BAB III

ANALISIS KORELASI KANONIK ROBUST DENGAN METODE *MINIMUM COVARIANCE DETERMINAN*

3.1 Deteksi Pencilan Multivariat

Pengidentifikasian pencilan pada kasus multivariat tidaklah mudah untuk dilakukan, hal ini dikarenakan adanya efek *masking* dan *swamping*. *Masking* terjadi pada saat pengamatan pencilan tidak terdeteksi karena adanya pengamatan pencilan lain yang berdekatan, sedangkan *swamping* terjadi pada saat pengamatan baik teridentifikasi sebagai pengamatan pencilan. Ada beberapa cara untuk mendeteksi pencilan dalam data multivariat, yaitu melalui jarak Mahalanobis atau melalui jarak *Robust*.

3.1.1 Jarak Mahalanobis

Pemeriksaan pencilan dalam kasus data multivariat dilakukan berdasarkan jarak kuadrat Mahalanobis dari $\boldsymbol{x}_i$ ke $\boldsymbol{\mu}$, yang didefinisikan sebagai berikut:

$$MD_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}), \quad i = 1,2, \dots, n$$

di mana :

$\boldsymbol{\mu}$ =vektor rata-rata

$\boldsymbol{\Sigma}$ =matriks kovarian

n =banyaknya pengamatan.

Sebuah pengamatan $\boldsymbol{x}_i$ diidentifikasikan sebagai pencilan jika jarak Mahalanobisnya lebih besar dari nilai khi-kuadrat tabel dengan taraf signifikansi

$(1-\alpha)$ dan derajat bebas p ($MD_i^2 > \chi^2_{p,(1-\alpha)}$ atau setara dengan $MD_i > \sqrt{\chi^2_{p,(1-\alpha)}}$).

Penggunaan jarak Mahalanobis dalam mendeteksi pencilan dapat dikatakan kurang baik. Hal ini dikarenakan adanya efek *masking* dan *swamping*. Efek *masking* dapat menurunkan jarak Mahalanobisnya sehingga jarak antar titik terpencil saling berdekatan, sedangkan efek *swamping* dapat meningkatkan jarak Mahalanobisnya sehingga dimungkinkan terjadinya kesalahan berupa pengamatan baik yang teridentifikasi sebagai pengamatan pencilan.

3.1.2 Jarak *Robust*

Jarak *robust* merupakan suatu pendekatan untuk mengidentifikasi pencilan pada data multivariat, yaitu dengan menggunakan penaksir dari $\boldsymbol{\mu}$ dan $\boldsymbol{\Sigma}$ pada metode *robust*. Sehingga metode ini mampu meminimumkan pengaruh dari adanya efek *masking* dan *swamping* dalam pendeteksian pencilan (Rencher, 2002).

Pemeriksaan pencilan dalam kasus data multivariat dilakukan berdasarkan jarak kuadrat *robust* dari $\boldsymbol{x}_i$ ke $\boldsymbol{T}$, yang didefinisikan sebagai berikut:

$$RD_i^2 = (\boldsymbol{x}_i - \boldsymbol{T})^t \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T}), \quad i = 1,2,\dots,n$$

di mana :

$\boldsymbol{T}$ =vektor rata-rata dari estimasi *robust*

$\boldsymbol{C}$ =matriks kovarian dari estimasi *robust*

n =banyaknya pengamatan.

Sebuah pengamatan $\boldsymbol{x}_i$ diidentifikasikan sebagai pencilan jika jarak *robust*nya lebih besar dari nilai khi-kuadrat tabel dengan taraf signifikansi $(1-\alpha)$ dan derajat bebas p ($RD_i^2 > \chi^2_{p,(1-\alpha)}$ atau setara dengan $RD_i > \sqrt{\chi^2_{p,(1-\alpha)}}$).

3.2 Analisis Korelasi Kanonik

3.2.1 Pengertian dan Tujuan Analisis Korelasi Kanonik

Analisis korelasi kanonik adalah salah satu teknik analisis statistik, yang digunakan untuk melihat hubungan antara himpunan variabel bebas $(X_1, X_2, \dots, X_p)$ dengan himpunan variabel terikat $(Y_1, Y_2, \dots, Y_q)$. Analisis ini dapat mengukur tingkat keeratan hubungan antara himpunan variabel bebas dengan himpunan variabel terikat. Analisis korelasi kanonik berfokus pada korelasi antara kombinasi linier dari himpunan variabel terikat dengan kombinasi linier dari himpunan variabel bebas. Ide utama dari analisis ini adalah mencari pasangan dari kombinasi linier ini yang memiliki korelasi terbesar.

3.2.2 Asumsi-asumsi dalam Analisis Korelasi Kanonik

Adapun asumsi-asumsi dalam analisis korelasi kanonik adalah sebagai berikut :

1. Banyaknya Variabel

   Variabel bebas dan variabel terikat terdiri dari lebih dari satu variabel dan berskala interval. Jika data berskala ordinal, maka data tersebut harus ditransformasi terlebih dahulu ke skala interval.

2. Uji Multikolinieritas

Menurut Hair (1998), multikolinieritas terjadi ketika dua atau lebih variabel memiliki nilai korelasi yang tinggi. Pengujian multikolinieritas dapat dilakukan dengan melihat besarnya nilai korelasi antar variabel bebasnya dan antar variabel terikatnya. Rumus dari nilai korelasi yang digunakan adalah nilai korelasi r Pearson.

Menurut Hocking (2003) pengujian multikolinieritas dengan menggunakan nilai korelasi antar variabel bebasnya atau veriabel terikatnya dapat menggunakan kriteria berikut :

Jika nilai $r_{ij} > 0{,}95$ maka terdapat kolinieritas yang tinggi

Jika dalam suatu data terdapat kolinieritas yang tinggi, maka menurut Nachrowi (2008) salah cara untuk mengatasinya adalah dengan tidak mengikutsertakan salah satu variabel yang kolinier.

3. Uji Normalitas

Johnson dan Winchern (2007) mengemukakan bahwa untuk menguji apakah suatu himpunan data berdistribusi normal multivariat adalah dengan menggunakan Q-Q plot yang didasarkan pada jarak kuadrat Mahalanobis. Adapun langkah-langkah untuk membuat Q-Q plot adalah sebagai berikut:

1) Hitung nilai d_j^2 di mana

$$d^2_{\sim} = (x_j - \bar{x})^t S^{-1} (x_j - \bar{x}) \qquad j = 1,2,\dots,n$$

$n =$ banyaknya pengamatan.

2) Urutkan jarak kuadrat Mahalanobis tersebut dari yang terkecil sampai terbesar

$d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$.

3) Setiap $d_{(j)}^2$ dihitung nilai $\frac{\left(j-\frac{1}{2}\right)}{n}$, di mana (j) adalah indeks bawah yang menunjukkan peringkat ke-j.

4) Hitung nilai $q_{j,p}\left(\frac{\left(j-\frac{1}{2}\right)}{n}\right)$, yaitu nilai khi-kuadrat dari $\frac{\left(j-\frac{1}{2}\right)}{n}$ dengan derajat bebas p, di mana p adalah banyaknya variabel

5) Gambar plot tersebut dengan koordinat $\left(q_{j,p}\left(\frac{\left(j-\frac{1}{2}\right)}{n}\right), d_{(j)}^2\right)$

Data dikatakan berdisribusi normal jika plot membentuk garis lurus (linier) atau paling tidak 50% dari nilai d_j^2 lebih kecil dari $\chi_{p,0.05}^2$ $\left(d_j^2 \leq \chi_{p,0.05}^2\right)$ (Anderson, 1999).

4. Uji Linieritas

Linearitas dapat dikatakan penting untuk analisis korelasi kanonik dan itu mempengaruhi dua aspek hasil korelasi kanonik. Pertama, koefisien korelasi kanonik antara sepasang variabel kanonik adalah berdasarkan hubungan linier Jika variabel kanonik berhubungan secara nonlinier, maka koefisien korelasi kanonik tidak akan menangkap hubungan tersebut. Kedua, analisis korelasi kanonik memaksimalkan hubungan linier antara variabel kanonik. Jadi, meskipun analisis korelasi kanonik adalah metode multivariat yang paling umum, masih dibatasi untuk mengidentifikasi hubungan linier. Jika hubungan tidak linier,. maka satu atau kedua variabel kanonik harus diubah, itupun jika memungkinkan (Hair, 1988).

Pengujian linieritas dilakukan antara sepasang variabel kanoniknya dan dapat dilihat dari nilai korelasi kanoniknya. Jika nilai tersebut tergolong signifikan secara statistik maka dapat dipastikan bahwa asumsi linieritas telah dipenuhi untuk pasangan variabel kanonik tersebut.

3.2.3 Penentuan Korelasi Kanonik dan Koefisien Variabel Kanonik

Analisis korelasi kanonik adalah suatu teknik yang digunakan untuk menentukan tingkatan asosiasi linier antara dua himpunan variabel, di mana masing-masing himpunan terdiri dari beberapa variabel. Kelompok pertama dari p variabel diwakili oleh $(p \times 1)$ vektor acak $\boldsymbol{X}$. Kelompok kedua dari q variabel diwakili oleh (械× 1) vektor acak $\boldsymbol{Y}$. Asumsikan, dalam pengembangan teoritis, bahwa X mewakili himpunan yang lebih kecil, sehingga $p \leq q$.

Misalkan untuk vektor acak $\boldsymbol{X}$ dan $\boldsymbol{Y}$:

$\langle(\boldsymbol{X}) = \boldsymbol{\mu_X}$; $\quad Cov(\boldsymbol{X}) = \boldsymbol{\Sigma_{XX}}$

$E(\boldsymbol{Y}) = \boldsymbol{\mu_Y}$; $\quad Cov(\boldsymbol{Y}) = \boldsymbol{\Sigma_{YY}}$

$$Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma_{XY}} = \boldsymbol{\Sigma_{YX}}.$$

Vektor acaknya :

$$\boldsymbol{W} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_q]^t,$$

vektor rata-ratanya :

$$\boldsymbol{\mu} = \begin{bmatrix} E(\boldsymbol{X}) \\ E(\boldsymbol{Y}) \end{bmatrix} = \begin{bmatrix} \mu_{\boldsymbol{X}} \\ \mu_{\boldsymbol{Y}} \end{bmatrix}$$

dan matriks kovariannya:

$$\boldsymbol{\Sigma} = E(\boldsymbol{W} - \boldsymbol{\mu})(\boldsymbol{W} - \boldsymbol{\mu})^t = \begin{bmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{YX}} \\ \boldsymbol{\Sigma_{XY}} & \boldsymbol{\Sigma_{YY}} \end{bmatrix}$$

Tugas pokok dari analisis korelasi kanonik adalah meringkaskan kumpulan antara himpunan $\boldsymbol{X}$ dan $\boldsymbol{Y}$. Kombinasi linear menyediakan ringkasan sederhana mengukur suatu himpunan dari variabel. Himpunan $\boldsymbol{U} = \boldsymbol{\alpha^t X}$ dan $\boldsymbol{V} = \boldsymbol{\beta^t Y}$ dengan $\boldsymbol{\alpha}$ dan $\boldsymbol{\beta}$ merupakan koefisien kanonik dan

$$Var(\boldsymbol{U}) = \boldsymbol{\alpha^t \Sigma_{XX} \alpha}$$

$$Var(\boldsymbol{V}) = \boldsymbol{\beta^t \Sigma_{YY} \beta}$$

$$Cov(\boldsymbol{U}, \boldsymbol{V}) = \boldsymbol{\alpha^t \Sigma_{XY} \beta}.$$

Kemudian dapat dicari koefisien vektor $\boldsymbol{\alpha}$ dan $\boldsymbol{\beta}$ sedemikian sehingga,

$$Corr(\boldsymbol{U}, \boldsymbol{V}) = \frac{\boldsymbol{\alpha^t \Sigma_{XY} \beta}}{\sqrt{(\boldsymbol{\alpha^t \Sigma_{XX} \alpha})(\boldsymbol{\beta^t \Sigma_{YY} \beta})}} \tag{3.1}$$

sebisa mungkin bernilai besar. Kombinasi linier yang dapat dibentuk adalah sebanyak min (p, q) pasang.

Teorema Korelasi Kanonik

Misalkan $p \leq q$ dan vektor acak $\boldsymbol{X}$ dan $\boldsymbol{Y}$ mempunyai, $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma_{XX}}$, $Cov(\boldsymbol{Y}) = \boldsymbol{\Sigma_{YY}}$ dan $Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma_{XY}} = \boldsymbol{\Sigma_{YX}}$ di mana $\boldsymbol{\Sigma}$ mempunyai rank lengkap. Untuk koefisien vektor $\boldsymbol{\alpha}$ dan $\boldsymbol{\beta}$, bentuk kombinasi linear $\boldsymbol{U} = \boldsymbol{\alpha^t X}$ dan $\boldsymbol{V} = \boldsymbol{\beta^t Y}$. Maka

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} Corr(\boldsymbol{U}, \boldsymbol{V}) = \rho_1$$

diperoleh dengan kombinasi linear (variabel kanonik bagian pertama).

$$\boldsymbol{U}_1 = \boldsymbol{e}_1^t \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-\frac{1}{2}} \boldsymbol{X} \text{ dan } \boldsymbol{V}_1 = \boldsymbol{f}_1^t \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-\frac{1}{2}} \boldsymbol{Y}.$$

Bagian ke-k dari variabel kanonik, $k = 2,3, \ldots, p$,

$$\boldsymbol{U}_k = \boldsymbol{e}_k^t \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-\frac{1}{2}} \boldsymbol{X} \text{ dan } \boldsymbol{V}_k = \boldsymbol{f}_k^t \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-\frac{1}{2}} \boldsymbol{Y}$$

memaksimumkan

$$Corr(\boldsymbol{U_k}, \boldsymbol{V_k}) = \rho_k^*$$

di antara kombinasi linear yang tidak berkorelasi dengan variabel kanonik sebelumnya. ${\rho_1}^{*2} \geq {\rho_2}^{*2} \geq \ldots \geq {\rho_p}^{*2}$ adalah nilai eigen dari $\boldsymbol{\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}}$ dan $\boldsymbol{e_1, e_2, \ldots, e_p}$ adalah vektor eigen $(p \times 1)$. (banyaknya ${\rho_1}^{*2}, {\rho_2}^{*2}, \ldots, {\rho_p}^{*2}$ juga nilai eigen p paling besar dari matriks $\boldsymbol{\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}}$ yang berkorespondensi dengan $(q \times 1)$ vektor eigen $\boldsymbol{f_1, f_2, \ldots, f_p}$. Tiap $\boldsymbol{f_i}$ adalah proporsi untuk $\boldsymbol{\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}e_1}$).

Variabel kanonik mempunyai sifat sebagai berikut:

$$Var(\boldsymbol{U_k}) = Var(\boldsymbol{V_k}) = 1$$

$$Cov(\boldsymbol{U_k}, \boldsymbol{U_l}) = Corr(\boldsymbol{U_k}, \boldsymbol{U_l}) = 0 \qquad k \neq l$$

$$Cov(\boldsymbol{V_k}, \boldsymbol{V_l}) = Corr(\boldsymbol{V_k}, \boldsymbol{V_l}) = 0 \qquad k \neq l$$

$$Cov(\boldsymbol{U_k}, \boldsymbol{V_l}) = Corr(\boldsymbol{U_k}, \boldsymbol{V_l}) = 0 \qquad k \neq l$$

untuk $k, l = 1,2, \ldots, p$

Dengan bahasa yang lebih sederhana Dillon dan Goldstein dalam Kumaat (2001) mengemukakan bahwa, untuk mendapatkan fungsi kanonik, langkah-langkah yang harus ditempuh adalah :

1. Untuk memperoleh koefisien korelasi kanonik langkah-langkahnya adalah dengan menyusun matriks kovarian $(\boldsymbol{S})$ atau matriks korelasi $(\boldsymbol{R})$. Matriks $\boldsymbol{S}$

dipakai apabila data yang diolah memiliki satuan yang sama, sedangkan matriks $\boldsymbol{R}$ dipakai bila data tersebut tidak memiliki satuan yang sama.

2. Mencari nilai eigen (λ) berdasarkan matriks $\boldsymbol{S}$ atau $\boldsymbol{R}$ pada langkah 1 dengan menggunakan rumus :

$$\left|\mathbf{R}_{\boldsymbol{XX}}^{-\frac{1}{2}}\mathbf{R}_{\boldsymbol{XY}}\mathbf{R}_{\boldsymbol{YY}}^{-1}\mathbf{R}_{\boldsymbol{YX}}\mathbf{R}_{\boldsymbol{XX}}^{-\frac{1}{2}}-\lambda\boldsymbol{I}\right|=0$$

Nilai eigen tersebut digunakan untuk memperoleh vektor eigen, di mana vektor eigen merupakan koefisien variabel kanonik

3. Mencari vektor eigen berdasarkan nilai eigen yang telah diperoleh pada langkah 2 dengan persamaan berikut :

$$\left(\mathbf{R}_{\boldsymbol{XX}}^{-\frac{1}{2}}\mathbf{R}_{\boldsymbol{XY}}\mathbf{R}_{\boldsymbol{YY}}^{-1}\mathbf{R}_{\boldsymbol{YX}}\mathbf{R}_{\boldsymbol{XX}}^{-\frac{1}{2}}-\lambda\boldsymbol{I}\right)\boldsymbol{a}=0$$

$$\left(\mathbf{R}_{\boldsymbol{YY}}^{-\frac{1}{2}}\mathbf{R}_{\boldsymbol{YX}}\mathbf{R}_{\boldsymbol{XX}}^{-1}\mathbf{R}_{\boldsymbol{XY}}\mathbf{R}_{\boldsymbol{YY}}^{-\frac{1}{2}}-\lambda\boldsymbol{I}\right)\boldsymbol{b}=0$$

Vektor eigen tersebut dinotasikan dengan $\boldsymbol{a}$ dan $\boldsymbol{b}$, yang merupakan nilai koefisien variabel kanonik atau disebut juga sebagai pembobot kanonik. Variabel kanonik yang dapat dibentuk berdasarkan vektor eigen tersebut ada sebanyak min (p, q) pasang

4. Setelah memperoleh vektor eigen, selanjutnya dicari korelasi kanonik yang dapat dihitung dengan menggunakan rumus :

$$Corr(\boldsymbol{U},\boldsymbol{V})=\frac{\boldsymbol{a}^{\boldsymbol{t}}\mathbf{R}_{\boldsymbol{XY}}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}^{\boldsymbol{t}}\mathbf{R}_{\boldsymbol{XX}}\boldsymbol{a})(\boldsymbol{b}^{\boldsymbol{t}}\mathbf{R}_{\boldsymbol{YY}}\boldsymbol{b})}}$$

5. Mencari proporsi atau keragaman data yang dijelaskan oleh setiap pasangan variabel kanonik, dengan menggunakan rumus :

$$e_i = \frac{\lambda_i}{1-\lambda_i}\,, i = 1,2,\dots,\min(p,q)$$

$$\text{Keragaman data} = \frac{e_i}{\Sigma e_i}$$

Keragaman data ini digunakan untuk memilih pasangan variabel kanonik mana yang akan dianalisis lebih lanjut. Batasan minimum keragaman kumulatif yang dikemukakan oleh Dillon dan Goldstein (1984) adalah 80%.

6. Melakukan pengujian hipotesis untuk setiap korelasi kanonik. Berdasarkan Johnson dan Winchern (2007), pengujian korelasi kanonik secara individu dilakukan melalui pendekatan distribusi khi-kuadrat dengan rumusan hipotesisnya adalah :

   H_0 : $\rho_i = 0$, artinya tidak ada hubungan yang signifikan antar pasangan variabel kanonik ke-i

   H_1 : $\rho_i \neq 0$, artinya ada hubungan yang signifikan antar pasangan variabel kanonik ke-i

   Kriteria yang digunakan adalah tolak H_0 pada tingkat signifikansi *á*, jika

   $$-\left(n-1-\frac{1}{2}(p+q+1)\right)\ln\prod_{k=i}^{p}(1-\rho_k^2) > \div^2_{(p-i-1)(q-i-1)}(\acute{a}).$$

Guna memudahkan dalam pencarian korelasi kanonik, berikut algoritma untuk analisis korelasi kanonik klasik dan analisi korelasi kanonik *robust*.

Analisis korelasi kanonik klasik :

1. Uji Asumsi
2. Deteksi Pencilan
3. Menentukan matriks kovarians
4. Menentukan nilai eigen dan vektor eigen

5. Menentukan korelasi kanonik dan pembobot kanonik
6. Menentukan proporsi keragaman

Analisis korelasi kanonik *robust*:

1. Uji Asumsi
2. Deteksi Pencilan
3. Menentukan matriks kovarians dengan metode MCD
4. Menentukan nilai eigen dan vektor eigen
5. Menentukan korelasi kanonik dan pembobot kanonik
6. Menentukan proporsi keragaman

3.2.4 C*anonical Loadings* dan *Cross Loadings*

Canonical loadings merupakan korelasi sederhana antara variabel asal dengan masing-masing variabel kanoniknya. Semakin besar nilai *canonical loadings* menunjukkan semakin dekat hubungan antara variabel asal dengan variabel kanoniknya. Menurut Hair (1998) *canonical loadings* variabel terikat diperoleh dengan rumus

$$\boldsymbol{R_{YV} = R_{YY} b}$$

R_{YY} merupakan korelasi sederhana antar variabel $\boldsymbol{Y}$ dan $\boldsymbol{b}$ merupakan vektor koefisien kanonik variabel $\boldsymbol{V}$. Sedangkan *canonical loadings* untuk variabel bebas diperoleh dengan rumus

$$\boldsymbol{R_{XU} = R_{XX} a}$$

R_{XX} merupakan korelasi sederhana antar variabel $\boldsymbol{X}$ dan $\boldsymbol{a}$ merupakan vektor koefisien kanonik variabel $\boldsymbol{U}$.

Canonical Cross loadings merupakan korelasi sederhana antara variabel asal dengan masing-masing variabel kanonik lawannya. Semakin besar nilai *canonical cross loadings* menunjukkan semakin kuat hubungan variabel asal dengan variabel kanonik lawannya. *Canonical Cross loadings* diperoleh dengan cara

$$\boldsymbol{R_{XV} = R_{XU}\rho_k}$$

$$\boldsymbol{R_{YU} = R_{YV}\rho_k}$$

$\boldsymbol{\rho_k}$ adalah nilai korelasi kanonik dari variabel kanonik ke-k.

3.2.5 Redudansi

Redudansi merupakan sebuah indeks yang menghitung proporsi keragaman yang dapat dijelaskan oleh variabel kanonik yang dipilih baik variabel kanonik terikat maupun variabel kanonik bebas.

Proporsi keragaman variabel asal yang diterangkan oleh variabel kanoniknya diperoleh dari perhitungan rata-rata *canonical loadings* yang dikuadratkan.

$$R^2_{\boldsymbol{X|U_i}} = \frac{\sum_{j=1}^{k} R^2_{\boldsymbol{X_jU_i}}}{k}$$

$$R^2_{\boldsymbol{Y|V_i}} = \frac{\sum_{j=1}^{k} R^2_{\boldsymbol{Y_jV_i}}}{k}$$

Proporsi keragaman variabel asal yang diterangkan oleh variabel kanonik lawannya diperoleh melalui perkalian kuadrat korelasi kanonik dengan rata-rata *canonical loadings* yang dikuadratkan, atau dapat dituliskan

$$R^2_{\boldsymbol{X|V_i}} = \rho_k^2 R^2_{\boldsymbol{X|U_i}}$$

$$R^2_{\boldsymbol{Y}|\boldsymbol{U_i}} = \rho_k^2 R^2_{\boldsymbol{Y}|\boldsymbol{V_i}}$$

$\boldsymbol{\rho_k}$ adalah nilai korelasi kanonik dari variabel kanonik ke-k.

3.3 *MINIMUM COVARIANCE DETERMINANT*

3.3.1 Definisi MCD

Misalkan $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_n)^t$ merupakan kumpulan data sejumlah n pengamatan terdiri dari p-variabel di mana $n \geq p + 1$. Penaksir MCD merupakan pasangan $\boldsymbol{T} \in \mathbb{R}^p$ dan $\boldsymbol{C}$ adalah matriks definit positif simetris berdimensi $p \times p$ dari suatu sub sampel berukuran h pengamatan di mana $\frac{(n+p+1)}{2} \leq h \leq n$ dengan

$$\boldsymbol{T} = \frac{1}{h}\sum_{i=1}^{h} \boldsymbol{w}_i$$

dan

$$\boldsymbol{C} = \frac{1}{h}\sum_{i=1}^{h} (\boldsymbol{w}_i - \boldsymbol{T}_i)(\boldsymbol{w}_i - \boldsymbol{T}_i)^t$$

yang meminimumkan determinan $\boldsymbol{C}$ (Buttler, dkk. 1993).

Dalam menentukan penaksir MCD, jika jumlah pengamatan (n kecil) maka penaksir MCD dapat segera ditentukan. Namun jika jumlah pengamatan (n besar) maka akan membutuhkan waktu yang cukup lama untuk menentukan penaksir MCD. Karena keterbatasan ini maka Rousseeuw dan Van Driessen (1999) mengembangkan suatu algoritma FAST-MCD yaitu Teorema C-Steps berikut.

Teorema C-Steps:

Misalkan $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_n)^t$ merupakan himpunan sejumlah n pengamatan terdiri dari p-variabel. Misal $H_1 \subset \{1,2, \dots, n\}$ dengan sejumlah elemen H_1, jumlah $|H_1| = h$, tetapkan

$$\boldsymbol{T}_1 = \frac{1}{h}\sum_{i=1}^{h} \boldsymbol{w}_i$$

dan

$$\boldsymbol{C}_1 = \frac{1}{h}\sum_{i=1}^{h} (\boldsymbol{w}_i - \boldsymbol{T}_i)(\boldsymbol{w}_i - \boldsymbol{T}_i)^t$$

Jika $det(\boldsymbol{C}_1) \neq 0$ maka jarak relatif,

$$d_1(i) = \sqrt{(\boldsymbol{w}_i - \boldsymbol{T}_1)^t \boldsymbol{C}_1{}^{-1} (\boldsymbol{w}_i - \boldsymbol{T}_1)} \ , i = 1,2, \dots, n.$$

Selanjutnya ambil H_2 sedemikian sehingga $\{d_1(i); i \in H_2\} \coloneqq \{(d_1)_{1:n}, (d_1)_{2:n}, \dots, (d_1)_{h:n}\}$, di mana $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{n:n}$ menyatakan urutan jarak, dan hitung $\boldsymbol{T_2}$ dan $\boldsymbol{C_2}$ berdasarkan H_2, maka

$$det(\boldsymbol{C_2}) \leq det(\boldsymbol{C}_1)$$

dan akan sama jika dan hanya jika $\boldsymbol{T}_1 = \boldsymbol{T_2}$ dan $\boldsymbol{C}_1 = \boldsymbol{C_2}$ (Driessen, 1999).

Untuk pengamatan $\{w_1, w_2, \dots, w\} \in \mathbb{R}^p$, MCD ditentukan dengan memilih subset dari observasi $\{w_{i_1}, w_{i_2}, \dots, w_{i_h}\}$ dari ukuran h, dengan $1 \leq h \leq n$ di mana memiliki determinan matriks kovarian yang minimum.

3.3.2 *Affine Equivariance*

Affine Equivariance memiliki implikasi bahwa suatu estimator dapat bertransformasi dengan baik dalam suatu tranformasi linier nonsingular. Sehingga meskipun data dirotasi atau ditranslasi tidak akan memiliki pengaruh pada pendeteksian pencilan (Hubert, 2009). Transformasi linier yang dimaksud adalah pada analisis diskriminan, analisis faktor, analisis korelasi kanonik (Yohai, 2006).

Estimator MCD dari rata-rata dan kovarian merupakan *affine equivariance*. Maksud dari *affine equivariance* adalah untuk suatu matriks nonsingular $\boldsymbol{A}$ dan vektor konstan $\boldsymbol{b} \in \mathbb{R}^d$

$$\boldsymbol{T}_{MCD}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{AT}_{MCD}(\boldsymbol{X}) + \boldsymbol{b}$$

$$\boldsymbol{C}_{MCD}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{AC}_{MCD}(\boldsymbol{X})\boldsymbol{A}^t.$$

3.3.3 *Breakdown Point*

Breakdown point adalah jumlah pengamatan minimal yang dapat menggantikan sejumlah pengamatan awal yang berakibat pada nilai taksiran yang dihasilkan sangat berbeda dari taksiran sebenarnya (Lopuhaa dan Rousseeuw, 1991). *Brekdown point* juga merupakan alat untuk mengukur ke*robust*an dari suatu penaksir.