BAB III

MINIMUM VOLUME ELLIPSOID

PADA ANALISIS KOMPONEN UTAMA ROBUST

Pada bab ini akan dikaji bahasan utama yaitu pencilan dan analisis komponen utama robust sebagai konsep pendukung serta metode *Minimum Volume Ellipsoid* sebagai konsep utama pada tugas akhir ini.

3.1. Analisis Komponen Utama Robust

Analisis komponen utama merupakan metode yang baik untuk menerangkan variasi total dari variabel-variabel yang mewakili suatu data melalui sebagian kecil komponen utama. Namun ketika terdapat pencilan dalam data, analisis komponen utama klasik yang diperoleh dapat menyimpang dari hasil yang diharapkan.

3.1.1. Pencilan

Dalam Hampel et al (1986: 1), pencilan didefinisikan sebagai observasi yang terletak sangat jauh dari bagian terbesar data, dan membahayakan beberapa prosedur statistik klasik. Dengan kata lain, pencilan merupakan penyimpangan beberapa observasi dalam suatu data yang terletak cukup jauh dari pusat observasi. Adapun jenis-jenis pencilan yaitu:

1. *Good leverage* merupakan observasi yang berada di ruang distribusi tetapi sudah tidak berada di daerah mayoritas data.

2. *Bad leverage* merupakan observasi yang tidak berada baik dalam ruang distribusi observasi maupun daerah mayoritas data.
3. Pencilan ortogonal merupakan observasi yang mempunyai jarak yang sangat besar dari daerah mayoritas data sehingga observasi tersebut sudah tidak dapat dilihat dalam ruang distribusinya.

Pendeteksian pencilan sangat penting dilakukan dalam prosedur statistik karena pencilan dapat mempengaruhi informasi yang terdapat di dalam data. Sebagai contoh, suatu data berat badan (dalam kilogram) dari lima orang siswa dalam kelompok A masing-masing 43, 39, 39, 38, dan 41, mean dari data tersebut adalah 40 kg. Jika kemudian ada seorang siswa baru masuk ke dalam kelompok tersebut dan dia memiliki berat badan 100 kg, maka mean berat badan dalam kelompok A menjadi 50 kg. Hasil mean akhir tersebut tidak dapat mewakili informasi sebenarnya, karena mayoritas siswa dalam kelompok A memiliki berat badan di bawah 50 kg. Hal ini menunjukkan bahwa pencilan dapat mempengaruhi keseimbangan data.

Pencilan yang terdeteksi dari suatu sampel mungkin saja dihilangkan dari data observasi agar dapat dilakukan analisis lebih lanjut. Tindakan tersebut dapat dilakukan apabila hanya terdapat satu buah pencilan, namun hal ini tidak mungkin dilakukan jika pencilan yang terdeteksi lebih dari satu, karena pencilan dapat mengandung informasi yang penting dari suatu data. Oleh karena itu, dalam mengatasi pencilan ini dibutuhkan suatu metode penaksir robust yang tangguh terhadap pencilan sehingga analisis komponen utama yang dilakukan tidak lagi dipengaruhi pencilan.

3.1.2. Jarak Mahalanobis

Umumnya, metode yang digunakan untuk mendeteksi pencilan dalam analisis multivariat berdasarkan pada jarak Mahalanobis. Bentuk dari persamaan jarak Mahalanobis adalah

$$d_i = \{(\boldsymbol{x}_i - \boldsymbol{a})\boldsymbol{S}^{*-1}(\boldsymbol{x}_i - \boldsymbol{a})^t\}^{1/2} \tag{3.1}$$

dengan $\boldsymbol{x}_i$ adalah vektor observasi, $\boldsymbol{a}$ adalah estimasi mean, dan $\boldsymbol{S}^*$ adalah estimasi dari matriks varians kovarians. Bentuk persamaan ini sama dengan T^2 Hotelling untuk vektor observasi individual, perbedaannya yaitu bahwa $\boldsymbol{a}$ dan $\boldsymbol{S}^*$ bukan lagi penaksir standar tetapi merupakan bagian dari proses robust. Kebanyakan teknik yang dipakai pada proses robust merupakan iterasi sehingga $\boldsymbol{a}$ dan $\boldsymbol{S}^*$ dapat berubah pada setiap iterasi. Jarak Mahalanobis ini dipakai untuk memberi indeks $w(d_i)$ yang digunakan untuk menghasilkan $\boldsymbol{a}$ dan $\boldsymbol{S}^*$

$$\boldsymbol{a} = \sum_{i=1}^{n} w_a(d_i)\frac{x_i}{\sum_{i=1}^{n} w_a(d_i)} \tag{3.2}$$

$$\boldsymbol{S}^* = \frac{\sum_{i=1}^{h} w_s(d_i)[\boldsymbol{x}_i - \boldsymbol{a}]^t[\boldsymbol{x}_i - \boldsymbol{a}]}{f\{w_s(d_i)\}}. \tag{3.3}$$

Definisi 3.2.1 Jarak Mahalanobis (Chen Y, Chen X, Xu, 2008: 222)

Untuk matriks data $\boldsymbol{X}_{nxp}$ dengan n observasi dan p variabel, jarak Mahalanobis dihitung sebagai

$$MD^2(\boldsymbol{x_i}, \boldsymbol{X}) = \big(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X})\big)\boldsymbol{S}(\boldsymbol{X})^{-1}(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X}))^t$$

untuk masing-masing observasi, dengan **T(X)** adalah penaksir lokasi multivariat (dalam kasus ini merupakan *arithmetic mean*) dan $\boldsymbol{S}(\boldsymbol{X})$ adalah penaksir varians kovarians klasik. Titik-titik dengan $MD^2(\boldsymbol{x_i}, \boldsymbol{X})$ yang besar diidentifikasi sebagai

titik-titik tak beraturan (pencilan) dan ditaksir menggunakan distribusi χ^2 dengan derajat kebebasan yang sesuai.

3.1.3. Estimasi Robust dengan Penaksir *Affine Equivariant*

Dalam Rousseeuw dan Leroy (1987: 248), sebagai penaksir lokasi multivariat, persamaan untuk T yang merupakan *translation equivariant* dapat dituliskan sebagai

$$T(\boldsymbol{x}_1 + \boldsymbol{b}, \dots, \boldsymbol{x}_n + \boldsymbol{b}) = T(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \tag{3.4}$$

untuk beberapa vektor $\boldsymbol{b}$ berdimensi p. Persamaan tersebut juga disebut sebagai *location equivariance*. Sedangkan penaksir lokasi multivariat adalah mean aritmetik

$$\boldsymbol{T}(\boldsymbol{X}) = \bar{\boldsymbol{x}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_\mathrm{i} \tag{3.5}$$

yang merupakan penaksir *least square* dalam kerangka ini karena persamaan tersebut akan meminimalkan $\sum_{\mathrm{i}=1}^{\mathrm{n}} \|\mathbf{x}_\mathrm{i} - \mathbf{T}\|^2$ dengan $\|\dots\|$ adalah norm Euclid.

T disebut *affine equivariant* jika dan hanya jika

$$T(\boldsymbol{x}_1\boldsymbol{A} + \boldsymbol{b}, \dots, \boldsymbol{x}_n\boldsymbol{A} + \boldsymbol{b}) = T(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)\boldsymbol{A} + \boldsymbol{b} \tag{3.6}$$

untuk semua vektor $\boldsymbol{b}$ dan sebarang matriks nonsingular $\boldsymbol{A}$.

Untuk penaksir varians kovarians, *affine equivariance* berarti bahwa

$$\boldsymbol{C}(\{\boldsymbol{x}_1\boldsymbol{A} + \boldsymbol{b}, \dots, \boldsymbol{x}_n\boldsymbol{A} + \boldsymbol{b}\}) = \boldsymbol{A}^t\boldsymbol{C}(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\})\boldsymbol{A} \tag{3.7}$$

dengan $\boldsymbol{A}$ adalah matriks nonsingular berukuran pxp dan $\boldsymbol{b}$ merupakan vektor. Pada $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, maximum likelihood menghasilkan penaksir equivariant

$$\boldsymbol{T}(\boldsymbol{X}) = \bar{\boldsymbol{x}}$$

$$\boldsymbol{S}(\boldsymbol{X}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathbf{x}_{\mathrm{i}} - \mathbf{T}(\mathbf{X})\right)^{\mathrm{t}}\left(\mathbf{x}_{\mathrm{i}} - \mathbf{T}(\mathbf{X})\right). \tag{3.8}$$

Untuk menghasilkan penaksir tak bias dari $\boldsymbol{\Sigma}$, penyebut pada persamaan $\boldsymbol{S}(\boldsymbol{X})$ dari n diganti dengan $n-1$, dan $\boldsymbol{x_i}$ menjadi vektor sampel ke-i yang dihitung sebagai

$$\boldsymbol{T}(\boldsymbol{X}) = \bar{\boldsymbol{x}}_{\boldsymbol{i}}$$

$$\boldsymbol{S}(\boldsymbol{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X})\right)^{t}\left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X})\right). \tag{3.9}$$

Untuk menggeneralisasi pendekatan maksimum likelihood, digunakan langkah iteratif *affine equivariant* untuk menaksir $\boldsymbol{\mu}$ dan $\boldsymbol{\Sigma}$. Definisi solusi simultan dari sistem persamaan

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_1\left([(\boldsymbol{x}_i - \boldsymbol{T})\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{T})^t]^{\frac{1}{2}}\right)(\boldsymbol{x}_i - \boldsymbol{T}) = \mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_2((\boldsymbol{x}_i - \boldsymbol{T})\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{T})^t)\,(\boldsymbol{x}_i - \boldsymbol{T})^t(\boldsymbol{x}_i - \boldsymbol{T}) = \boldsymbol{S}. \end{cases} \tag{3.10}$$

Sebagai ukuran tingkat kerobustan suatu metode terhadap keberadaan pencilan digunakan *breakdown point*. *Breakdown point* untuk mean adalah 1 yang artinya dengan hanya menggantikan 1 nilai ekstrim pada data asal, maka akan didapati perubahan mean yang sangat besar (Suryana, 2008). *Breakdown point* merupakan persentase terkecil pada penyimpangan yang dapat memiliki efek besar pada penaksir. *Breakdown point* pada semua *affine equivariant* yang banyak dipakai adalah $1/(p+1)$ atau sebesar 50% *breakdown point* (Rousseeuw dan Leroy, 1987: 253).

3.2. Pendeteksian Pencilan Menggunakan Metode MVE

Penggunaan jarak Mahalanobis klasik pada data multivariat berdasarkan mean sampel dan matriks varians kovarians sampel terkadang tidak efektif ketika berhadapan dengan sekelompok pencilan, dengan kata lain data masih dipengaruhi oleh pencilan. Akibatnya, jika terdapat pencilan dalam data, pencilan tersebut dapat mempengaruhi mean sampel dan matriks varians kovarians sampel sedemikian sehingga pencilan-pencilan tersebut memiliki jarak Mahalanobis yang kecil. Karena itu, pencilan tetap tidak terdeteksi. Hal ini disebut dengan efek *masking*. Rousseeuw dan Leroy (1987: 256) mengusulkan untuk menghitung jarak robust berdasarkan penaksir lokasi dan *scatter* dengan *high breakdown point*.

3.2.1. MVE dan Resampling

MVE yang diperkenalkan oleh Rousseeuw merupakan penaksir robust dengan *high breakdown* pertama pada lokasi dan *scatter* multivariat yang sering digunakan. MVE terkenal karena kepekaannya terhadap pencilan yang membuatnya dapat diandalkan untuk mendeteksi pencilan dan tersedia secara luas, serta implementasi MVE mudah dioperasikan pada algoritma perhitungannya (Rousseeuw dan Van Aelst, 2009: 1).

Menurut Rousseeuw, Croux, dan Haesbroeck (2002: 192), MVE dapat didefinisikan sebagai elipsoid terkecil yang mencakup paling sedikit h elemen pada himpunan $\boldsymbol{X} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \mathbb{R}^p$, dengan penaksir lokasi MVE merupakan pusat elipsoid dan penaksir *scatter* MVE berkoresponden menjadi matriks bentukan (*shape matrix*).

Definisi 3.2.1.1 (Rousseeuw dan Van Aelst, 2009: 72)

Pasangan $(\boldsymbol{T}, \boldsymbol{S})$ dengan penaksir lokasi $\boldsymbol{T}(\boldsymbol{X})$ dan penaksir *scatter* $\boldsymbol{S}(\boldsymbol{X})$ meminimalkan determinan $\boldsymbol{S}$ pada kondisi

$$\{i|(\boldsymbol{X}_i - \boldsymbol{T})\boldsymbol{S}^{-1}(\boldsymbol{X}_i - \boldsymbol{T})^t \leq c^2\} \geq h$$

dengan n adalah jumlah observasi, p adalah jumlah variabel, $h = [(n + p + 1)/2]$, c adalah konstanta dan $\boldsymbol{X}$ adalah himpunan data observasi. Proses meminimalkan tersebut berada pada setiap $\boldsymbol{T} \epsilon \mathbb{R}^p$ dan $\boldsymbol{S} \epsilon PDS(p)$. $PDS(p)$ adalah himpunan matriks simetri definit positif berukuran pxp.

Nilai c adalah konstanta tetap yang dipilih untuk menentukan besar dari $\boldsymbol{S}_n$. Umumnya, c dipilih sedemikian sehingga $\boldsymbol{S}_n$ merupakan penaksir konsisten pada matriks varians kovarians untuk data dari distribusi normal multivariat, dengan kata lain $c = \sqrt{\chi^2_{p,\alpha}}$ dengan $\alpha = h/n$. Dari definisi ini, jelas bahwa MVE menaksir pusat dan *scatter* untuk h lebih memusatkan observasi dalam data. Nilai h dapat dipilih dan menentukan ke-robust-an untuk menghasilkan estimasi MVE.

Persamaan standar untuk nilai h yang sering dipakai dinyatakan dengan $h = \left[\frac{n+p+1}{2}\right] \approx \frac{n}{2}$ karena ini dapat menghasilkan *breakdown point* yang maksimal. Tetapi jika diketahui pecahan pencilan sebagian besar berada pada $0 < \alpha < \frac{1}{2}$, maka dapat digunakan estimator MVE (α) dengan $h = [n(1 - \alpha)]$. Untuk mendapatkan hasil yang baik, nilai $\alpha = \frac{1}{4}$ adalah pilihan yang tepat.

MVE memiliki *breakdown point* mendekati 50%, yang berarti bahwa $\boldsymbol{T}(\boldsymbol{X})$ dapat tetap terbatas dan besar nilai eigen dari $\boldsymbol{S}(\boldsymbol{X})$ akan tetap jauh dari nol

dan tidak terhingga ketika kurang dari separuh dari jumlah data digantikan oleh nilai-nilai sebarang (Rousseeuw dan Van Aelst, 1991: 196).

Definisi 3.2.1.2 (Rousseeuw dan Van Aelst, 2009: 73)

Jarak robust $RD(x_i)$ yang berhubungan dengan MVE didefinisikan

$$RD(x_i) = \sqrt{(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X}))\boldsymbol{S}(\boldsymbol{X})^{-1}(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X}))^t}, \qquad i = 1, \dots, n.$$

Karena dalam mencari solusi eksak untuk menyelesaikan masalah pada MVE agak sulit, sebagai alternatifnya akan dicari pendekatan MVE menggunakan algoritma resampling yang disebut juga algoritma $(p+1)$-subset. Algoritma ini memiliki keuntungan dalam penaksirannya karena masih bersifat *affine equivariant* dan masih mempertahankan *high breakdown point*. Dalam tugas akhir ini digunakan adaptasi dari penaksir $(p+1)$-subset untuk MVE dengan *high breakdown point*. Gambaran utama dari algoritma resampling ini yaitu dengan merata-ratakan lebih dari beberapa nilai percobaan dibandingkan dengan hanya mengambil salah satu yang optimal.

Penaksir dasar dalam penentuan lokasi dan *scatter* multivariat adalah *affine equivariance*, yang berarti bahwa penaksir menunjukkan dengan baik di bawah transformasi *affine* pada data. Sehingga, penaksir $\boldsymbol{T}$ dan $\boldsymbol{S}$ pada lokasi dan *scatter* multivariat yang *affine equivariant* berikut pada beberapa matriks data $\boldsymbol{X}$

$$\boldsymbol{T}(\boldsymbol{XA} + \mathbf{1}_n\boldsymbol{v}^t) = \boldsymbol{A}^t\boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{v}$$

$$\boldsymbol{S}(\boldsymbol{XA} + \mathbf{1}_n\boldsymbol{v}^t) = \boldsymbol{A}^t\boldsymbol{S}(\boldsymbol{X})\boldsymbol{A} \qquad (3.11)$$

untuk semua matriks nonsingular pxp $\boldsymbol{A}$ dan $\boldsymbol{v}\epsilon\mathbb{R}^p$. Vektor $\mathbf{1}_n = (1,1,\dots,1)^t\epsilon\mathbb{R}^n$. *Affine equivariance* dalam penaksir cukup penting karena dapat membuat analisis

menjadi independen pada skala pengukuran variabel baik dalam translasi maupun rotasi pada data.

3.2.2. Metode MVE

Dari definisi 3.2.1.1 dapat diketahui bahwa perhitungan MVE yang tepat untuk himpunan $\boldsymbol{X}_n$ akan menuntut pengujian semua $\binom{n}{h}$ elipsoid yang mengandung h observasi pada $\boldsymbol{X}_n$ untuk mencari elipsoid dengan volume terkecil. Penyelesaian masalah kombinatorial ini mungkin dilakukan jika himpunan data kecil dan berdimensi rendah. Karena jumlah elipsoid umumnya sangat besar, maka untuk menyelesaikan persoalan pada data yang berdimensi tinggi dilakukan pendekatan algoritma.

Patokan algoritma MVE membatasi pencarian elipsoid yang ditentukan oleh subset yang memuat $p+1$ (p merupakan jumlah variabel pada matriks $\boldsymbol{X}_n$) observasi yang berbeda diberikan indeks $\boldsymbol{J} = \{i_i, i_2, \dots, i_{p+1}\} \subset \{1, \dots, n\}$, kemudian berdasarkan persamaan (3.10), hitung mean dan matriks varians kovarians sampel sehingga persamaan menjadi

$$\boldsymbol{T}_{\mathcal{H}} = \bar{\boldsymbol{x}}_{\boldsymbol{J}} = \frac{1}{p+1} \sum_{i \in \mathcal{H}}^{p+1} \boldsymbol{x}_i \tag{3.12}$$

$$\boldsymbol{S}_{\boldsymbol{J}} = \frac{1}{p} \sum_{i \in \boldsymbol{J}}^{p+1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\boldsymbol{J}})^t (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\boldsymbol{J}}) \tag{3.13}$$

dengan matriks varians kovarians $\boldsymbol{S}_{\boldsymbol{J}}$ non singular untuk $\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \dots, \boldsymbol{x}_{i_{p+1}}$. Jika $p+1$ bukan pada posisi umum, maka observasi pada $\boldsymbol{X}_n$ ditambahkan hingga subset dengan matriks varians kovarians sampel nonsingular diperoleh (atau

sebuah subsampel singular berukuran h diperoleh). Elipsoid ditentukan oleh $\boldsymbol{T_J}$ dan $\boldsymbol{S_J}$ yang kemudian dinaikkan atau diturunkan hingga mengandung tepat h titik. Sedangkan faktor penskalaan diberikan oleh persamaan D_J^2/c^2 dengan $c = \sqrt{\chi^2_{p,\alpha}}$ dan

$$D_J^2 = [(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_J)\boldsymbol{C}_J^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_J)^t]_{h:n} \tag{3.14}$$

dengan $h = (n + p + 1)/2$ dan $h{:}\,n$ menunjukkan jarak kuadrat terkecil ke-h di antara jarak kuadrat pada n observasi dalam $\boldsymbol{X}_n$. Elipsoid yang dihasilkan dapat memenuhi definisi 3.4.1.1 dan volumenya proporsional dengan menggunakan persamaan

$$P_J = \left(\det\left(D_J^2 \boldsymbol{S_J}\right)\right)^{0,5}. \tag{3.15}$$

Ulangi langkah-langkah tersebut sebanyak subsampel **J**, dan pilih salah satu dengan P_J terendah. Kemudian untuk subsampel $\boldsymbol{J}$ yang telah diperoleh, hitung

$$\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T_J}$$

dan

$$\boldsymbol{S}(\boldsymbol{X}) = c^2(n,p)\left(\chi^2_{p,\alpha}\right)^{-1} D_J^2 \boldsymbol{S_J} \tag{3.16}$$

dengan $c^2(n,p)$ adalah *correction term* sampel kecil yang dihitung sebagai $\left[1 + \frac{15}{n-p}\right]^2$ dan $\chi^2_{p,\alpha}$ merupakan median dari distribusi χ^2 dengan derajat kebebasan p. Hal ini menunjukkan bahwa sampling intensif dan perhitungan membutuhkan pencarian solusi dari analisis MVE. Total jumlah subsampel bergantung pada n dan p. Berdasarkan estimasi MVE mean $\boldsymbol{T}(\boldsymbol{X})$ dan varians kovarians $\boldsymbol{S}(\boldsymbol{X})$, pemberian indeks dapat dihitung menggunakan

$$W_i = \left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X})\right)\boldsymbol{S}(\boldsymbol{X})^{-1}\left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X})\right)^t \quad (3.17)$$

untuk observasi i, dengan $W_i > \chi^2_{p;\alpha}$ didefinisikan sebagai pencilan. PCA yang digambarkan dengan pencilan yang didefinisikan melalui MVE memiliki indeks 0 dan data normal diberikan indeks 1.

Dari uraian tersebut, algoritma pendeteksian pencilan pada metode MVE adalah:

1. Bentuk subsampel yang mengandung p+1 observasi disimbolkan dengan indeks $\boldsymbol{J}$ sebanyak $\binom{n}{h}$.
2. Hitung mean dan matriks varians kovarians koresponden dengan $\boldsymbol{S_J}$ nonsingular.
3. Hitung $D_J^2 = [(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\boldsymbol{J}})\boldsymbol{S_J^{-1}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\boldsymbol{J}})^{\boldsymbol{t}}]_{h:n}$
4. Hitung $P_J = \left(\det\left(D_J^2 \boldsymbol{S_J}\right)\right)^{0,5}$ untuk menghasilkan elipsoid.
5. Ulangi langkah 1-4 sebanyak subsampel $\boldsymbol{J}$, kemudian pilih P_J dengan nilai terendah.
6. Hitung $\boldsymbol{T}(\boldsymbol{X})$ dan $\boldsymbol{S}(\boldsymbol{X}) = c^2(n,p)\left(\chi^2_{p;\alpha}\right)^{-1} D_J^2 \boldsymbol{S_J}$ dari P_J yang akan digunakan sebagai estimasi mean dan matriks varians kovarians.
7. Observasi dengan $W_i > \chi^2_{p;\alpha}$ diidentifikasi sebagai pencilan.

3.3. *Weighted* PCA

Untuk menangani masalah pencilan, selanjutnya digunakan *Weighted Principal Component Analysis* (WPCA) yang diperkenalkan oleh Joaquim F. Pinto da Costa et al (2011) dalam jurnal *"A Weighted Principal Component Analysis and Its Application to Gene Expression Data"*. PCA klasik tidak cukup tepat untuk menangani masalah pencilan. Dalam tugas akhir ini akan digunakan koefisien korelasi baru sebagai alternatif korelasi Pearson yang disebut *Weighted* PCA (WPCA).

Pada PCA klasik, vektor-vektor eigen dari matriks varians kovarians atau matriks korelasi Pearson memuat koefisien-koefisien dari kombinasi linier dari matriks asli yang berkorespondensi dengan variabel-variabel baru (komponen-komponen utama). Seperti telah diketahui koefisien korelasi Pearson sangat sensitif terhadap kehadiran pencilan dan gangguan. Untuk menangani hal ini, akan digunakan *rank* (rangking) dari setiap observasi, yang dimulai dengan merangking observasi untuk setiap variabel dari satu (rangking tertinggi) sampai n (rangking terendah).

Misalkan diberikan data bivariat (X_i, Y_i), $i = 1,2, \dots, n$, beri nama R_i untuk rangking variabel X_i dan Q_i untuk rangking variabel Y_i. Jika akan menghitung koefisien korelasi Pearson untuk rangking data, maka yang diperoleh adalah koefisien korelasi rank Spearman (r_S), yaitu :

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(Q_i - \bar{Q})^2}} \tag{3.18}$$

dengan $\bar{R}$ dan $\bar{Q}$ adalah rata-rata *rank*. (Rachmatin, 2011: 2)

R_i dan Q_i untuk merepresentasikan *rank* dari dua variabel yang berkorespondensi dengan observasi (sampel) i. Untuk tujuan komputasi, persamaan dalam mencari koefisien korelasi *rank* adalah

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n^3 - n} \qquad (3.19).$$

Definisi 3.2.3.1 (da Costa, 2011: 247)

$$W_2 D_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2$$

yang mewakili lebih dari WD_i^2 penyesuaian yang lebih penting pada *rank-rank* atas. Biasanya koefisien korelasi *rank* didefinisikan seperti koefisien Spearman, sebagai fungsi linier dari jarak antara dua vektor *rank*.

Da Costa mengusulkan bahwa transformasi mengandung pensubstitusian nilai observasi i pada variabel pertama dengan nilai $R_i' = R_i(2n + 2 - R_i)$, dengan R_i merupakan *rank* dari observasi. Begitu pula untuk variabel lainnya, diperoleh dari $Q_i' = Q_i(2n + 2 - Q_i)$. Untuk menghitung rata-ratanya digunakan persamaan berikut

$$\overline{R'} = \frac{1}{n}\sum_{i=1}^{n} R_i(2n + 2 - R_i) = \frac{(n+1)(4n+5)}{6}$$

$$\overline{Q'} = \frac{1}{n}\sum_{i=1}^{n} Q_i(2n + 2 - Q_i) = \frac{(n+1)(4n+5)}{6} \qquad (3.20)$$

Algoritma WPCA yang diperkenalkan da Costa (Rachmatin, 2011) adalah sebagai berikut:

Misalkan diberikan data matriks $\boldsymbol{X}$ dengan ukuran nxp (n observasi dan p variabel).

1. Transformasi $\boldsymbol{X}$ menjadi $\boldsymbol{R}$ (nxp) → matriks *rank* untuk data.

$\boldsymbol{X} = [x_{ij}]$ $\quad i = 1,2, \dots, n$

$\quad j = 1,2, \dots, p$

$\boldsymbol{R} = [R_{ij}]$ dengan R_{ij} = *rank* observasi ke-i kolom ke-j.

2. Transformasi $\boldsymbol{R}$ menjadi $\boldsymbol{R}'$ dengan $\boldsymbol{R}' = [R_{ij}{}']$

$$R_{ij}{}' = R_{ij}(2n + 2 - R_{ij})$$

3. Standardisasi $R_{ij}{}'$ dengan transformasi

$$R_{ij}{}' = \frac{R_{ij}{}' - \bar{R}_j'}{S_{R_j}\sqrt{n}}$$

dengan $\bar{R}_j' = \frac{1}{n}\sum_{i=1}^{n} R_{ij}$

$$S_{R_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_{ij}{}' - R_{ij}\right)}$$

4. Jika $\boldsymbol{X}'$ menyatakan matriks data hasil transformasi langkah ke-3, hitung $(\boldsymbol{X}')^t.\boldsymbol{X}$ --- "r_{w2}".

Tentukan nilai eigen dan vektor eigen dari $(\boldsymbol{X}')^t.\boldsymbol{X}$.