,

didapatkan $\mathbf{\Lambda}'\mathbf{I}_m = \mathbf{\Lambda}'$. Karena matriks $\mathbf{\Lambda}'$ adalah simetris sehinggga $\mathbf{\Lambda}' = \mathbf{\Lambda}$

jadi $F'\mathbf{D}_r F = \mathbf{\Lambda}$.

inersia utama kolom adalah $G'\mathbf{D}_c G = \mathbf{\Lambda}$

bukti $G'\mathbf{D}_c G = \mathbf{\Lambda}$, akan ditunjukkan sebagai berikut :

$$G'\mathbf{D}_c G = \left(\mathbf{D}_c^{-1}\mathrm{B}\mathbf{\Lambda}\right)' \mathbf{D}_c \left(\mathbf{D}_c^{-1}\mathbf{B}\right)$$

$$= \mathbf{\Lambda}'\mathbf{B}' \left(\mathbf{D}_c^{-1}\right)^{-1} \mathbf{I}\mathbf{B}$$

$= \mathbf{\Lambda}'\mathbf{B}'\mathbf{D}_r^{-1}\mathbf{B}$, dengan menggunakan $\mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}_m$, didapatkan $\mathbf{\Lambda}'\mathbf{I}_m = \mathbf{\Lambda}'$. Karena matriks $\mathbf{\Lambda}'$ adalah simetris sehingga $\mathbf{\Lambda}' = \mathbf{\Lambda}$ jadi $G'\mathbf{D}_c G = \mathbf{\Lambda}$.

Besaran $\lambda_1^2, \ldots, \lambda_i^2$ dapat diinterpretasikan sebagai besarnya kontribusi yang diberikan pada total *inersia* oleh masing-masing dimensi pertama, kedua dan seterusnya. Sehingga besaran relatif untuk mengukur besarnya kehilangan informasi dapat dirumuskan sebagai berikut:

$$L = 1 - \frac{\sum_{i=1}^{d} \lambda_i^2}{\sum_{i=1}^{m} \lambda_i^2}$$

Uji Kesesuaian Kai Kuadrat (*Test of Goodness of Fit*)

Sebuah metode analisis non-parametrik yang digunakan ialah uji Kai Kuadrat. Uji Kai Kuadrat tidak dibatasi oleh asumsi-asumsi ketat tentang jenis populasi maupun parameter populasi. Metode ini sangat bermanfaat ketika data yang tersedia hanya berupa frekuensi, misalnya banyaknya subjek dalam kategori sakit dan tidak sakit, atau banyaknya penderita *diabetes mellitus* dalam kategori I, II, III, IV menurut tingkat penyakitnya.

Teknik yang dikembangkan oleh *Pearson* tahun 1900 itu melibatkan perhitungan suatu kuantitas yang disebut Kai Kuadrat, berasal dari huruf Yunani "Chi" $\left(\chi\right)$.

Uji Kai Kuadrat berguna untuk tiga macam kebutuhan:

1. Menguji kesesuaian (*test of goodness of fit*). Dengan uji kesesuaian, suatu distribusi sampel dievaluasi apakah sesuai (*fit*) dengan distribusi populasi tertentu.
2. Menguji ketergantungan (*test of independent*). Dengan uji *independent* diperiksa apakah dua buah variabel dari sebuah sampel saling tergantung atau tidak saling tergantung.
3. Menguji homogenitas (*test of homogeneity*). Dengan uji homogenitas, beberapa sampel dievaluasi apakah berasal dari populasi-populasi yang sama (homogen) dalam hal variabel tertentu.

Dalam melakukan uji Kai Kuadrat, ada syarat-syarat yang perlu dipenuhi:

1. Sampel yang dipilih acak
2. Semua pengamatan dilakukan independen
3. Setiap sel paling sedikit berisi frekuensi harapan sebesar 1. Sel-sel dengan frekuensi harapan kurang dari 5 tidak melebihi 20% dari total sel.Untuk tabel 2×2, berarti tidak satu sel pun boleh berisi frekuensi harapan < 5.
4. Sampel sebaiknya lebih dari 40.

Uji yang sesuai untuk mengetahui ada tidaknya hubungan antara dua variabel kategori yang berupa tabel kontingensi, adalah *Pearson Chi-Square test* statistik ujinya adalah :

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(\mathrm{p}_{ij} - m_{ij}\right)^2}{m_{ij}}$$

keterangan :

p_{ij} = jumlah pengamatan pada baris ke-i dan kolom ke-j

$p_{i.}$ = jumlah pengamatan pada baris ke-i

$p_{.j}$ = jumlah pengamatan pada kolom ke-j

m_{ij} = frekuensi harapan

a = banyaknya baris

b = banyaknya kolom

Uji indenpendensi *Pearson Chi-Square* dapat digunakan jika nilai harapan kurang dari (m_{ij} <5) tidak lebih dari 20% (maksimal 20%).

Koefisien Kontingensi

Untuk melihat keeratan hubungan atau kecenderungan antara variabel satu dengan yang lainnya. Dengan menggunakan rumusan koefisien kontingensi sebagai berikut :

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

keterangan:

χ^2 = Statistik uji Kai Kuadrat

N = banyaknya populasi sampel

Nilainya $0 \leq C < 1$

Analisis Biplot

Analisis Biplot merupakan suatu cara untuk memberikan peragaan grafis dari suatu matriks data X dalam suatu plot. Peragaan secara grafis ini dilakukan dengan menumpang tindihkan vektor baris dari matriks X dengan vektor kolom dari matriks X. Vektor baris dari matriks X menggambarkan objek yang diamati sedangkan vektor kolom dari matriks X menggambarkan variabel. Dari peragaan ini diharapkan akan diperoleh gambaran tentang objek, seperti kedekatan antar objek dan gambaran tentang variabel baik tentang keragaman maupun korelasi, serta keterkaitan antar objek-objek dengan variabel-variabel.

Landasan analisis ini ialah setiap matriks $n \text{ X } p$ yang berpangkat r dengan $r \le \min\{n, p\}$ dapat digambarkan dalam ruang berdimensi r. Bagi matriks yang berpangkat r dan akan digambarkan dalam ruang berdimensi k dengan $k \le r$, maka terlebih dahulu dilakukan pendekatan optimum dengan suatu matriks berpangkat k. Pendekatan optimum ini berdasarkan pada perbedaan kuadrat norma terkecil antara kedua matriks tersebut. Dari matriks hasil pendekatan tersebut, dapat digambarkan konfigurasi objek dan variabel dalam ruang berdimensi k. Untuk memudahkan pemahaman tentang masalah ini, maka dapat dimisalkan dengan $k = 2$. Dengan menggunakan nilai k ini maka pendekatan tersebut dapat digambarkan dalam bidang atau ruang berdimensi dua.

Menurut Sartono, dkk (2003), analisis Biplot didasarkan pada *Singular Value Decomposition* (SVD). Bentuk umum SVD oleh Greenacre (1984) dijelaskan sebagai berikut. Misalkan suatu matriks data X berukuran $n \text{ X } p$ dimana n adalah pengamatan dan p adalah variabel yang dikoreksi terhadap nilai rata-rata. Matriks X ini mempunyai pangkat r, dan dapat dituliskan menjadi:

$$X = ULA'$$

matriks U merupakan matriks vektor singular yang berukuran $(n \text{ X } r)$ dan matriks A merupakan matriks vektor singular yang berukuran $(p \text{ X } r)$ sehingga $U'U = A'A = I_r$ (matriks identitas berdimensi r). Sedangkan L adalah matriks diagonal yang berukuran $(r \text{ X } r)$ dengan unsur-unsur diagonal adalah akar kuadrat dari akar ciri-akar ciri $X'X$ atau XX', sehingga $\sqrt{\lambda_1} \ge \sqrt{\lambda_2} \ge \ldots \ge \sqrt{\lambda_n}$. Unsur-unsur diagonal ini disebut nilai singular matriks X.

Kolom-kolom matriks A disebut vektor singular baris yang merupakan landasan ortonormal baris-baris matriks X dalam ruang berdimensi p. Kolom-kolom matriks U disebut vektor singular kolom yang merupakan landasan ortonormal kolom-kolom matriks X dalam ruang berdimensi n. Dengan penjabaran, persamaan diatas menjadi:

$$X = UL^{\alpha} L^{1-\alpha} A'$$

Misalkan $G = UL^{\alpha}$ dengan G adalah matriks berukuran n X r dan $H' = L^{1-\alpha} A'$ dengan H' adalah matriks berukuran p X r. α adalah nilai faktorisasi yang besarnya $0 \le \alpha \le 1$, sehingga diperoleh

$$X = GH'$$

dimana

$$G = \begin{bmatrix} g'_1 \\ g'_2 \\ . \\ . \\ g'_n \end{bmatrix} \text{ dan } H = \begin{bmatrix} h'_1 \\ h'_2 \\ . \\ . \\ h'_n \end{bmatrix}$$

atau

$$X_{ij} = g'_i h_j$$

dimana: $i = 1,2,3,...,n; \quad j = 1,2,3,...,p.$ Dengan g'_i adalah baris-baris matriks G dan h'_j baris-baris matriks H.

Jika X berpangkat dua, maka vektor pengaruh baris g_i dan vektor pengaruh kolom h_j dapat digambarkan dalam ruang berdimensi dua. Namun jika X berpangkat lebih dari dua, maka diselesaikan dengan matriks X berpangkat dua. Sehingga persamaan terakhir menjadi:

$${}_2X_{ij} = g_i^{*'} h_j^{*}$$

dengan ${}_2X_{ij}$ merupakan unsur pendekatan matriks X pada dimensi dua, sedangkan g_i^{*} dan h_j^{*} mengandung dua unsur pertama vektor g_i dan h_j.

Nilai α yang digunakan dapat merupakan nilai sembarang antara $0 \le \alpha \le 1$. Akan tetapi pengambilan nilai-nilai ekstrim $\alpha = 0$ dan $\alpha = 1$ akan berguna

dalam interpretasi Biplot. Jika $\alpha = 0$ yang digunakan maka diambil $G = U$ dan $H = AL.$ Sehingga diperoleh:

$$\begin{aligned} X'X &= \left(GH'\right)'\left(GH'\right) \\ &= HG'GH' \\ &= HU'UH' \\ &= HH' \end{aligned}$$

Karena $X'X = HH' = (n-1)S,$ maka hasil kali $h_j'h_k$ akan sama dengan $(n-1)$ kali varian S_{jk} dengan $h_j'h_j$ menggambarkan keragaman variabel ke- $j.$ Oleh karena itu, korelasi antara variabel ke- j dan ke- k ditunjukkan oleh nilai sudut kosinus antara vektor h_j dan $h_k.$ Jarak Euclid antara objek ke- h dan ke- i dalam Biplot akan sebanding dengan jarak Mahalanobis antara pengamatan ke- h dan ke- $i.$

Jika α yang digunakan adalah $\alpha = 1,$ maka dipilih $G = UL$ dan $H = A,$ sehingga diperoleh hubungan berikut.

$$\begin{aligned} XX' &= \left(GH'\right)\left(GH'\right)' \\ &= GH'HG \\ &= GA'AG' \\ &= GG' \end{aligned}$$

Pada keadaan ini, jarak Euclid antara g_h dan g_i akan sama dengan jarak Euclid antara x_h dan $x_i.$ Selain itu vektor pengaruh baris ke- i sama dengan skor komponen utama untuk objek ke- i dari hasil analisis komponen utama. Hal ini dikarenakan pengambilan $G = UL$ sehingga unsur ke- k dari g_i adalah $u_{ik}\sqrt{\lambda_{ik}} = Z_{ik}$ yang merupakan skor komponen utama ke- k dari pengamatan ke- $i,$ dan dari $H = A$ diperoleh bahwa vektor pengaruh kolom h_j sama dengan $a_j,$ yaitu vektor pembobot variabel ke- j pada komponen uatama ke- $k.$

Pembuatan Biplot

Secara umum, langkah-langkah pembuatan Biplot ini adalah sebagai berikut:

1. Transformasi matriks X.
2. Menentukan matriks S*ingular Value Decomposition* (SVD) ULA'.
3. Menghitung faktor pembobot λ untuk baris dan kolom.

   $$\lambda_{b,1} = \sigma_1^{\tau} \qquad \lambda_{b,2} = \sigma_2^{\tau}$$
   $$\lambda_{c,1} = \sigma_1^{1-\tau} \qquad \lambda_{c,2} = \sigma_2^{1-\tau}$$

   dimana σ_1 dan σ_2 adalah nilai singular pertama dan kedua dan τ adalah *split factor*.
4. Menghitung nilai-nilai dari setiap baris matriks. Nilai setiap baris dihitung dengan menggunakan:

   $$xb_i = U_{i1}\lambda_{b,1} \qquad yb_i = U_{i2}\lambda_{b,2}$$
5. Menghitung nilai-nilai dari setiap kolom matriks. Nilai setiap kolom dihitung dengan menggunakan:

   $$xc_j = A_{i1}\lambda_{c,1} \qquad yc_j = A_{i2}\lambda_{c,2}$$
6. Menghubungkan nilai (X, Y) untuk baris dan kolom.
7. Kemudian semua nilai yang ada dihubungkan dengan garis lurus untuk menggambarkan keadaan setiap variabel.

Pembuatan Biplot dapat dilakukan dengan melakukan transformasi terlebih dahulu. Transformasi yang diperbolehkan antara lain transformasi fungsional seperti logaritma, sinus dan lain-lain, transformasi matriks seperti rasio frekuensi yang diamati (Rasio Kontingensi Pearson), rasio kemungkinan, pemberian pembobot dari baris dan atau kolom, pemusatan baris atau kolom yang terboboti, pemusatan ganda terboboti, dan normalisasi baris atau kolom yang terboboti.

Dalam beberapa kasus pemilihan transformasi ini berhubungan dengan prosedur standar dalam analisis yang disebut *Canned Analysis*. Dalam Biplot XLS terdapat beberapa definisi awal untuk transformasi yang mempunyai tipe data yang berbeda-beda, yaitu:

a. Analisis Komponen Utama (AKU) tanpa normalisasi. Transformasi ini dimulai dengan pemusatan kolom-kolom ⬭ SVD, dan pembuatan Biplot.
b. Analisis Komponen Utama (AKU) dengan normalisasi. Transformasi ini dimulai dengan pemusatan kolom-kolom ⬭ Normalisasi dari kolom-kolom ⬭ SVD, dan pembuatan Biplot.
c. Analisis Koresponden Sederhana. Transformasi ini diawali dengan melakukan transformasi rasio kontingensi Pearson ⬭ penerapan pembobot baris dan kolom ⬭ pemusatan ganda terboboti ⬭ SVD yang terboboti, dan pembuatan Biplot.

d. *Logratio analysis*. Transformasi ini diawali dengan melakukan transformasi logaritma ⬭ pemusatan ganda ⬭ SVD, dan pembuatan Biplot.

e. *Ratio maps*. Transformasi ini diawali dengan melakukan transformasi logaritma ⬭ penerapan pembobot baris dan kolom ⬭ pemusatan ganda terboboti ⬭ SVD terboboti, dan pembuatan Biplot.

Dua tipe transformasi pertama yang dijelaskan diatas digunakan untuk Biplot Komponen Utama. Sedangkan tipe transformasi ketiga, hanya digunakan untuk menggambarkan frekuensi dari tabel kontingensi. Data yang ada pada tabel ini akan ditransformasikan kedalam analisis Biplot koresponden.

Tipe-Tipe Biplot

Biplot mempunyai beberapa tipe. Perbedaan tipe ini berdasarkan pada nilai α yang digunakan. Nilai α yang digunakan dalam Biplot adalah $0 \le \alpha \le 1.$ Namun nilai α yang lazim digunakan adalah $\alpha = 1; \alpha = 0.5;$ dan $\alpha = 0.$ Jika α yang digunakan adalah $\alpha = 1$ maka Biplot yang dibentuk disebut Biplot RMP (*Row Metric Preserving*). Biplot RMP ini digunakan untuk menduga jarak Euclid secara optimal. Biplot dengan $\alpha = 1$ disebut juga dengan Biplot komponen utama.

Jika α yang digunakan adalah $\alpha = 0,$ maka akan terbentuk tipe Biplot yang disebut Biplot CMP (*Column Metric Preserving).* Salah satu contoh dalam interpretasi hubungan antar variabel ini adalah interpretasi dari matriks kovarian atau matriks korelasi. Selain itu, Biplot ini disebut juga Biplot faktor komponen utama. Nilai α lain yang digunakan dalam pembuatan Biplot yaitu $\alpha = 0{,}5.$ Untuk nilai α ini, Biplot yang dibentuk disebut Biplot Simetri atau Biplot SQRT (S*quare Root Biplot)*. Tipe Biplot ini merupakan tipe Biplot yang membuat kesamaan penskalaan atau pemobot untuk baris dan kolom. Kesamaan penskalaan ini diperlukan dalam interpretasi hubungan dari dua faktor pengamatan.

Biplot RMP dan Biplot CMP dapat dikembangkan menjadi suatu tipe Biplot baru. Tipe Biplot baru tersebut disebut Biplot RCMP (*Row-Column Metric Preserving)*. Biplot ini mempunyai nilai $\alpha > 1.$ Penggunaan dua tipe Biplot yang berbeda ini dilakukan dengan menggunakan data asal. Keuntungan dari Biplot ini adalah dapat memberikan gambaran maksimum atau dapat menggambarkan kedua informasi yang terdapat pada baris dan kolom matriks. Pada Biplot ini, agar dapat menggambarkan setiap nilai yang ada pada matriks maka dimensi yang digunakan berada pada ruang dengan dimensi rendah.

Analisis Konjoin

Pada riset pemasaran banyak ditemukan bagaimana cara mendesain suatu produk yang banyak diminati oleh konsumen, salah satunya adalah produk *flash disk*. Sebagaimana lazimnya sebuah produk, terdapat beberapa atribut yang mempengaruhi konsumen untuk membeli *flash disk* yaitu kapasitas, ukuran, harga, fitur tambahan, dan bahan. Pengukuran dan analisis dalam penelitian pemasaran untuk memilih suatu produk biasanya dilakukan dengan menggunakan analisis konjoin.

Analisis konjoin (*Conjoint Analysis)* merupakan suatu metode analisis dalam analisis multivariat, analisis ini mulai diperkenalkan pada tahun 1970-an (Cattink and Wittink, 1992). Analisis ini digunakan untuk membantu mendapatkan atau komposisi atribut-atribut suatu produk baik baru maupun lama yang paling banyak disukai konsumen. Hasil utama konjoin adalah suatu bentuk (desain) produk barang atau jasa atau objek tertentu yang diinginkan oleh sebagian besar responden (Santoso, 2002). Menurut Hair *et al*., (1998), dalam prosesnya analisis konjoin akan memberikan ukuran kuantitatif terhadap tingkat kegunaan (*utility)* dan kepentingan relatif (*relatif importance*) suatu atribut dari produk

Terdapat beberapa ketentuan dalam memilih metode yang akan digunakan dalam analisis konjoin(Hair *et al*., 1998), yaitu :

- Choice-Based Conjoint (CBC). Digunakan apabila jumlah atribut ≤ 6
- Traditional Conjoint, digunakan apabila jumlah atribut < 10
- Adaptive Conjoint Analysis (ACA), digunakan apabila jumlah atribut ≥ 10

Merancang Kombinasi Atribut

Metode yang digunakan untuk merancang kombinasi taraf dari atribut pada penelitian ini adalah dengan menggunakan *Choice-Based Conjoint (CBC)*. Metode ini mulai popular pada awal tahun 1990-an, dan saat ini banyak digunakan serta mendapat perhatian yang sangat besar oleh kalangan peneliti dan praktisi pemasaran..

Menurut Hair *et al.*, (1998) keunggulan utama *Choice-Based* dibandingkan metode lain adalah prosedur pengumpulan datanya secara langsung mencerminkan prilaku pasar. Metode *Choice-Based Conjoint,* tidak tepat digunakan dalam penelitian dengan jumlah atribut yang banyak. Green dan Srinivasan (1990), menyatakan bahwa 6-10 atribut adalah jumlah maksimum atribut yang dapat menggunakan konsep *full-profile* dalam analisis konjoin.

Metode Analisis

Konjoin termasuk dalam *Multivariate Dependence Method***,** yaitu hubungan antara variabelnya dependensi. Hubungan dependensi yaitu jika variabel-variabel yang diteliti secara teoritis dapat dipisahkan kedalam variabel-variabel respon dan variabel penjelas (Santoso, 2002) dengan model sebagai berikut :

$$Y = X_1 + X_2 + X_3 + \ldots + X_k$$

(metrik atau *non*metrik) (*non*metrik)

Variabel independen (X) adalah faktor dan berupa data *non*metrik. Termasuk disini adalah bagian dari faktor (level). Sedangkan variabel dependen (Y) adalah pendapat keseluruhan (*overall preference*) dari seorang responden terhadap sekian faktor dan level pada sebuah produk.

Secara umum model dasar analisis konjoin untuk pilihan responden (r_i) untuk setiap faktor dan level ditulis dalam bentuk :

$$r_i = \beta_0 + \sum_{j=1}^{p} u_{jk_{ji}} \qquad (1)$$

Keterangan:

r_i : Kegunaan atau Utility total

β_0 : Intersep model respon

$u_{jk_{ji}}$: Sumbangan *the part worth* atau utility dari factor ke-*j* level ke- k_{ji} .

p : Jumlah atribut.

fungsi kegunaan (*utility function*) yang nantinya akan diduga sangat dipengaruhi oleh model yang akan digunakan dibawah ini :

1. Faktor diskret

$$\hat{u}_{jk} = \begin{cases} \hat{\alpha}_{jk} & untuk, \quad k = 1,\ldots,m_j - 1 \\ -\sum_{j=1}^{m_j-1} \hat{\alpha}_{jk}, & untuk \quad k = m_j \end{cases} \qquad (2)$$

2. Faktor linier $\hat{u}_{jk} = \hat{\beta}_j x_k$ (3)

3. Faktor ideal-point $\hat{u}_{jk} = \hat{\gamma}_{j1} z_{jk} + \hat{\gamma}_{j2} z^2_{jk}$ (4)

Saat ini terdapat beberapa metode atau prosedur yang dapat digunakan untuk menyelesaikan model dari analisis konjoin, salah satunya adalah metode regresi dengan variabel *dummy*.

Untuk atribut ke- j dengan k_j level, variabel *dummy*nya adalah :

Tabel.1 Variabel *Dummy* Atribut ke- j **dan Level** k_j

Level	x_1	x_2	⋯	X_{k_j-1}
1	1	0	⋯	0
2	0	1	⋯	0
3	0	0	⋯	0
⋮	⋮	⋮		⋮
$k_j - 1$	0	0	⋯	1
k_j	0	0	⋯	0

Langkah yang paling penting dalam analisis konjoin adalah mengestimasi kegunaan (*utility function*) atau tingkat kepentingan relatif individu (*individual level part worth*).

Untuk mendapatkan nilai-nilai $u_{jk_{ji}}$ tersebut, langkah yang harus dilakukan adalah mengestimasi model dasar analisis konjoin dengan persamaan regresi linier ganda dengan variabel *dummy*. Maka persamaan regresinya adalah :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \quad (5)$$

Untuk menaksir parameter pada persamaan (5) maka akan digunakan metode kuadrat terkecil.

Pentingnya suatu atribut, misalnya RANGE_i dinyatakan dalam kisaran *Part Worth* melintasi level dari atribut, yaitu :

$$\text{RANGE}_i = \{\max(u_{jk_{ji}}) - \min(u_{jk_{ji}})\}, \text{ untuk setiap i} \quad (6)$$

Selanjutnya, pentingnya suatu atribut digunakan untuk menyakinkan kepentingan relatif dengan atribut lainnya. Kepentingan relatif disimbolkan dengan IMP yang ditentukan melalui formula berikut :

$$IMP_i = \frac{RANGE}{\sum_{i=1}^{p} RANGE_i} \times 100\% \quad (7)$$

Setelah didapatkan nilai-nilai $u_{jk_{ji}}$, maka kisaran *part worth* RANGE_i dan timbangan kepentingan relatif IMP_i akan diperoleh. Kisaran *part worth* dan timbangan kepentingan relatif ini memberikan dasar untuk menginterpretasikan hasil. Angka IMP_i yang terbesar menunjukkan preferensi terbesar terhadap level-level pada atribut tertentu.

Uji Realibilitas dan Validitas

Tahapan selanjutnya yang perlu dilakukan dalam analisis konjoin adalah menilai keandalan dan kesahihan. Jika perosedur konjoin menggunakan regresi dengan variabel dummy, maka ketepatan/kecocokan dari estimasi model digunakan nilai koefisien determinasi berganda R^2 (Supranto, 2004). Koefisien determinasi (R^2) adalah persentase keragaman variabel bebas yang dapat dijelaskan oleh model persamaan regresi. Nilai (R^2) persamaan regresi yang makin mendekati 100% menunjukkan bahwa makin banyak keragaman variabel bebas yang dapat dijelaskan dari persamaan regresi tersebut.

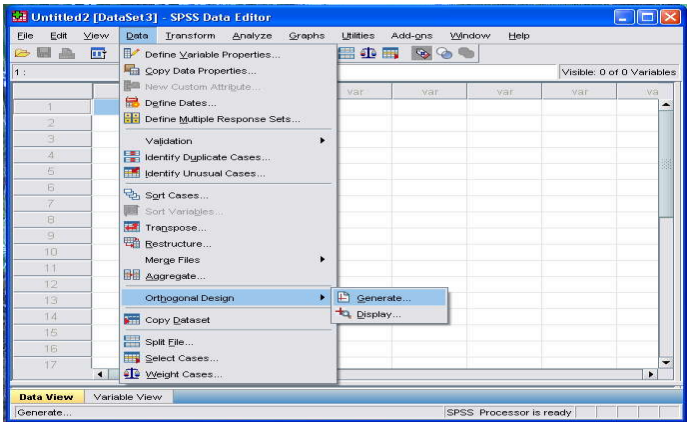
ANALISIS DATA

Analisis konjoin untuk mengetahui preferensi mahasiswa Matematika FMIPA yang menggunakan *flash disk* dilakukan dengan menggunakan Program SPSS Versi 16.0. Proses analisis konjoin dilakukan melalui tiga langkah yaitu:

1. Merancang Kartu Stimuli

Buka Program SPSS, untuk membuat stimuli dengan orthogonal design

►Dari Menu, buka **Data** kemudian pilih **Orthogonal Design**, lalu klik **Generate...**

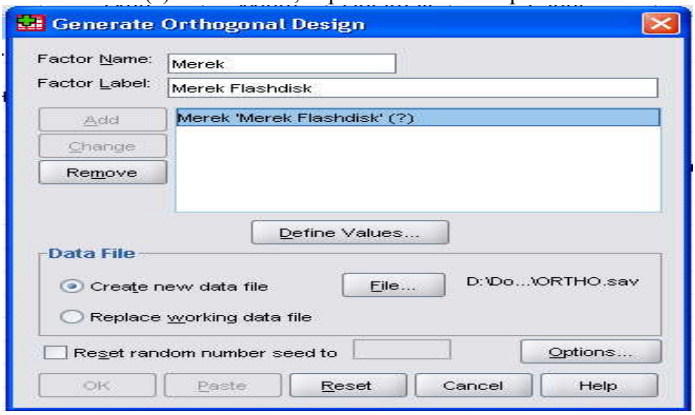


Gambar 1. Tampilan SPSS Data Editor

- Setelah itu akan tampil kotak dialog "**Generate Orthogonal Design**",
- Langkah selanjutnya setelah tampilan muncul kemudian mendeskripsikan variabel-variabel yang akan digunakan dalam analisis. Ketik variabel MEREK pada Factor Name, kemudian Merek *Flash disk* pada **Factor Label**, untuk lebih jelasnya dapat dilihat dari tampilan berikut ini :

Gambar 3. Pemberian Nama Atribut Pada Kotak Dialog Generate Orthogonal Design

- Selanjutnya klik **Add**, kemudian akan tampil MEREK 'Merek Flashdisk'(?) klik item ini, dapat dilihat dari tampilan berikut ini :



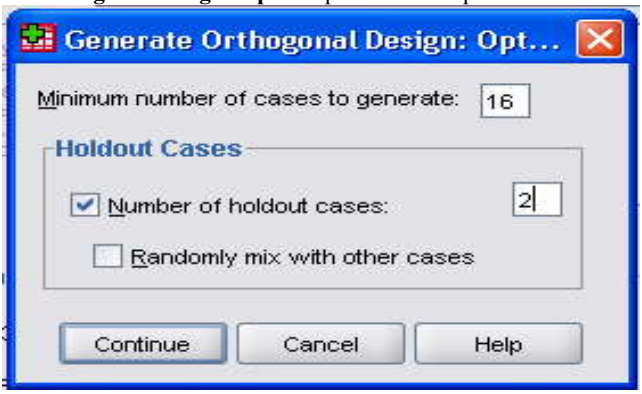
Gambar 4. Tampilan Nama Atribut Pada Kotak Dialog Generate Orthogonal Design

- Selanjutnya klik **Define Values....** Maka akan tampil kotak dialog **Generate Design Define Values**. Masukkan masing-masing nilai pada kotak **Value** dan nama pada kotak **Label** untuk setiap taraf dari faktor. Untuk faktor merek, value dan label. Kemudian klik **Continue**.

- Kemudian ulangi langkah Langkah 2-5 tersebut untuk mendefenisikan faktor

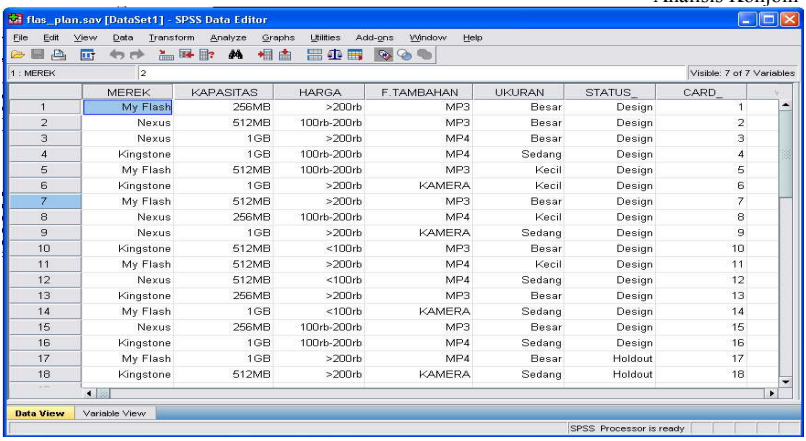
Gambar 5. Atribut Yang Sudah di Definisikan Pada Kotak Dialog

- Pilih **Replace working data file** pada kotak pilihan **Data File**
- Pilih **Reset random number seed to** dan masukkan angka **2000000.**
- Klik menu **Options**, sehingga akan tampil kotak dialog **Generate Orthogonal Design: Opt...** dapat dilihat tampilan berikut :



Gambar 6. kotak Dialog Generate Orthogonal Design: Opt...

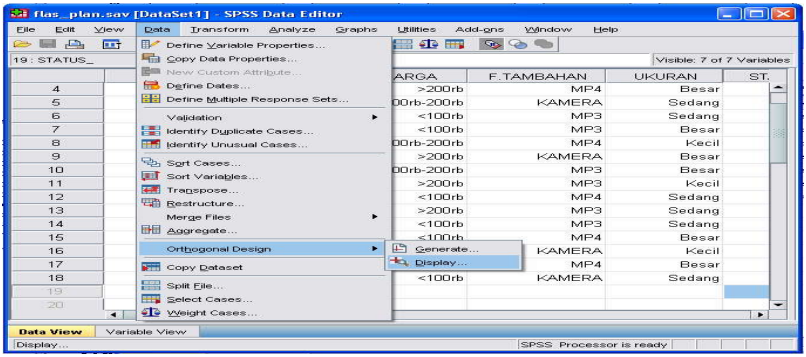
- Langkah selanjutnya karena pada penelitian ini diinginkan adanya 18 stimuli, masukkan angka 18 pada **Minimum number of cases to generate**. Pada kotak **Holdout Cases**, pilih **Number of houldout cases** dan masukkan angka 2.
- Lalu klik **Continue**. Pada kotak dialog **Generate Orthogonal Design**, klik **OK** Sehingga pada Data Editor akan tampil data sebagai berikut :



Gambar 7. Rancangan Stimuli Pada Data Editor

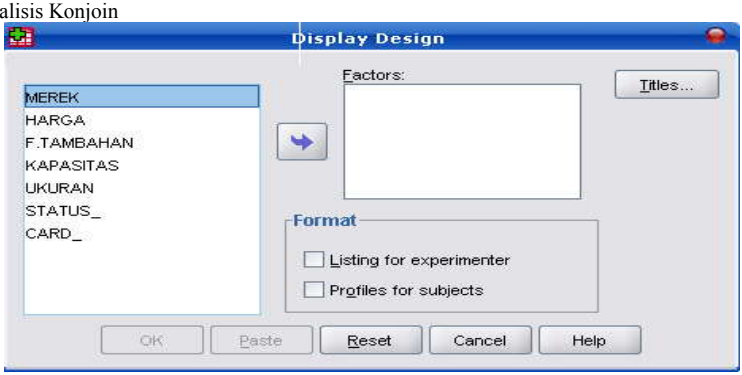
- Pilih menu **File**-> **Save** dan simpan Data Editor dengan nama file **flas_plan.sav**
- Kemudian langkah selanjutnya untuk display orthogonal design :

► Dari menu buka **Data** kemudian pilih **Orthogonal Design**, lalu klik **Display**.....



Gambar 8. Tampilan SPSS Data Editor Untuk Display

- Setelah itu akan muncul kotak dialog **Display Design**, seperti berikut :



Gambar 9. Kotak Dialog Display Design

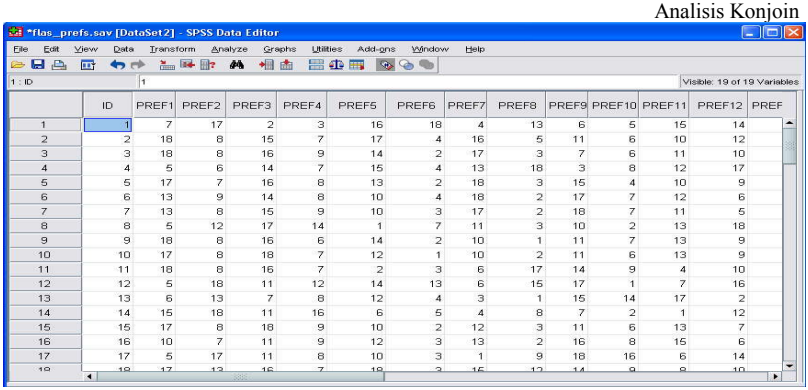
- Masukkan variabel **Merek**, **Kapasitas**, **Harga**, **F.Tambahan** dan **Ukuran** ke dalam kotak **Factors**. Klik item variabel, kemudian klik tanda  . Pada kotak **Format**, pilih **Listing for experimenter** untuk menampilkan seluruh stimuli ke dalam satu tabel dan **Profiles for subjects** untuk menampilkan setiap stimuli ke dalam tabel-tabel terpisah. Sehingga kotak dialog menjadi seperti berikut :

Gambar 10. Kotak Format Pada Kotak Dialog Display Design

- Klik **OK** untuk menjalankan analisis.

2. Membuat Data Preferensi

Setelah dilakukan pengumpulan data, maka data preferensi dimasukkan ke satu Data Editor dengan nama **flas_prefs.sav**. tampilan data pada program SPSS adalah sebagai berikut :



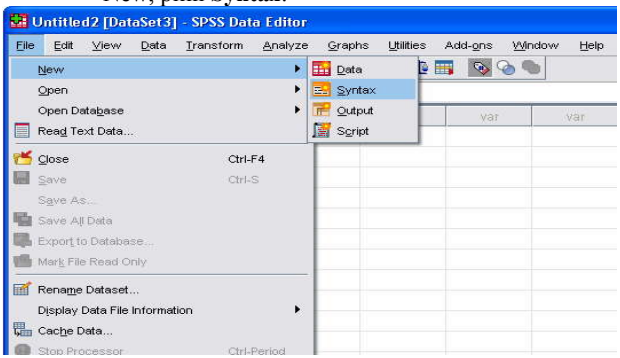
	ID	PREF1	PREF2	PREF3	PREF4	PREF5	PREF6	PREF7	PREF8	PREF9	PREF10	PREF11	PREF12	PREF
1	1	7	17	2	3	16	18	4	13	6	5	15	14	
2	2	18	8	15	7	17	4	16	5	11	6	10	12	
3	3	18	8	16	9	14	2	17	3	7	6	11	10	
4	4	5	6	14	7	15	4	13	18	3	8	12	17	
5	5	17	7	16	8	13	2	18	3	15	4	10	9	
6	6	13	9	14	8	10	4	18	2	17	7	12	6	
7	7	13	8	15	9	10	3	17	2	18	7	11	5	
8	8	5	12	17	14	1	7	11	3	10	2	13	18	
9	9	18	8	16	6	14	2	10	1	11	7	13	9	
10	10	17	8	18	7	12	1	10	2	11	6	13	9	
11	11	18	8	16	7	2	3	6	17	14	9	4	10	
12	12	5	18	11	12	14	13	6	15	17	1	7	16	
13	13	6	13	7	8	12	4	3	1	15	14	17	2	
14	14	15	18	11	16	6	5	4	8	7	2	1	12	
15	15	17	8	18	9	10	2	12	3	11	6	13	7	
16	16	10	7	11	9	12	3	13	2	16	8	15	6	
17	17	5	17	11	8	10	3	1	9	18	16	6	14	

Gambar 11. Data Editor Untuk Data Preferensi Responden

3. Membuat Syntax Konjoin

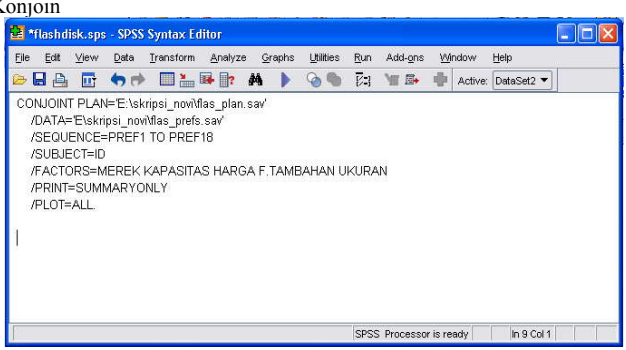
. Langkah-langkah penulisan dan eksekusi syntax dilakukan dengan cara sebagai berikut :

- Pada Data Editor yang kosong, klik **File**. Kemudian pada menu **New**, pilih **Syntax**.



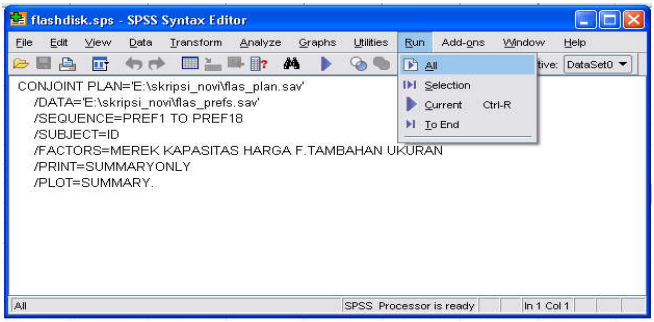
Gambar 12. Menampilkan Syntax Editor Pada SPSS 16.0

- Pada layar Syntax Editor masukkan kalimat perintah untuk melakukan analisis konjoin,:



Gambar 13. CONJOINT Command Pada Syntax Editor

- Dari tampilan Syntax Editor SPSS 16 , klik menu **Run**, kemudian klik **All**.



Gambar 14. Mengeksekusi CONJOINT Command Pada Syntax Editor

Hasil dan Interpretasi Analisis Konjoin

1. Nilai *Utility* (Kegunaan)

Nilai (*utility*) adalah nilai yang menyatakan utilitas masing-masing level dalam faktor. Apabila dalam grafik *utility* adalah positif, maka berarti responden tersebut menyukai level tersebut, dan apabila negatif berarti responden tidak menyukai level tersebut. Nilai *utility* secara umum dapat dilihat pada tabel berikut:

Tabel 2. Rata-rata Nilai *Utility* (Kegunaan) Pada Atribut *Flash Disk*.

Utility

		Utility Estimate	Std. Error
MEREK	Kingstone	.268	.403
	My Flash	.006	.278
	Nexus	-.274	.328
KAPASITAS	256MB	-.154	.429
	512MB	-.168	.279
	1GB	.322	.462
HARGA	<100rb	.011	.359
	100rb-200rb	.031	.406
	>200rb	-.043	.291
F.TAMBAHAN	MP3	.022	.249
	MP4	.002	.292
	KAMERA	-.024	.310
UKURAN	Besar	.064	.253
	Sedang	-.184	.302
	Kecil	.120	.289
(Constant)		8.482	.216

Masing–masing nilai *utility* adalah variabel x_{ij} atribut ke-i level ke-j dengan nilai konstanta $\beta_0 = 8.482$, maka model analisis konjoin untuk preferensi mahasiswa Matematika FMIPA dalam pemilihan *flash disk* adalah :

$$r_i = 0.268\,x_{11} + 0.006\,x_{12} - 0.274\,x_{13} - 0.154\,x_{21} - 0.168\,x_{22} + 0.322\,x_{23} + 0{,}011\,x_{31} + 0.031\,x_{32} - 0.043\,x_{33} + 0.022\,x_{41} + 0.002\,x_{42} - 0.024\,x_{43} + 0.064\,x_{51} - 0.168\,x_{52} + 0.12\,x_{53} + 8.482$$

2. Nilai *Importance* (Kepentingan)

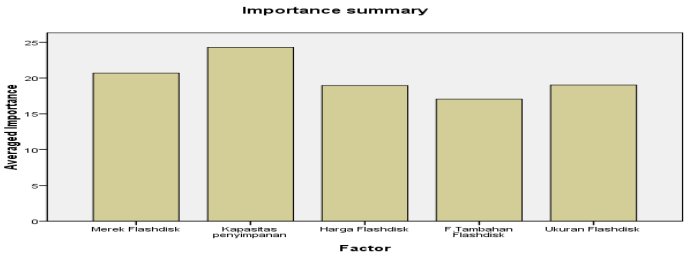
Dari analisis diperoleh nilai *importance* (kepentingan), yang mana nilai tersebut merupakan gabungan pendapat responden terhadap faktor yang dimaksud. Nilai *importance* digunakan untuk mengetahui faktor mana yang dianggap terpenting oleh responden dalam memilih *flash disk*. Nilai yang tertinggi dianggap faktor yang terpenting dalam memilih suatu produk. Hasil analisis konjoin untuk nilai *importance* secara umum dapat dilihat pada tabel dibawah ini :

Tabel 3. Rata-rata Nilai *Importance* (Kepentingan) Atribut *Flash Disk*

	Importance Values
MEREK	20.690
KAPASITAS	24.289
HARGA	18.944
F.TAMBAHAN	17.046
UKURAN	19.031

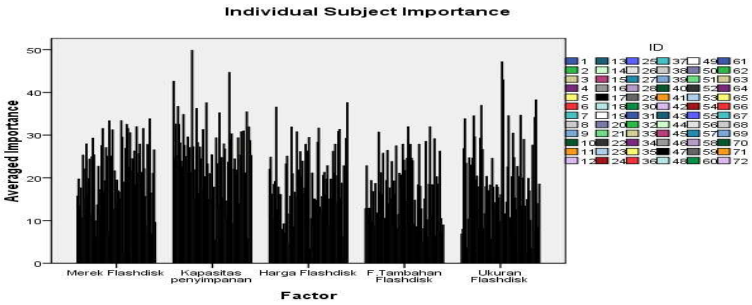
Averaged Importance Score

Nilai *importance* di atas dapat disajikan pada gambar berikut :



Gambar 15. Nilai *Importance* Atribut *Flash disk*

Sedangkan grafik nilai *importance* untuk 72 responden gambar berikut:



Gambar 16. Distribusi Penilaian Setiap Responden Terhadap Faktor

Berdasarkan tabel korelasi dapat diketahui hubungan (korelasi) antara data responden dengan data sebenarnya yang bertujuan untuk mengukur ketepatan/kecocokan estimasi model. Output untuk nilai korelasi adalah sebagai berikut:

Tabel 4. Nilai Correlation Responden Terhadap Atribut *Flash Disk*

Correlations[a]

	Value	Sig.
Pearson's R	.512	.021
Kendall's tau	.310	.048
Kendall's tau for Holdouts	1.000	.

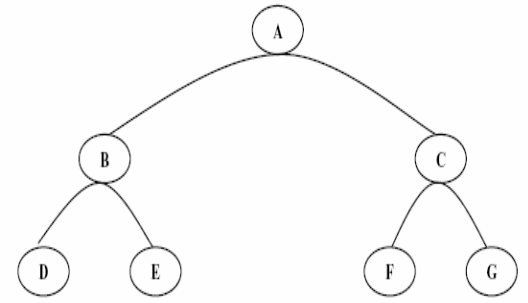
a. Correlations between observed and estimated preferences

Pada tabel korelasi angka signifikan untuk uji Pearson's R dan Kendall's tau dibawah **0,05** maka kedua uji tersebut berada pada taraf signifikan, maka Ho ditolak. Hal ini berarti memang ada korelasi yang nyata antara hasil konjoin dengan pendapat responden tersebut. Dengan demikian bahwa pendapat 72 responden tersebut bisa diterima untuk menggambarkan keinginan populasi pembeli *flash disk*.

Berdasarkan hasil analisis konjoin rata-rata responden menganggap kapasitas lebih penting dibanding fitur tambahan dan atribut lainnya. Diketahui responden lebih senang *flash disk* dengan merek kingstone, harga murah, memiliki kapasitas 1G, dengan fitur tambahan MP3, dan mempunyai ukuran yang kecil sehingga mudah dibawah kemana-mana.

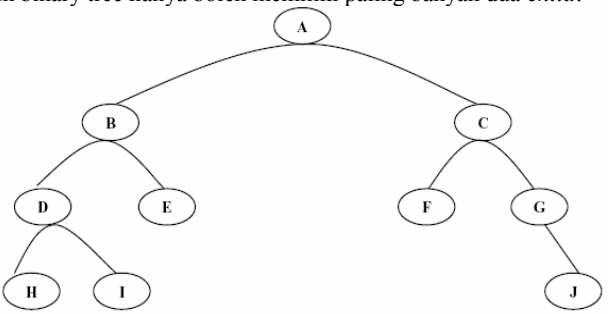
Regresi Pohon

Pohon atau tree adalah kumpulan node yang saling terhubung satu sama lain dalam suatu kesatuan yang membentuk struktur sebuah pohon. Struktur pohon adalah suatu cara merepresentasikan suatu struktur hirarki (one-to-many) yang secara grafis mirip sebuah pohon (Rachmat, 2007). Node dalam sebuah tree antara lain *root*, *leaf* dan *internal node* dimana node root dalam sebuah tree adalah suatu node yang memiliki hirarki tertinggi, leaf adalah node yang tidak memiliki cabang (sering juga disebut *terminal node*), sedangkan internal node adalah node dalam yang bukan merupakan leaf. Node-node lain di bawah node root yang saling terhubung satu sama lain disebut dengan *subtree*.



Gambar 1. Contoh Tree
(*Sumber : Handout Struktur Data, Rachmat. 2007)*

Tree terdiri atas berbagai jenis, salah satu diantaranya adalah *binary tree*. Binary tree merupakan suatu tree dengan syarat bahwa setiap node hanya memiliki maksimal dua subtree dan kedua subtree tersebut harus terpisah. Tiap node dalam binary tree hanya boleh memiliki paling banyak dua *child*.



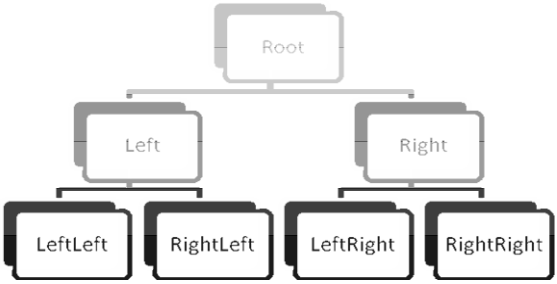
Gambar 2. Contoh Binary Tree
(*Sumber : Handout Struktur Data, Rachmat, 2007)*

Model berdasarkan pohon (*tree-based model*) adalah teknik analisis data eksplorasi yang dapat digunakan untuk :

a. Membangun dan mengevaluasi model prediksi multivariat.
b. Meringkas himpunan data multivariat yang besar.
c. Menaksir kecukupan model linier.
d. Menemukan kaidah prediksi yang dapat dievaluasi dengan cepat dan berulang kali.

Model berdasarkan pohon digunakan untuk masalah klasifikasi dan regresi. Pada suatu data pengamatan, jenis data untuk variabel prediktor X_i berupa data kualitatif, sedangkan untuk variabel respon Y berupa data kuantitatif atau kualitatif (numerik atau kategorik).

Dalam model tree-based, *root* merupakan node teratas sedangkan *leaf* adalah terminal node dari suatu tree. *Root* atau *parent* memiliki dua anak (*child*) yang terdiri dari anak kiri dan anak kanan. Anak kiri dan kanan tersebut akan dibagi kembali menjadi anak-anak berikutnya. Proses ini berlangsung terus hingga dicapai terminal node (*leaf*).



Gambar 3. Ilustrasi Tree dengan anak kiri dan anak kanan

Pengenalan Pohon Keputusan

Decision tree (pohon keputusan) adalah cara merepresentasikan kumpulan aturan yang mengacu ke suatu nilai atau kelas. Pohon keputusan digunakan dalam *data mining* dan *machine learning*. Data mining adalah proses yang menggunakan berbagai perangkat analisis data untuk menemukan pola dan hubungan dalam data yang digunakan untuk membuat prediksi yang valid.

Dalam data mining, pohon mempunyai 3 kategori (Anonim, 2007b) :

1. *Classification Tree Analysis*, merupakan analisis data yang memprediksi data ke dalam suatu kelas-kelas tertentu.
2. *Regression Tree Analysis*, merupakan analisis data yang memprediksi data ke dalam suatu bilangan riil (misalnya harga sebuah rumah atau lamanya pasien dirawat di rumah sakit).
3. *Classification and Regression Tree (CART) Analysis*, merupakan analisis data yang digunakan untuk mengarahkan kedua prosedur di atas. Analisis ini pertama kali diperkenalkan oleh Breiman.

Pohon klasifikasi dan regresi adalah kumpulan dari banyak kaidah yang ditunjukkan oleh bentuk pohon biner (*binary tree*). Kaidah tersebut ditentukan oleh suatu prosedur yang dikenal dengan *recursive partitioning* (penyekatan berulang).

Misalkan data awal berbentuk : $(X, Y) = (X_1, X_2, X_3, \ldots, X_n, Y)$ dimana $X_i, i = 1, 2, \ldots, n$ merupakan himpunan variabel prediktor yang datanya dapat berbentuk kategorik maupun numerik. Sedangkan *Y* merupakan variabel respon yang ingin dibuat pohon klasifikasi atau pohon regresinya.

Pohon klasifikasi dan pohon regresi menjadi semakin populer untuk mempartisi data dan mengidentifikasi struktur lokal pada dataset besar dan kecil. Pada pohon klasifikasi variabel responnya adalah kategorik, sebaliknya pohon regresi variabel responnya adalah kuantitatif (interval atau rasio).

Classification and Regression Trees (CART) Analysis

Classification and Regression Trees adalah metode klasifikasi menggunakan data historis untuk membangun suatu pohon keputusan. Metodologi CART mulai dikembangkan pada tahun 80-an oleh Breiman, Friedman, Olshen, dan Stone dalam makalah mereka yang berjudul "*Classification and Regression Trees*" (1984).

CART adalah suatu analisis diskriminan non-parametrik yang dirancang untuk menyajikan kaidah keputusan berbentuk pohon biner yang membagi data pada learning sampel dalam batasan linier univariat. Analisis ini menghasilkan kelompok data hirarkis yang dimulai dari *node root* untuk keseluruhan learning sampel dan berakhir pada kelompok kecil pengamatan yang homogen. Pada setiap *terminal node* diberikan label kelas atau nilai yang diramalkan, sehingga menghasilkan struktur pohon yang dapat ditafsirkan sebagai pohon keputusan.

Pohon keputusan ditunjukkan oleh suatu himpunan pertanyaan yang membagi learning sampel ke dalam bagian yang lebih kecil lagi. CART akan mencari semua variabel yang mungkin dan semua nilai yang mungkin dalam rangka menemukan pembagian yang terbaik. Pertanyaan kemudian membagi data ke dalam dua bagian dengan homogenitas maksimum. Proses tersebut kemudian diulangi untuk masing-masing pecahan data hasil (Timofeev, 2004).

Menurut Andriyashin (2005), keuntungan dari penggunaan analisis CART adalah sebagai berikut :

1. Merupakan bentuk statistika non-parametrik, sehingga tidak memerlukan asumsi sebaran dan uji hipotesis.
2. Tidak memerlukan variabel untuk dipilih sebelumnya.
3. Sangat efisien dalam terminologi perhitungan.
4. Dapat menangani dataset dengan struktur yang kompleks.
5. Sangat tangguh dalam menangani outlier, umumnya algoritma pemisahan akan mengisolasikan outlier pada individu node atau beberapa node.
6. Dapat menggunakan sembarang kombinasi data kontinu/numerik dan kategorik.

7. Hasilnya invarian dengan transformasi monoton dari variabel respon, artinya penggantian sembarang variabel dengan algoritmanya atau nilai akar kuadrat, tidak akan menyebabkan struktur pohon berubah.

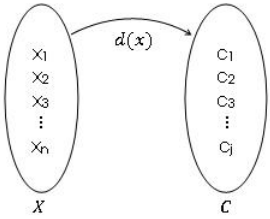
Pohon Klasifikasi

Classifier atau kaidah klasifikasi adalah suatu cara sistematik dalam memprediksi suatu kasus masuk dalam kelas tertentu. Untuk memberikan formulasi yang lebih tepat, maka disusun suatu himpunan pengukuran $\{x_1, x_2, \dots, x_n\}$ sebagai faktor pengukuran (*measurement vector*). Semua vektor pengukuran yang mungkin berada di dalamnya didefinisikan sebagai ruang pengukuran X.

Andaikan suatu kasus atau objek mempunyai *J* kelas yaitu $1, 2, \dots, j$ dan misalkan C adalah himpunan kelas tersebut dengan $C = \{1, 2, \dots, j\}$. Suatu cara sistematik dalam memprediksi anggota kelas tersebut adalah dengan menggunakan suatu aturan yang menempatkan anggota kelas dalam C tersebut pada setiap vektor pengukuran x dalam X.

Definisi 1.

"Suatu *classifier* atau aturan klasifikasi adalah suatu fungsi $d(x)$ pada X sehingga untuk setiap x, $d(x)$ adalah sama dengan salah satu dari $\{1, 2, \dots, j\}$."



Gambar 4. Definisi *Classifier* pada suatu fungsi $d(x)$

Cara lain untuk melihat *classifier* adalah dengan mendefinisikan A_j sebagai subset dari X dimana $d(x)$ sama dengan *j* sehingga $A_j = \{x ; d(x) = j\}$.

Himpunan $A_1, \dots, A_j$ adalah disjoint dan

$$X = \bigcup_j A_j$$

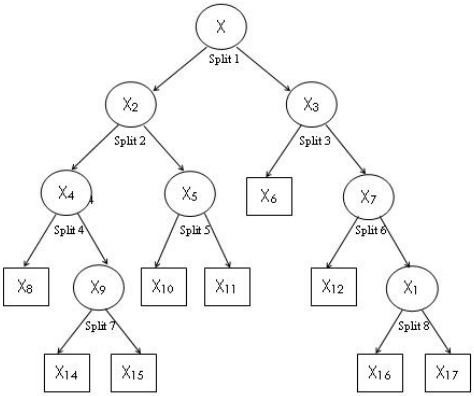
sehingga A_j adalah bentuk partisi dari X.

Definisi 2.

"*Classifier* adalah suatu partisi pada X dalam J yang memisahkan himpunan bagian $A_1, \ldots, A_j \ni X = \bigcup A_j$. Sehingga untuk setiap $x \in A_j$ kelas prediksinya adalah *j*."

Struktur klasifikasi pohon biner dibangun dengan cara pemisahan berulang oleh himpunan bagian X ke dalam dua keturunan (node anak), dimulai dengan X itu sendiri.

Sebagai ilustrasi dapat dilihat pada gambar di bawah ini :



Gambar 5. Struktur Klasifikasi Pohon Biner

keterangan :

□ :terminal node ○ : node lainnya

↙↘ : pemisahan berulang menjadi 2 node anak

Pada gambar 5 tersebut, X_2 dan X_3 adalah disjoint dengan $X = X_2 \cup X_3$. Sama halnya dengan X_4 dan X_5 dimana $X_4 \cup X_5 = X_2$ dan $X_6 \cup X_7 = X_3$. Himpunan bagian yang tidak dipisahkan yaitu : $X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16}$, dan X_{17} disebut dengan *terminal node*.

Masing-masing terminal node merupakan bentuk partisi pada X dan ditunjukkan oleh suatu label kelas. Mungkin saja terdapat dua atau lebih terminal node dengan label kelas yang sama. Partisi yang berhubungan dengan *classifier* didapatkan dengan meletakkan semua terminal node pada kelas yang sama dalam pohon klasifikasi tersebut.

Dalam konstruksi klasifikasi sistematik, semua data historis dirangkum dalam suatu *learning sample*. Learning sampel merupakan sampel data yang digunakan untuk membangun pohon klasifikasi.

		Node						
		1	2	⋯	s	⋯	t	
Kelas	1				$N_1(s)$			
	2				$N_2(s)$			
	⋯				⋮			
	i	$N_j(1)$	$N_j(2)$	⋯	$N_j(s)$	⋯	$N_j(t)$	N_j
	⋯				⋮			
	k				$N_k(s)$			
					$N(s)$			N

Gambar 6. Ilustrasi Perhitungan Rumus Pohon Klasifikasi

keterangan :

C : Himpunan kelas pada learning sampel dari $1, 2, \dots, j, \dots, k$.

T : Himpunan node pada learning sampel dari $1, 2, \dots, s, \dots, t$.

N : jumlah keseluruhan pengamatan.

N_j : jumlah pengamatan yang berada pada kelas ke-*j*.

$N(s)$: jumlah pengamatan pada node *s*.

$N_j(t)$: pengamatan pada node *t* kelas ke-*j*.

$N_k(s)$: pengamatan pada node *s* kelas ke-*k*.

Dari gambar di atas, didefinisikan rumus-rumus sebagai berikut :

1. Proporsi pengamatan pada kelas ke-*j* terhadap jumlah keseluruhan pengamatan.

$$\pi(j) = \frac{N_j}{N} \tag{1}$$

2. Jumlah pengamatan pada kelas ke-*j*.

$$N_j = \sum_{s=1}^{t} N_j(s) \tag{2}$$

3. Jumlah pengamatan pada node *s*.

$$N(s) = \sum_{j=1}^{k} N_j(s) \tag{3}$$

4. Peluang pengamatan pada node *s*.

$$p(s) = \frac{N(s)}{N} \tag{4}$$

5. Peluang bersama pengamatan pada node *s* kelas ke-*j*.

$$p(j,s) = \frac{N_j(s)}{N} \tag{5}$$

6. Peluang bersyarat pengamatan pada node *s* kelas ke-*j*.

$$p(j|s) = \frac{p(j,s)}{p(s)} = \frac{\frac{N_j(s)}{N}}{\frac{N(s)}{N}} = \frac{N_j(s)}{N(s)} \tag{6}$$

Dari persamaan (3) dan (6) diperoleh rumus berikut :

$$p(1|s) + p(2|s) + p(3|s) + \ldots + p(k|s) = \sum_{j=1}^{k} p(j|s) = \sum_{j=1}^{k} \frac{N_j(s)}{N(s)}$$

$$= \frac{1}{N(s)} . N(s) = 1$$

Suatu ukuran impurity pada node *t* disimbolkan dengan $i(t)$, dimana $i(t)$ merupakan suatu fungsi peluang kelas $p(1|t),\ p(2|t),\ldots,p(k|t)$. Sehingga secara matematis dapat dituliskan dengan :

$$i(t) = f\,[p(1|t), p(2|t), \ldots, p(k|t)] \tag{7}$$

Definisi 3.

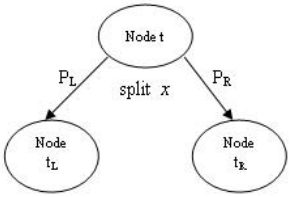
"*Impurity function* (fungsi impurity) adalah suatu fungsi *f* yang didefinisikan pada himpunan $(p_1, p_2, \ldots, p_k)$ yang memenuhi $p_j \geq 0,\ j = 1, \ldots, k,\ \sum_j p_j = 1$ dengan kriteria sebagai berikut :

1. *f* akan *maksimum unik* pada titik $\left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$. Dengan kata lain, masing-masing kelas dalam populasi memiliki peluang yang sama.
2. *f* akan *minimum unik* pada titik $(1,0,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,0,0,\ldots,1)$.
3. *f* adalah fungsi simetrik dari $p_1, p_2, \ldots, p_k$."

Misalkan jumlah terminal node pada pohon klasifikasi adalah $\tilde{T}$ dan diketahui himpunan $I(t) = i(t).p(t)$. Rumus impurity pohon (*tree impurity*) didefinisikan dengan :

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t).p(t) \tag{8}$$

Suatu pohon klasifikasi dibangun berdasarkan aturan pemisahan (*splitting rule*), yaitu aturan yang memisahkan learning sampel ke dalam bagian yang lebih kecil. Setiap kali data yang ada harus dibagi menjadi dua bagian dengan homogenitas maksimum.



Gambar 7. Algoritma pemisahan pada CART

keterangan :

t : node ayah
t_L : node anak kiri
t_R : node anak kanan
p_L : peluang node kiri
p_R : peluang node kanan

Anggap t adalah node ayah yang dipisahkan oleh pembagian x menjadi dua node anak, yaitu node anak kiri t_L dan node anak kanan t_R. Masing-masing node anak tersebut mempunyai peluang, p_L dan p_R dengan $p_R = 1 - p_L$. Pada sembarang terminal node, akan dipilih pembagian yang paling mengurangi nilai $I(t)$ dengan kata lain akan akuivalen dengan memaksimalkan perubahan fungsi impurity node *t* pada *x* sebagai berikut :

$$\Delta I(x,t) = I(t) - I(t_L) - I(t_R)$$

atau

$$\Delta i(x,t) = i(t) - p_L \,.\, i(t_L) - p_R \,.\, i(t_R) \qquad (9)$$

Nilai pemisahan terbaik $\Delta i(x,t)$ menunjukkan perubahan impurity node *t* pada *x* dengan $t_L \cup t_R = 1$. Oleh karena itu, pemisahan terbaik dari *t* adalah :

$$\arg\max_x \{\Delta\, i(x,t)\} = \arg\max_x \left\{i(t) - p_L\, i(t_L) - p_R\, i(t_R)\right\} \qquad (10)$$

Nilai optimal x^* dapat ditentukan dengan cara memaksimalkan $\Delta i(x,t)$ dengan *x* yang berbeda pada masing-masing node *t*. Prosedur semacam

ini memungkinkan untuk membangun pohon keputusan dari sembarang bentuk pohon maksimum.

Karena nilai $i(t)$ pada kenyataannya adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$x^* = \underset{x}{\operatorname{argmax}}\ \Delta\, i(x,t) = \underset{x}{\operatorname{argmax}}\ \left\{- p_L\, i(t_L) - p_R\, i(t_R)\right\}$$
$$= \underset{x}{\operatorname{argmin}}\ \left\{ p_L\, i(t_L) + p_R\, i(t_R)\right\} \qquad (11)$$

dimana t_L dan t_R adalah fungsi eksplisit dari *x*.

Selanjutnya akan dibahas cara untuk menggambarkan fungsi impurity $i(t)$ tersebut. Di dalam teori ada beberapa fungsi impurity, tetapi hanya dua yang secara luas digunakan dalam praktek, yaitu Kaidah Pemisahan Gini (*Gini Splitting Rule*) dan Kaidah Pemisahan Twoing (*Twoing Splitting Rule*).

Gini Splitting Rule

Gini Splitting Rule atau disebut juga indeks Gini (*Gini Index*) adalah kaidah yang paling umum digunakan dalam memecahkan permasalahan pohon klasifikasi. Data impurity didefinisikan dengan menggunakan ukuran varian (*variance measure*). Misalkan 1 adalah semua pengamatan pada node *t* kelas ke-*j* dan 0 untuk yang lainnya. Kemudian estimasi varian contoh untuk node *t* pengamatan sebagai berikut :

$$p(j|t)\big(1 - p(j|t)\big)$$

Indeks Gini pada node *t* kelas ke-*j* didefinisikan dengan rumus sebagai berikut :

$$i(t) = \sum_{j=1}^{k} p(j|t)\left(1 - p(j|t)\right) = \sum_{j=1}^{k} p(j|t) - p^2(j|t)$$
$$= \sum_{j=1}^{k} p(j|t) - \sum_{j=1}^{k} p^{\ 2}(j|t)$$
$$= 1 - \sum_{j=1}^{k} p^2(j|t) \qquad (12)$$

Sehingga diperoleh perubahan fungsi impurity node *t* oleh pemisahan *x* sebagai berikut :

$$\Delta i(x,t) = i(t) - p_L\, .i(t_L) - p_R\, .i(t_R)$$
$$= 1 - \sum_{j=1}^{k} p^2(j|t) - p_L \cdot \left(1 - \sum_{j=1}^{k} p^2(j|t_L)\right) - p_R \cdot \left(1 - \sum_{j=1}^{k} p^2(j|t_R)\right)$$
$$= 1 - \sum_{j=1}^{k} p^2(j|t) - p_L + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) - p_R + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R)$$
$$= 1 - \sum_{j=1}^{k} p^2(j|t) - p_L + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) - p_R + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R)$$

$$= 1 - p_L - p_R - \sum_{j=1}^{k} p^2(j|t) + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R)$$

$$= -\sum_{j=1}^{k} p^2(j|t) + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R) \quad (13)$$

Sehingga didapatkan pemisahan terbaik dari *t* sebagai berikut :

$$\underset{x}{\operatorname{argmax}}\, \Delta\, i(x,t) = \underset{x}{\operatorname{argmax}} \left\{ -\sum_{j=1}^{k} p^2(j|t) + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R) \right\} \quad (14)$$

Indeks Gini akan mencari learning sampel untuk kelas paling besar dan mengisolasinya dari sisa data tersebut. Kaidah ini bekerja dengan baik pada data berukuran besar.

Peluang bersyarat pengamatan pada node *t* kelas ke-*j* juga dapat dihitung sebagai berikut :

$$\begin{aligned} p(j|t) &= \frac{p(j,t)}{p(t)} \\ &= \frac{p(j,t_L) + p(j,t_R)}{p(t)} \\ &= \frac{p(j,t_L)}{p(t)} + \frac{p(j,t_R)}{p(t)} \\ &= \frac{p(t_L)}{p(t)} \cdot \frac{p(j,t_L)}{p(t_L)} + \frac{p(t_R)}{p(t)} \cdot \frac{p(j,t_R)}{p(t_R)} \\ &= p_L \cdot p(j|t_L) + p_L \cdot p(j|t_R) \end{aligned} \quad (15)$$

Indeks Gini juga dapat dianggap sebagai suatu fungsi $f(p_1, p_2, \ldots, p_k)$ yang dapat diubah menjadi polinomial derajat-dua dengan koefisien tidak negatif. Untuk masing-masing fungsi konveks dengan syarat $\forall \alpha \geq 0$ akan memenuhi pertidaksamaan :

$$f(\alpha.p_1 + (1-\alpha)\, p_1', \alpha.p_2 + (1-\alpha)\, p_2', \ldots, \alpha.p_k + (1-\alpha)\, p_k') > \alpha\, f(p_1, \ldots, p_k) + (1-\alpha)\, f(p_1', \ldots, p_k') \quad (16)$$

Dengan menggunakan persamaan (7) dan (12) akan dibuktikan :

$$p_L\, i(t_L) + p_R\, i(t_R) = p_L \cdot f\left(p(1|t_L), \ldots, p(k|t_L)\right) + p_R \cdot f\left(p(1|t_R), \ldots, p(k|t_R)\right) \leq f\left(p_L\, p(1|t_L) + p_R\, p(1|t_R), \ldots, p_L\, p(k|t_L) + p_R\, p(k|t_R)\right) \quad (17)$$

Bukti :

Perhatikan ruas kiri dari pertidaksamaan (16) :

$$f(\alpha.p_1 + (1-\alpha)\, p_1', \alpha.p_2 + (1-\alpha)\, p_2', \ldots, \alpha.p_k + (1-\alpha)\, p_k')$$

$$= f\left((\alpha.p_1, \alpha.p_2, \dots, \alpha.p_k) + \left((1-\alpha)\,p'_1, (1-\alpha)\,p'_2, \dots, (1-\alpha)\,p'_k\right)\right)$$

$$= f(\alpha.p_1, \alpha.p_2, \dots, \alpha.p_k) + f\left((1-\alpha)\,p'_1, (1-\alpha)\,p'_2, \dots, (1-\alpha)\,p'_k\right)$$

$$= \left[1 - \alpha \sum_{j=1}^{k} {p_C}^2\right] + \left[1 - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2\right]$$

$$= 2 - \alpha \sum_{j=1}^{k} {p_C}^2 - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2 \qquad *)$$

Perhatikan kembali ruas kanan dari pertidaksamaan (16) :

$$\alpha\, f(p_1, \dots, p_k) + (1-\alpha)\, f(p'_1, \dots, p'_k)$$

$$= \left[\alpha\left(1 - \sum_{j=1}^{k} {p_C}^2\right)\right] + \left[(1-\alpha)\left(1 - \sum_{j=1}^{k} {p'_C}^2\right)\right]$$

$$= \left[\alpha - \alpha \sum_{j=1}^{k} {p_C}^2\right] + \left[(1-\alpha) - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2\right]$$

$$= 1 - \alpha \sum_{j=1}^{k} {p_C}^2 - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2 \qquad **)$$

Dari *) dan **), diperoleh fungsi konveks berikut :

$$2 - \alpha \sum_{j=1}^{k} {p_C}^2 - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2 > 1 - \alpha \sum_{j=1}^{k} {p_C}^2 - (1-\alpha) \sum_{j=1}^{k} {p'_C}^2$$

Misalkan $\alpha = p_L$ dan $(1-\alpha) = p_R$, sehingga pertidaksamaan menjadi :

$$f\left(p_L\, p(1|t_L) + p_R\, p(1|t_R), \dots, p_L\, p(k|t_L) + p_R\, p(k|t_R)\right) > p_L \,.f\left(p(1|t_L), \dots, p(k|t_L)\right) + p_R \,.f\left(p(1|t_R), \dots, p(k|t_R)\right)$$

atau

$$p_L \,.f\left(p(1|t_L), \dots, p(k|t_L)\right) + p_R \,.f\left(p(1|t_R), \dots, p(k|t_R)\right) \leq f\left(p_L\, p(1|t_L) + p_R\, p(1|t_R), \dots, p_L\, p(k|t_L) + p_R\, p(k|t_R)\right)$$

Pertidaksamaan ini akan berubah menjadi suatu persamaan dengan syarat $\forall\, c \in C$ berlaku $p(C|t_L) = p(C|t_R)$.

Bukti :

$$p_L \,.f\left(p(1|t_L), \dots, p(k|t_L)\right) + p_R \,.f\left(p(1|t_R), \dots, p(k|t_R)\right) = p_L \,.f\left(p(C|t_L)\right) + p_R \,.f\left(p(C|t_R)\right)$$

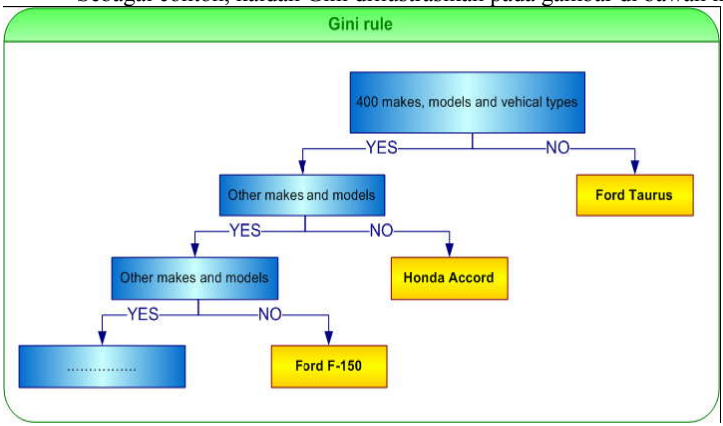
$$f\left(p_L\, p(1|t_L) + p_R\, p(1|t_R), \dots, p_L\, p(k|t_L) + p_R\, p(k|t_R)\right)$$

$$= f\left(p_L\, p(1|t_L), \ldots, p_L\, p(k|t_L) + p_R\, p(1|t_R), \ldots, p_R\, p(k|t_R)\right)$$
$$= f\left(p_L\left(p(1|t_L), \ldots, p(k|t_L)\right) + p_R\left(p(1|t_R), \ldots, p(k|t_R)\right)\right)$$
$$= f\left(p_L\left(p(C|t_L)\right) + p_R\left(p(C|t_R)\right)\right)$$
$$= p_L \cdot f\left(p(C|t_L)\right) + p_R \cdot f\left(p(C|t_R)\right)$$

Sehingga terbukti :

$$p_L\, i(t_L) + p_R\, i(t_R) = p_L \cdot f\left(p(1|t_L), \ldots, p(k|t_L)\right) + p_R \cdot f\left(p(1|t_R), \ldots, p(k|t_R)\right)$$
$$= f\left(p_L\, p(1|t_L) + p_R\, p(1|t_R), \ldots, p_L\, p(k|t_L) + p_R\, p(k|t_R)\right) \quad (18)$$

Sebagai contoh, kaidah Gini diilustrasikan pada gambar di bawah ini :



Gambar 8. Pohon Klasifikasi yang dibangun menggunakan Kaidah Gini (*Sumber : Financial Application of Classification and Regression Trees, Andriyashin. 2005)*

Twoing Splitting Rule

Tidak seperti kaidah *Gini*, kaidah *Twoing* tidak mencari nilai maksimal dari ukuran impurity. Sebagai gantinya kaidah ini mencoba untuk menyeimbangkan konstruksi pohon dengan cara seolah-olah membagi learning sampel menjadi dua kelas. Sehingga pengamatan dapat dibedakan antara faktor umum yang berada pada tingkat teratas dan karakteristik khusus yang berada pada tingkat yang lebih rendah.

Menurut Andiyashin (2005), misalkan terdapat himpunan kelas learning sampel $C = \{1, 2, \ldots, k\}$.Himpunan tersebut dibagi menjadi dua bagian yaitu: $C_1 = \{c_1, c_2, \ldots, c_n\}$ dan $C_2 = C \setminus C_1$ sedemikian sehingga semua pengamatan yang berada pada C_1 mempunyai kelas dummy 1, sedangkan sisanya mempunyai kelas dummy 2.

Kemudian akan dihitung nilai $\Delta\, i(x,t)$ untuk *x* yang berbeda "jika hanya ada dua kelas dummy". Karena nilai $\Delta\, i(x,t)$ bergantung pada C_1, maka nilai $\Delta\, i(x,t,C_1)$ adalah maksimal. Dengan kata lain, kaidah Twoing adalah suatu aturan yang digunakan untuk menemukan kombinasi superkelas pada setiap node seolah-olah kenaikan impurity telah dimaksimalkan hanya oleh dua kelas $C = \{1,2\}$.

Walaupun kaidah Twoing dapat diterapkan terutama untuk data dengan jumlah kelas yang besar, kelemahannya terdapat pada kecepatan perhitungan. Asumsikan bahwa learning sampel mempunyai *J* kelas, kemudian himpunan C dipisahkan menjadi C_1 Dan C_2 dengan 2^{J-1} cara. Pada kasus dimana terdapat 11 data kelas pada learning sampel, maka akan terbentuk 1024 kombinasi.

Seperti yang telah disebutkan sebelumnya, kaidah Twoing merupakan kaidah pemisahan yang tidak tergantung pada ukuran impurity $i(t)$. Menurut Sezgin (2006), pada kaidah twoing, tidak ada ukuran impurity yang spesifik. Sehingga untuk sembarang node, pemisahan yang terbaik ditentukan dengan cara memaksimalkan perubahan impurity pada node anak kanan t_R dan node anak kiri t_L. Ini mengakibatkan timbulnya perbedaan definisi kaidah twoing oleh para peneliti, antara lain :

1. Chee Jen Chang (2002)

$$i(t) = \frac{P_L.P_R}{4}\left(\sum_j (|p(j|t_L) - p(j|t_R)|)\right)^2 \tag{19}$$

2. David Feldman (2003)

$$d_T(t) = \frac{P_L.P_R}{4}\sum_{j=1}^{J}|p(j|t_L) - p(j|t_R)| \tag{20}$$

3. Roman Timofeev (2004)

$$\Delta i(t) = \frac{P_L.P_R}{4}\left[\sum_{k=1}^{k}|p(k|t_L) - p(k|t_R)|\right]^2 \tag{21}$$

4. Anton Andriyashin (2005)

$$S_1(s) = \{j : p(j|t_L) \geq p(j|t_R)\}$$

$$\max_{S_1} \Delta\, i(s,t,S_1) = \frac{p_L \cdot p_R}{4}\left[\sum_{j=1}^{J} |p(j|t_L) - p(j|t_R)|\right]^2 \quad (22)$$

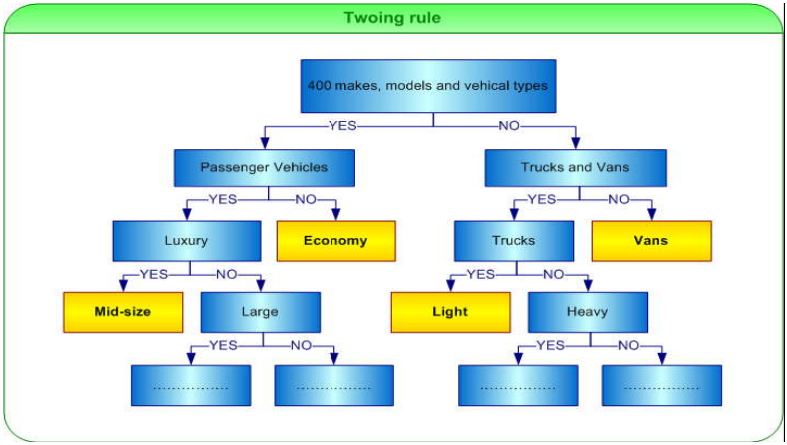
5. Ozge Sezgin (2006)

$$\Delta i(t) = \frac{P_L \cdot P_R}{4}\left[\sum_{j=1}^{J} p(j|t_L)\, p(j|t_R)\right]^2 \quad (23)$$

6. Zambon, et al (2006)

$$Twoing(t) = \frac{P_L \cdot P_R}{4}\left(\sum_{i} (|p(i|t_L) - p(i|t_R)|)\right)^2 \quad (24)$$

Sebagai contoh, kaidah Twoing diilustrasikan pada gambar di bawah ini :



Gambar 9. Pohon Klasifikasi yang dibangun menggunakan Kaidah Twoing
(*Sumber : Financial Application of Classification and Regression Trees, Andriyashin. 2005)*

Pohon Regresi

Konstruksi pohon pada pohon klasifikasi dan pohon regresi tidak jauh berbeda, yang membedakannya adalah jenis variabel responnya. Pohon regresi adalah jenis pohon keputusan dengan variabel respon kontinu. Tujuan utama pohon regresi adalah untuk menghasilkan struktur pohon prediktor atau kaidah prediksi (Breiman et al, 1984). Prediktor ini menjalankan dua tujuan utama, yaitu : (1) untuk memprediksi ketelitian variabel respon atau nilai baru dari variabel

prediktor, (2) untuk menjelaskan hubungan antara variabel respon dan variabel prediktor.

Indeks Gini dan Kaidah Twoing yang dibahas sebelumnya berasumsi bahwa banyaknya kelas adalah terbatas sehingga dibuat beberapa ukuran berdasarkan $p(C|t)$. Akan tetapi, jika variabel responnya adalah kontinu dan tidak ada kelas, pendekatan ini tidak dapat digunakan lagi kecuali jika kelompok nilai-nilai kontinu tersebut diganti dengan kelas dummy.

Prediktor pada pohon regresi dibangun dengan mendeteksi heterogenitas (dalam kaitan dengan varian pada variabel prediktor) yang ada di dalam data tersebut. Pohon regresi melakukannya dengan penyekatan berulang data ke dalam kelompok atau terminal node yang secara internal lebih homogen dibandingkan node di atasnya. Pada masing-masing terminal node, nilai rata-rata dari variabel respon dianggap sebagai nilai prediksi.

Terdapat dua kaidah pemisahan atau fungsi impurity pada pohon regresi, yaitu *Least Square (LS) function dan the Least Absolute Deviation (LAD) function*. Karena mekanisme kedua aturan tersebut sama, maka disini hanya akan dibahas ukuran impurity LS. Menggunakan kriteria ini, impurity node diukur dengan jumlah kuadrat node, $SS(t)$ yang didefinisikan sebagai berikut :

$$SS(t) = \sum_{i=1}^{N(t)} \left(y_{i(t)} - \bar{y}_{(t)}\right)^2 \tag{25}$$

keterangan :

$SS(t)$: fungsi impurity.

$y_{i(t)}$: nilai individu dari variabel respon pada node *t*.

$\bar{y}_{(t)}$: nilai rata-rata variabel respon pada node *t*.

$N(t)$: jumlah pengamatan pada node *t*.

Misalkan diberikan fungsi impurity, $SS(t)$ yang dipisahkan oleh *x* pada node kiri t_L dan node kanan t_R. Sehingga pemisahan terbaik (*goodness of split*) dapat ditunjukkan dengan rumus sebagai berikut :

$$\underset{x}{\operatorname{argmax}} \{f(x,t)\} = \underset{x}{\operatorname{argmax}} \{SS(t) - SS(t_R) - SS(t_L)\} \tag{26}$$

Alternatif lain dari fungsi impurity pada pohon regresi adalah dengan menggunakan nilai varian dari node anak kiri dan node anak kanan. Nilai varian pada suatu node *t* didefinisikan dengan rumus berikut :

$$x^2(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} \left[y_{i(t)} - \bar{y}_{(t)}\right]^2 \tag{27}$$

Sehingga diperoleh perubahan fungsi impurity node *t* oleh pemisahan *x* berikut :

$$\Delta x^2(t) = x^2(t) - p_L . x^2(t_L) - p_R . x^2(t_R) \tag{28}$$

keterangan :

$x^2(t_L)$: varian variabel prediktor pada node anak kiri.

$x^2(t_R)$: varian variabel prediktor pada node anak kanan.

Nilai perubahan impurity $\Delta x^2(t)$ menunjukkan nilai pemisahan terbaik node *t* oleh pemisahan *x*, yang dirumuskan dengan :

$$\underset{x}{\text{argmax}}\left\{\Delta x^2(t)\right\} = \underset{x}{\text{argmax}}\left\{x^2(t) - p_L.x^2(t_L) - p_R.x^2(t_R)\right\} \quad (29)$$

Sama halnya dengan pohon klasifikasi, nilai optimal x^* dapat ditentukan dengan cara memaksimalkan $\Delta x^2(t)$ dengan *x* yang berbeda pada masing-masing node *t*. Karena nilai $x^2(t)$ adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$\begin{aligned} x^* &= \underset{x}{\text{argmax}}\left\{\Delta x^2(t)\right\} = \underset{x}{\text{argmax}}\left\{-p_L.x^2(t_L) - p_R.x^2(t_R)\right\} \\ &= \underset{x}{\text{argmin}}\left\{p_L.x^2(t_L) + p_R.x^2(t_R)\right\} \end{aligned} \quad (30)$$

Sehingga diperoleh kesimpulan bahwa pemisahan yang terbaik adalah nilai $\Delta x^2(t)$ yang paling tinggi dengan jumlah varian $[p_L.x^2(t_L) + p_L.x^2(t_R)]$ yang paling rendah. Prosedur ini menghasilkan node anak kiri dan node anak kanan yang lebih homogen dibandingkan node ayahnya. Masing-masing bagian membagi pengamatan pada node anak kiri dan node anak kanan sehingga nilai rata-rata dari variabel prediktor pada suatu terminal node adalah lebih rendah dibandingkan nilai rata-rata node ayah.

Langkah-langkah pembuatan pohon klasifikasi

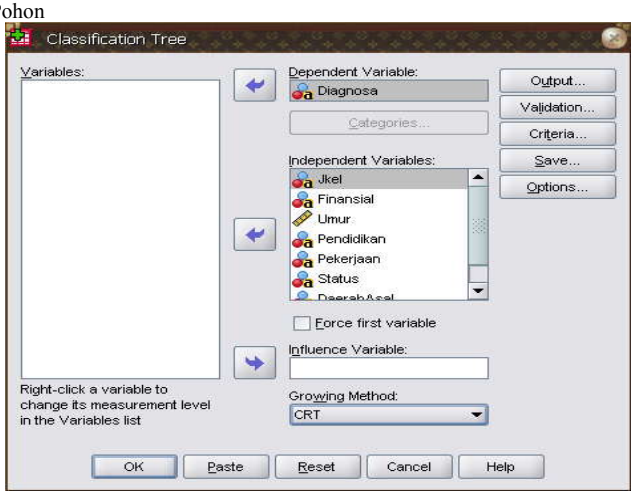
1. Sebagai langkah pertama, buka data hasil penelitian dengan menggunakan software SPSS versi 16.0. Tingkatan pengukuran (*measurement level*) dari data yang digunakan dapat berbentuk salah satu diantara :


Skalar


Nominal

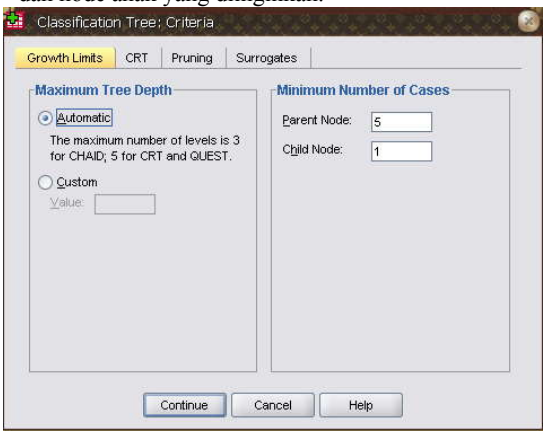

Ordinal

2. Untuk membuat pohon klasifikasi, pada menu utama pilih : Analyze – Classify – Tree. Selanjutnya pilih *dependent variable* (variable respon) dan *independent variables* (variabel prediktor) yang digunakan. Kemudian tentukan metode pertumbuhan pohon (*Growing Methods*). Terdapat empat metode pertumbuhan pohon, yaitu CHAID, Exhaustive CHAID, CRT, dan QUEST. Dalam analisis data penelitian ini digunakan metode CRT.



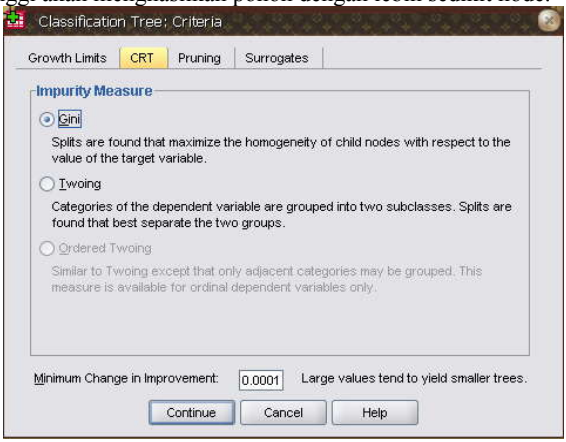
Gambar 10. Dialog Box untuk Pohon Klasifikasi

3. Selanjutnya klik dialog box Criteria, yang terdiri atas : Growth Limits, CRT, Pruning, dan Surrogates.
   a. Growth Limits berfungsi untuk membatasi jumlah level/ tingkatan pada pohon dan mengontrol jumlah minimum kasus pada node ayah dan node anak.
      i) Pada Maximum Tree Depth, pilih Automatic. Jika ingin menambah level di bawah node root, pilih Custom dan isi dengan nilai yang diinginkan.
      ii) Pada Minimum Number of Cases, isi jumlah minimum node ayah dan node anak yang diinginkan.



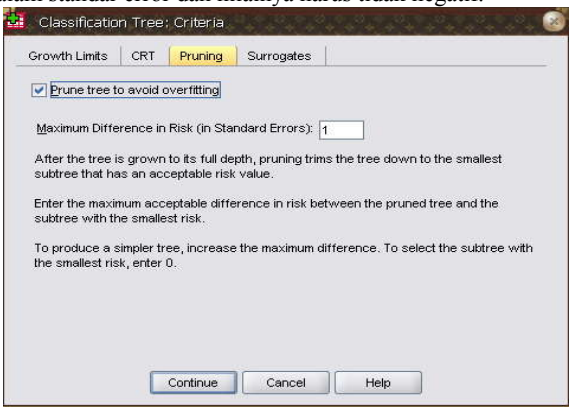
Gambar 11. Growth Limit pada dialog box Criteria

b. Metode pertumbuhan CRT mencoba untuk memaksimalkan node dalam homogenitas, artinya node yang tidak menunjukkan kasus yang homogen adalah suatu indikasi dari impurity. Metode yang dapat digunakan antara lain : metode Gini, Twoing atau Ordered Twoing. Disini juga dapat dipilih Minimum Change in Improvement, yaitu perubahan minimum dalam impurity yang diperlukan untuk memisahkan node, nilai yang ditetapkan adalah 0.0001. Nilai yang lebih tinggi akan menghasilkan pohon dengan lebih sedikit node.



Gambar 12. CRT pada dialog box Criteria

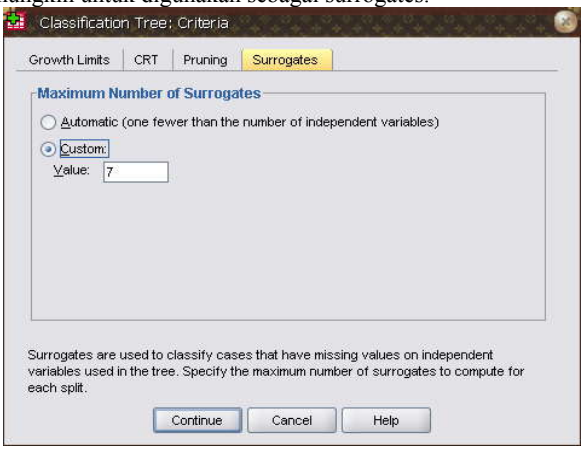
c. Pruning berfungsi untuk menghindari model overfitting dengan cara membabat ukuran pohon tersebut. Pohon tumbuh sampai kriteria tertentu kemudian secara otomatis menjadi subtree yang lebih kecil berdasarkan Maximum Difference in Risk. Nilai resiko dinyatakan dalam standar error dan nilainya harus tidak negatif.



Gambar 13. Pruning pada dialog box Criteria

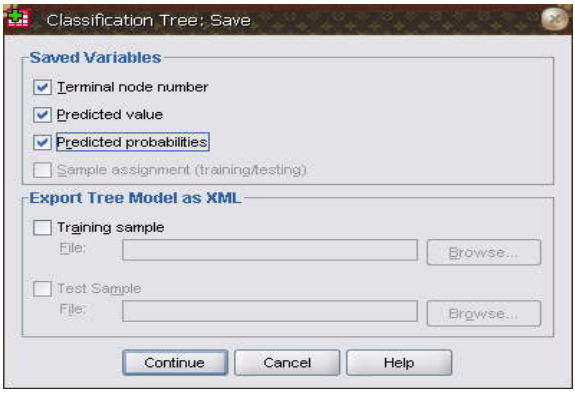
d. Surrogates pada CRT digunakan sebagai pengganti variabel prediktor. Untuk kasus dimana nilai dari variabel tersebut hilang, variabel

prediktor lain yang mempunyai asosiasi yang tinggi dengan variabel original digunakan untuk klasifikasi. Jumlah maksimum surrogates adalah satu kurangnya dari jumlah variabel prediktor. Dengan kata lain, untuk setiap variabel prediktor, semua variabel prediktor yang lain mungkin untuk digunakan sebagai surrogates.



Gambar 14. Surrogates pada dialog box Criteria

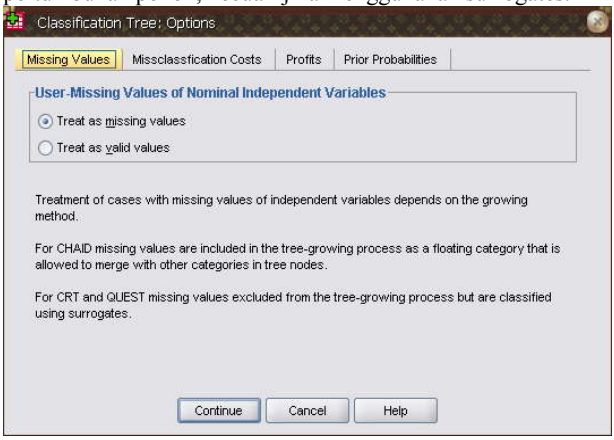
4. Berikutnya klik tab Save, bubuhkan check list (✓) jika ingin menyimpan jumlah terminal node, nilai prediksi dan peluang prediksi yang digunakan selama analisis.



Gambar 15. Save Dialog Box

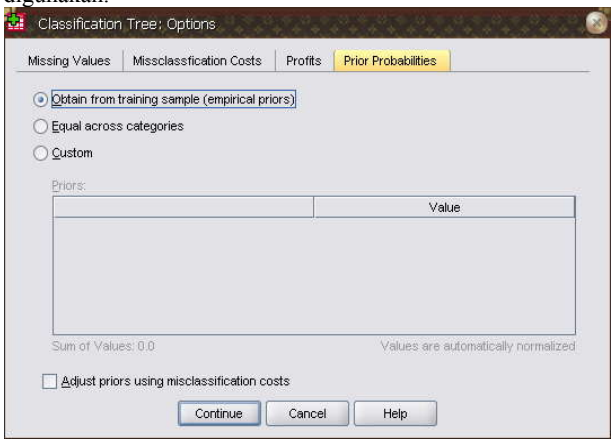
5. Selanjutnya klik tab Option, yang terdiri atas : Missing Values, Misclassification Costs, Profits, dan Prior Probabilities.
   a. Missing values pada variabel prediktor nominal memiliki dua pilihan, yaitu : Treat as missing values (artinya nilai hilang diperlakukan seperti nilai hilang pada suatu sistem) dan Treat as valid values (artinya nilai hilang diperlakukan seperti nilai-nilai biasa). Untuk metode CRT, kasus

dengan nilai hilang pada variabel prediktor dikeluarkan dari proses pertumbuhan pohon, kecuali jika menggunakan surrogates.



Gambar 16. Missing Values pada dialog box Option

b. Taraf misklasifikasi (*misclassification costs*) pada variabel respon kategorik digunakan untuk untuk memberikan informasi yang berhubungan dengan kesalahan klasifikasi.
c. Pada variabel respon kategorik, Profits berfungsi untuk menentukan hasil dan mengeluarkan nilai pada tingkatan/level tertentu.
d. Prior probabilities adalah perkiraan frekuensi relatif dari masing-masing kategori pada variabel respon terhadap variabel prediktor yang digunakan.



Gambar 17. Prior Probabilities pada dialog box Option

6. Kemudian klik tab Output yang berisi : Tree, Statistics, Plots dan Rules. Pilih sesuai dengan kriteria yang diinginkan, dan terakhir klik OK.

DAFTAR PUSTAKA

Adisantoso, J., A. Rambe, dan D. Kusnidar. 1997. Nem dan Status Akreditasi SMP Swasta di Propinsi Jawa Barat. *Forum Statistika dan Komputasi*, **Vol. 2 No. 1 : 24-35**

Afifi, A. A. 1990. *Computer-Aided Multivariate Analysis*. Van Nostrand Reinhold. New York.

Agustina, D. 2006. *Model Persamaan Struktural Faktor-Faktor yang Mempengaruhi Kepuasan Kerja (Studi Kasus di PT. Sinar Harapan Teknik Bengkulu)*. Skripsi pada Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Bengkulu. Tidak Dipublikasikan.

Alhusin, S. 2002. *Aplikasi Statistik Praktis Dengan Menggunakan SPSS 10 for Window*. Surakarta:Graha Ilmu.

Amenta, P. 1998. *Interpolative and Predictive Biplots applied toGeneralized PLS Discriminant Analysis.* http://www.mtisd06.unior.it/collegamenti/MTISD%202006/abstracts/01c_Amenta.pdf.

Ananta, A dan Sri, H. 1985. *Mutu Modal Manusia*. Lembaga Demografi Fakultas Ekonomi Universitas Indonesia. Jakarta.

Anderson, T.W. *An Introduction To Multivariate Statistical Analysis*. John Wiley & Sons, inc. Canada.

Andriyashin, A. 2005. *Financial Applications of Classification and Regression Trees*. Center of Applied Statistics and Economics Humboldt University. Berlin. **edoc.hu-berlin.de/master/andriyashin-anton-2005-03-24/PDF/andriyashin.pdf**

Anonim, 2004. *Cluster Analysis*. http://www.statsoft.com/textbook/stcluan.html.

Anonim, 2005. *Cluster Analysis*. http://149.170.199.144/multivar/ca.htm.

Anonim, 2005. *Cluster Analysis*. http://paleo.cortland.edu/class/stats/LectureNotes/11-cluster.pdf.

Anonim. 2001. http://www.qematel.com/edisi40/gema%2utama.

Anonim. 1977. *Vademecum Bimas Volume III.* Jakarta: Badan Pengendali Bimas.

Anonim. 1999. *Pemantauan Perkembangan Kesejahteraan Rakyat*. Badan Pusat Statistik. Jakarta.

Anonim. 1999. *Panduan Pengelolahan Data dengan Paket Program Minitab Windows*. Edisi Kedua. Jurusan Statistika FMIPA IPB

Anonim. 2000. *What is a Biplot?*. http://tukey.upf.es/xlsbiplot/usersmanual/node3.html.

Anonim. 2001. *Path Analysis*. www2.chass.ncsu.edo/garson/pa765/path.html .12 Mei 2006.

Anonim. 2002. *Data dan Informasi Kemiskinan Tahun 2002*. Badan Pusat Statistik. Jakarta.

Anonim. 2002. Structural Equation Modeling. http://www.statisticsolutions.com/Covariance.htm

Anonim. 2002. *Table of Real: Draw Biplot.* http://www.mathworks.com/access/helpdesk/help/toolbox/stats/index.html?/access/helpdesk/help/toolbox/stats/f72143.html. 3 Juni 2002.

Anonim. 2003. *Biplot*. http://www.geocities.com/bagusco4/mybook/9.html.

Daftar Pustaka

Anonim. 2003. Structural Equation Modeling. http://www2.chass.ncsu.edu/garson/pa765/structur.htm

Anonim. 2003a. *Analisis Peubah Ganda*. Jurusan Statistika IPB. Bogor.

Anonim. 2003b. *Indikator Kesejahteraan Rakyat 2003*. Badan Pusat Statistik. Jakarta.

Anonim. 2003c. *Modul Praktikum Pelatihan Analisis Multivariat*. Jurusan Statistika, Bogor.

Anonim. 2004a. *Cluster Analysis*. http://www.statsoft.com/textbook/stcluan.html.

Anonim. 2004b. *Human Development Report*. http://hdr.undp.org/reports/global/2004/?CFID=987575&CFTOKEN=32593276

Anonim. 2004c. *Kualitas SDM Indonesia Rendah*. http://www.pikiranrakyat.com.

Anonim. 2004d. *Statistik Kesejahteraan Rakyat 2004*. Badan Pusat Statistik . Jakarta.

Anonim. 2005. *Pengembangan Analisis Multivariat dengan SPSS 12*. Jakarta : Salemba Infotek.

Anonim. 2005a. *Cluster Analysis*. http://149.170.199.144/multivar/ca.html.

Anonim. 2005b. *Cluster Analysis*. http://paleo.cortland.edu/class/stats/LectureNotes/11-cluster.pdf.

Anonim. 2005c. *Factor Analysis (Chapter 15, Kachigan)*. http://www/mail.pl.itb.ac.id/~zpontoh/PL211/index.htm

Anonim. 2005d. *Laporan Perkembangan Pencapaian Tujuan Pembangunan Millenium Indonesia*. http://www.undp.org.id/pubs/imdg2004/BI/IndonesiaMDGIBackground.pdf

Anonim. 2006b. *How to Perform and Interpret Factor Analysis using SPSS*. http://www.ncl.ac.uk.htm.

Anonim. 2006d. *Lingkungan Strategis dan Permasalahan.* http://www.menkokesra.go.id/content/Januari2006.html.

Anonim. 2007. *Analisis Korelasi Kanonik* : *Analisis Pengolahan Data*. http//www.deptan.go.id/editama/statistik/web_statistik.doc.

Anonim. 2007. *Correlation Analisys With Sas : The Cancorr Procedure.*

Anonim. 2007. *Degree Of Freedom*. http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics). 26 Agustus 2007 20:30 WIB.

Anonim. 2007. *Penyajian Lebih Lanjut*.

Anonim.1999.*Path Analysis*.www.luna.cas.usf.edu/~mbrannic/files/regression/Pathan.html. 13 Mei 2006.

Anonim.2006a.*Exploratory Factor Analysis*.

Anonim. 2007a. *Classification and Regression Trees*. *www.cems.uwe.ac.uk/~rblawton/classification%20and%20regression%20trees.ppt*

Anonim. 2007b. *Decision Tree Learning*. www.wikipedia.com

Anonim. 2007e. *Tree Structured Classifier*. **www.stat.psu.edu/~jiali/ course/stat597e/notes2/trees.pdf**

Anonim. 2007f. *SPSS Classification Trees™ 13.0*. **https://www.washington.edu/uware/ spss/docs/ClassificationTrees13.0.pdf**

Anton, H. 1995. *Aljabar Linier Elementer*. Edisi Kelima. Erlangga. Jakarta

Ardianto, R. 2007. *Model Pengawasan Kemetrologian Pada Stasiun Pengisian Bahan Bakar Umum* (SPBU).

Argadiredja, D. 2003. *Program Penanggulangan Kemiskinan Bidang Kesehatan*. http://www.bppt.go.id/rakorbangnas03/depkes4.pdf.

Arikunto, S. 2002. *Prosedur Penelitian : Suatu Pendekatan Praktek*. Rineka Cipta, Jakarta.

Bachrudin, A. dan Harapan L. Tobing. 2003. *Analisis Data untuk Penelitian Survei dengan Menggunakan Lisrel 8, Dilengkapi Contoh Kasus*. Bandung : Universitas Padjajaran.

Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone. 1984. *Classification and Regression Trees.* Monterey, California, U.S.A.: Wadsworth, Inc.

Brotodiharjo,R.S. 1998. *Pengantar Ilmu Hukum Pajak*. Refika Aditama. Bandung

Busrah, E. 2004. *Analisis Produksi dan Pendapatan Usahatani Kacang Tanah dengan Sistem Olah Tanah dan Tanpa Olah Tanah serta Pemasarannya (Studi Kasus : Desa Retak Mudik Kecamatan Pondok Suguh Bengkulu Utara)*. Skripsi FP. Tidak dipublikasikan.

Carey, G.1998. *Multiple Regression and Path Analysis*. www.exeter.ac.uk/~SEGLea/multvar2/pathanal.html. 12 Mei 2006

Cattin, P., and D. R. Wittink. 1982. Commercial use of conjoint analysis: A survey. *Journal of Marketing*, 46:3, 44–53.

Chee, J. C. 2002. Partitioning Groups using Classification and Regression Tree in Biomedical Research**.** *binfo.ym.edu.tw/edu/seminars/pdf/CART_YMUBC.pdf*

Darlington, R. B. 2006. *FactorAnalysis*. http: //comp9.psych.cornell.edu/Darlington/factor.html.

Devore, J. L. 2004. *Probability and Statistics for Engineering and The Sciences*. Sixth Edition. Thomson Brooks/Cole : Canada.

Dillon, W.R. & M. Golstein. 1984. *Multivariate Analysis Method & Applications*. John Wiley & Sons, inc, Canada.

Djojonegoro, W. 1995. *Peningkatan Kualitas Sumber Daya Manusia untuk Pembangunan*. Jakarta. Depdikbud.

Draper, N. and H. Smith. 1992. *Analisis Regresi Terapan*. Jakarta : PT Gramedia.

Febriyani, C. 2001. *Pengembangan Produk Dengan Analisis Konjoin*. Jakarta

Ferdinand, A. 2002. *Structural Equation Modeling dalam Penelitian Manajemen*. Semarang : BP UNDIP

Feldman, D. & S. Gross. 2003. Mortgage Default : Classification Trees Analysis. The Pinhas Sapir Center for Development. Tel-Aviv University. sapir.tau.ac.il/papers/sapir-wp/3-03.pdf

Fitriyanti, W. 2006. *Deskripsi Tingkat Kepuasan Penumpang Garuda Indonesia Tahun 2006* (Skripsi STIS). Jakarta. www.youngstatistician.com forest.psych.unc.edu/research/vista-frames/pdf/chap11.pdf. 06 juni 2007. 13:29 WIB.

Galindo, M.P., F. Gomez, V. Villardon, A. Zarza dan M. Vallejo. 1999. *RCMP-Biplot as a Tool to Inspect Environmental Data.* **Error! Hyperlink reference not valid.**.

Gaspersz, V. 1992. *Teknik Analisis dalam Penelitian Percobaan*. Bandung : Tarsito.

Genovart, M., M. McMinn, & D. Bowler. 2003. A Discriminant for Predicting Sex in the Balearic Shearwater. *Waterbird.* Vol. 26 No.1 : 72-76

Gregorius,S. 2005. *Menanggulangi Kemiskinan Desa*. http/www. Ekonomirakyat .org/edisi_14/artikel_6. htm

Hadi, A.F. 2000. *Pendekatan Eksplorasi Peubah Ganda* (*Multivariate*) *untuk Penelitian Pemasaran.* http://72.14.235.104/search?q=cache:Edpy90sVrgUJ:unej.ac.id/fakultas/mipa/majalah_mat/2000/Pendekatan%2520Eksplorasi-Alf.pdf+analisis+peubah+ganda&hl=id&ct=clnk&cd=5&gl=id. 24 Agustus 2007 14:14 WIB.

Hadi, S. 1977. *Metodologi Research*. Jilid II. Yayasan Penerbitan Fakultas Psikologi Universitas Gajah Mada. Yogyakarta.

Hair, JF. *et al*. 1998. *Multivariate Analysis Fifth Edition*, New Jersey: Prentice. Hall International.

Hens, N., L. Bruckers, M. Arbyn, M. Aerts & G. Molenberghs. 2002. Classification Tree Analysis of Cervix Cancer Screening in the Belgian Health Interview Survey 1997. *Arch Public Health. 60 : 275-294.* www.iph.fgov.be/aph/pdf/aphfull60_275_294.pdf

Hubert, M.. & K.V. Driessen. 2002. *Fast and Robust Discriminant Analysis.* www. kuleuven.ac.be.

Ichsan, M.H. 1986. *Buku Materi Pokok Administrasi Perpajakan*. Universitas Terbuka. Depdikbud Jakarta.

Ilya, L dan E. P. Smith, 2001. *Biplot and Singular Value Decomposition Macros for Excel.* http://www.stat.org.vt.edu/vining/keying/biplot.doc. 5 September 2001

Ismail, Z. 1999. *Penanggulangan Kemiskinan Masyarakat Perkampungan Kumuh di Perkotaan. Puslitbang Ekonomi dan Pembangunan LIP*I. Jakarta

James, M.G & Weaver, W. 1987. *Aljabar Matriks Untuk Para Insinyur*. Edisi Kedua. Erlangga. Jakarta.

Jatmiko,P. B. 2004. *Pengelompokkan Desa/Kelurahan di Kabupaten Alor tahun 2002*. Skripsi Program Sarjana. STIS. Jakarta.

Johnson, D. E. 1998. *Applied Multivariate Methods for Analysis*. Kansas State University, USA.

Johnson, R.A & Wichern, D.W. 2002. *Multivariate Analysis Methods and Application*. By John Wiley & Sons.

Johnson, R.A dan D.W. Winchern. 2002. *Applied Multivariate Statistical Analysis. Fifth edition.* Prentice Hall. New Jersey.

Kenny, A. D. 1998. Multiple Factor Models. http://davidakenny.net/com/mfactor.htm.

Kenny, A. D. 2002. Single-Factor Model. http://davidakenny.net/com/1factor.htm.

Koentjoro. 1986. Pengaruh Tingkat Pendidikan Orangtua Terhadap Prestasi Belajar Anak di SD Pedesaan. Psikologi UGM. *Ringkasan Hasil Penelitian 1982-1986*. Depdikbud, Jakarta.

Kohler, U. 2004. *Biplot, revisited.* http://fmwww.bc.edu/repec/usug2004/biplot.pdf. 24 Juni 2004.

Kucukkocaoglu, G. & O. Sezgin. 2007. *IPO Mechanism Selection by Using Classification and Regression Trees (CART)*. Başkent University, Faculty of Economics and Administrative Sciences, Management Department, Bağlıca, Ankara. Turkey.

Kutner, M.H., C.J. Nachtsheim, J. Neter & W. Li. 2005. *Applied Linear Statistical Models*. Fifth Edition. Mc Graw-Hill International Edition.

Lee, bee-leng. 2007. *Correspondence Analysis*.

Mamesah, D.J. 1995. *Sistem Administrasi Keuangan Daerah*. Gramedia Pustaka Utama. Jakarta.

Mantir, S. 1986. Faktor-faktor yang Mempengaruhi Prestasi Belajar pada Mahasiswa Program Studi PLS FKIP Universitas Palangkaraya. FKIP Universitas Palangkaraya. *Ringkasan Hasil Penelitian 1982-1986*. Depdikbud, Jakarta.

Mardalis. 1989. *Metode Penelitian Suatu Pendekatan Proposal*. Jakarta : Bumi Aksara.

Mardalis. 1995. *Metode Penelitian*. Bumi Aksara. Jakarta.

Maryatin, D. 2002. *Analisis Korespondensi Data Kriminologi Polres Jember*. Jember:Universitas Jember.

Mawarti, I., IM Tirta, & Rita Ratih, T. 2003. *Variabel Boneka (Dummy Variable) dalam Analisis Regresi Linier (Dummy Variable To Linear Regression Analysis)*. www.unej.ac.id/fakultas/mipa/skripsi/matematika/indah98.pdf. 13 Mei 2006

Michael, G. 2007. *Correspondence Analysis In Practice. Barcelona: Universitat Pompeu Fabra.*

Michael, G. 2007. *Correspondence Analysis Versus Spectral Mapping/Weigted Log-Ratio Analysis*. http://www.econ.upf.es/~michael. 05 maret 2007

Milne, G.R. 2005. *ClusterAnalysis.* http://intra.som.umass.edu/georgemilne/powerpoint/cluster%analysis.ppt

Moehar, D. 2002. *Metode Penelitian Sosial Ekonomi.* Bumi Aksara. Bandung

Moloe,1999.----. http://www.sehat2010.com.

Montgomery, C.D. and W.W. Hines. 1990. *Probabilita dan Statistik dalam Ilmu Rekayasa dan Manajemen*. Jakarta : Universitas Indonesia.

Morrison, D.F. 1981. *Multivariate Statistical Methods*. Mc Graw Hill. New York.

Mulyani, E. 2006. *Model Log-Linier Beberapa Kasus Kriminologi yang terjadi di Wilayah Polres Bengkulu pada Tahun 2004/2005.* Skripsi pada Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Bengkulu. Tidak Dipublikasikan.

Murti, B. 1996. *Penerapan Metode Statistik Non Parametrik dalam Ilmu-Ilmu Kesehatan.* Jakarta: PT Gramedia Pustaka Utama.

Notoatmodjo, S. 2002. *Metodologi Penelitian Kesehatan*. Rineka Cipta. Jakarta.

Notoatmodjo, S. 2003. *Pengembangan Sumber Daya Manusia*. PT.Rineka Cipta. Jakarta.

Nugroho, B.A. 2005. *Strategi Jitu Memilih Metode Statistik Penelitian Dengan SPSS*. Yogyakarta: Andi.

Nugroho, S. 1989. *Sidik Komponen Utama dan Sidik Diskriminasi Faktor Keberhasilan Studi di Tahun Pertama Mahasiswa Faperta UNIB*. Laporan Penelitian, Universitas Bengkulu.

Nurdianto, A. 2004. *Analisis Kualitas Operasi Jasa Rumah Bersalin (Studi Kasus Rumah Bersalin Kasih Ibu).* Skripsi pada Jurusan Manajemen, Fakultas Ekonomi. Universitas Bengkulu. Bengkulu. Tidak Dipublikasikan.

Portier, K.M. 2003. *Some Old and New Approaches to Cluster Analysis.* Florida. http://www.ifasstat.ufl.edu/sta4702/pdf/lecture10.pdf.

Purba, T. 1986. Pengaruh Latar Belakang Pendidikan Mahasiswa Terhadap Hasil Belajar pada Matakuliah Gambar Teknik di Jurusan Seni Rupa. IKIP Jakarta. *Ringkasan Hasil Penelitian 1982-1986*. Depdikbud, Jakarta.

Rachmat, A. 2007. *Manipulasi Tree*. Handout Struktur Data Prodi Teknik Informatika UKDW.

Ratnasari, V. 2005. *Faktor-faktor yang Mempengaruhi Kepuasan dan Performansi Kerja Pegawai dengan Pendekatan Structural Equation Modeling di PT 'X'*: Prosiding Seminar Nasional Statistika VII. hal:247-259.

Rencher, A. C. 1995. *Method of Multivariate Analysis*. John Wiley & Sons,Inc.Canada

Ruseffendi. 1994. *Dasar-dasar Penelitian Pendidikan dan Bidang-bidang Non Eksakta Lainnya*. IKIP Semarang Press, Semarang.

Rusfidra, 2001. *Peranan Pendidikan Tinggi Jarak Jauh untuk Mewujudkan Knowledge Based Society*. http://www.depdiknas.go.id/jurnal/34/peranan-pendidikan-tinggi-jarak-jauh/html.

Sambamoorthi, N. 2005a. *Hierarchical Cluster Analysis: Some Basic and Algorithms*. http://www.crmportals.com.

Santoso, A., H. Wijayanto, I. D. Salvianti. 1997. Penggunaan Analisis Diskriminan Metode MDP *(Minimum Distance Probability)* pada Data Biner. *Forum Statistika dan Komputasi*, Vol. 2 No. 1 : 24-35.

Santoso, S. 2004. *Buku Latihan SPSS Statistik Multivariat*. Elex Media Komputindo, Jakarta.

Santoso, S. 2004. *SPSS : Statistik Multivariat*. PT Elex Media Komputindo, Jakarta.

Sartono, B., F. M. Affendi, U. D. Syafitri, I. M. Sumertajaya, dan Y. Angraeni. 2003. *Analisis Peubah Ganda*. Fakultas Matematika dan Ilmu Pengetahuan Alam. Institut Pertanian Bogor. Bogor.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York

Sembiring, R. K. 1995. *Analisis Regresi*. Penerbit ITB, Bandung.

Sezgin, O. 2006. *Statistical Methods In Credit Rating*. Department of Financial Mathematics. The Middle East Technical University. Turkey. www3.iam.metu.edu.tr/iam/ images/2/21/Özgesezginthesis.pdf

Singgih, S. 2004. *Buku Latihan SPSS Statistik Multivariat*. PT Elex Media Komputindo Kelompok Gramedia. Jakarta.

Sodarsono, A dan I N, Latra. 2005. *Analisis Hubungan Antara Kecamatan Dengan Jumlah Penderita Penyakit Jenis Communicable Diseases di Kabupaten Mojekerto*. Prosiding Seminar Nasional Statistika VII Jurusan Statistik. Surabaya: ITS.

Soekartawi. 1990. *Teori Ekonomi Produksi, Analisis Fungsi Cobb-Douglass*. Jakarta : Rajawali.

Soekartawi. 1993. *Teori Ekonomi Produksi*. Jakarta: Rajawali Pers.

Soekartawi. 2006. *Strategi Mengentaskan Kemiskinan di Indonesia Melalui Inpres Desa tertinggal*. Jurnal volume 7.2 hal 1-14

Sudjana. 2002. *Metoda Statistika*.Tarsito. Bandung

Sudjana. 2002. *Teknik Analisis Regresi dan Korelasi bagi Para Peneliti*. Tarsito, Bandung.

Sudjana. 2002. *Teknik Analisis Regresi dan Korelasi Bagi Para Peneliti*. Edisi ke-3. Penerbit Tarsito, Bandung.

Sugiyono. 2003. *Statistika Untuk Penelitian*. Alfabeta. Bandung.

Suharjo, B dan Siswadi. 1999. *Analisis Eksplorasi Data Peubah Ganda dan SPSS 7.5*. Bogor: FMIPA-IPB.

Sujamto. 1984. *Otonomi Daerah Yang Nyata dan Bertanggung Jawab*. Cetakan Pertama. Chalia Indonesia. Jakarta.

Sumarno. 1999. *Teknik Budidaya Kacang Tanah*. Bandung: Sinar baru.

Supranto, J. 1989. *Statistika : Teori dan Aplikasi, jilid 2*. Edisi ke-5. Penerbit Erlangga

Supranto, J. 2004. *Analisis Multivariat: Arti & Interprestasi*. Rineka Cipta. Jakarta

Suwarno. 1986. Studi Lingkungan Tempat Tinggal Terhadap Kegiatan Belajar Mahasiswa IKIP Yogyakarta. IKIP Yogyakarta. *Ringkasan Hasil Penelitian 1982-1986*. Depdikbud, Jakarta.

Taufiqurrahman. 2003. Skripsi "*Analisis Dana Alokasi Umum Propinsi Bengkulu dalam Kerangka Pelaksanaan Otonomi Daerah"*. UNIB. Bengkulu.

Timofeev, R. 2004. *Classification and Regression Trees (CART) Theory and Applications*. Center of Applied Statistics and Economics Humboldt University. Berlin. edoc.hu-berlin.de/master/ timofeev-roman-2004-12-20/PDF/timofeev.pdf

Tjiptoherijanto, P. 1993. *Sumber Daya Manusia dalam Pembangunan Nasional*. Lembaga Penerbit Fakultas Ekonomi Universitas Indonesia. Jakarta.

Udina. F. 2005. *Interactive Biplot Construction.* http://www.jstatsoft.org/. Februari 2005.

Umar, H. 2003. *Metode Penelitian Untuk Skripsi Dan Tesis Bisnis*. PT Grafindo Persada. Jakarta.

Venables, W.N. & B.D. Ripley. 1995. *Modern Applied Statistics with S-Plus*, Springer, Verlag. New York.

Wahyudian. 2003. *Analisis Faktor-Faktor yang Mempengaruhi Konsumsi Kopi dan Analisis Pemetaan Beberapa Merek Kopi dan Implikasinya Pada Pemasaran Kopi.* Jurnal Manajemen dan Agribisnis Vol. 1 No. 1 April 2003: 55-68

Wibowo, A. 2005. *Pengantar Analisis Jalur (Path Analysis)*. Surabaya : Lembaga Penelitian Universitas Airlangga.

Wibowo, A. 2005. *Pengantar Analisis Persamaan Struktural*. Surabaya: Lembaga Penelitian Universitas Airlangga.

Wibowo, W. 2002. Perbandingan Hasil Klasifikasi Analisis diskriminan dan Regresi Logistik pada Pengklasifikasian Data Respon Biner. *Kappa,* Vol 3, No 1: 36-45.

Wilkinson, L. 2005. *FactorAnalysis*. http://www.2.chass.ncsu.edu/garson/pa765/factor.html

William, R.D & M.Goldstein. 1984. Multivariate Analysis Methods And Applications. John, W & sons. Canada.

Wittink, D. *et al*. 1992. *The Number Of Levels Effect In Conjoint*. www.sawtoothsoftware.com

Yohannes, Y. & P. Webb. 1999. **Classification and Regression Trees, CART™: A User Manual For Identifying Indicators of Vulnerability to famine and Chronic Food Insecurity.** ***International Food Policy Research Institute. Washington, U.S.A. www.ifpri.org/pubs/microcom/micro3.pdf***

Yoo, S., S. Kim, dan K. Choi, 2000. *Closeness between Objects and Variables in a Biplot.* http://maths.fs.utm.my/matema~1/azme.pdf www.balipost.co.id/balipostcetaK/2005/8/3/k4.htm

Zambon, M., R. Lawrence, A. Bunn & S. Powell. 2006. *Effect of Alternative Splitting Rules on Image Processing Using Classification Tree Analysis*. Photogrammetric Engineering & Remote Sensing Vol. 72, No. 1 : 25–30.

Zhou, Z. H. 2007. *Data Mining Chapter 5 : Classification and Regression*. Department of Computer Science and Technology Nanjing University. China.

SIGIT NUGROHO, Ph.D. (University of Kentucky-USA, 1994) dilahirkan di Surakarta pada tanggal 30 Nopember 1960. Ia menyelesaikan Pendidikan Dasar dan Menengahnya di Yogyakarta. Setelah tamat **SMA Negeri III 'Padmanaba' Yogyakarta**, ia meneruskan studinya di Institut Pertanian Bogor pada tahun 1980 melalui jalur Proyek Perintis II. Lulus sebagai **Sarjana Statistika (Ir.)** tahun 1984 dari Jurusan Statistika – **Fakultas Matematika dan Ilmu Pengetahuan Alam – Institut Pertanian Bogor (FMIPA-IPB)**. Sejak awal 1986 ia bekerja sebagai staf pengajar pada Fakultas Pertanian Universitas Bengkulu (Faperta UNIB), yang selanjutnya pada tahun 2000 pindah ke Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Bengkulu. Sampai buku ini ditulis, jabatan akademiknya adalah **Lektor Kepala** dalam bidang Statistika.

Pada tahun 1987 ia melanjutkan studinya di Department of Statistics, University of Kentucky, U.S.A. dan meraih gelar **Master of Science (M.Sc.**) dalam bidang Statistika pada tahun 1989. Setelah dua tahun kembali ke Universitas Bengkulu mengasuh mata kuliah Matematika I dan II, Metode Statistika I dan II, serta Rancangan Percobaan di Faperta UNIB ia kembali meneruskan studinya pada tahun 1991 ke jenjang yang lebih tinggi di tempat yang sama (Department of Statistics, University of Kentucky, U.S.A). Dibawah bimbingan **Zakkula Govindarajulu, Ph.D.** (Minnesota, 1961), ia menyelesaikan disertasinya yang berjudul "*On the Locally Most Powerful Rank Test of the Two-way Experiment*" dan dinyatakan lulus pada tanggal 15 April 1994 dihadapan tim penguji yang terdiri dari: William S. Griffith, Ph.D., Henry Howard, Ph.D., William S. Rayens, Ph.D., Mokhtar Ali, Ph.D., Mai Zhou, Ph.D. dan mendapatkan gelar **Doctor of Philosophy (Ph.D.)** dalam bidang Statistika.

Pada tahun 1988 penulis mengikuti Kursus "*Analysis of Messy Data*" di Washington, D.C. yang diberikan langsung oleh penulis buku tentang analisis tersebut, yaitu: George M. Milliken, Ph.D. dan Dallas T. Johnson, Ph.D.

Selain sebagai staf pengajar (**Lektor Kepala** dalam bidang Statistika) Universitas Bengkulu, ia juga sebagai dosen tamu pada program doktor di Jurusan Statistika IPB (2003) dan beberapa pendidikan tinggi lainnya. Sebagai tambahan, ia juga sebagai konsultan *Data Analysis* (Matematika, Operation Research dan Statistika). Pada tahun 2003-2006 penulis juga menjadi **Senior Instruktur** pada Divisi Pendidikan dan Pelatihan **PT. Bank Rakyat Indonesia (Persero) Tbk.** Sampai dengan tahun 1997 penulis juga menjadi anggota *American Statistical*

Association. Berbagai kegiatan seminar dalam bidang statistika telah diikutinya baik lokal, nasional, regional, ataupun internasional.

Beberapa Publikasi Jurnal yang berhubungan dengan bidang ilmunya:

1. Uji Nonparametrik Perlakuan Acak dalam Rancangan Acak Kelompok Lengkap. *Forum Statistika dan Komputasi IPB* (1997) **2** (1), 10-14 ISSN 0853-8115 Tests
2. Tests for Random Effects in Two-way Experiment with One Observation per Cell. *Indian Journal of Mathematics* vol **41** No 1 January 1999. B.N. Prasad Birth Centenary Commemoration Volume. (with Dr. Z. Govindarajulu, Univ of Kentucky)
3. Nonparametric tests for random effects in the balanced incomplete block design. *Statistics & Probability Letters* **56**, 431-437 2002. (with Dr. Z. Govindarajulu, Univ of Kentucky)
4. Some Notes on Nonparametric Test of Random Treatment Effects in One-way and Two-way Experiments. *Journal of Quantitative Methods* nol **3** no 2. 2007 (with Dr. Z. Govindarajulu, Univ of Kentucky)

SIGIT NUGROHO, Ph.D.
Website: http://www.stasignug.cjb.net/
Email : snugroho@unib.ac.id
snsw1960@telkom.net dan sinugsta@yahoo.com

Untuk data obervasi dari beberapa peubah / variabel yang diamati dari satu obyek, kita kenal dengan data peubah banyak atau multivariat. Bila dalam statistika satu peubah atau univariat kita dapat memandang data berada dalam sebuah garis lurus, maka untuk statistika multivariat, data dengan *k* peubah akan berada dalam ruang berdimensi-*k*

Berbagai alat analisis dikenalkan dalm buku ini, mulai dari bagaimana mereduksi variabel, mencari faktor penentu dari sekelompok variabel, menentukan posisi sebuah data berdasarkan urutan suatu kriteria, mengelompokkan data berdasarkan sifat kemiripan, dan beberapa pengembangan alat analisis multivariat lainnya.

Ilustrasi penggunaan paket program untuk analisis konjoin dan regresi pohon juga diberikan dalam tahapan-tahapan untuk memudahkan pengguna.

ISBN 978-979-9431-36-3
9 789799 431363
Stat Multivariat Terapan

UNIB Press
Jl. WR Supratman
Bengkulu 38371