BAB 2
KAJIAN PUSTAKA DAN DASAR TEORI

2.1 Analisis Diskriminan

Diskriminan merupakan metode analisis multivariat yang bertujuan untuk memisahkan objek pengamatan yang berbeda dan mengalokasikan objek pengamatan baru ke dalam kelompok yang telah didefinisikan (Johnson dan Wichern, 2002: 581). Misal sebuah populasi Ω terdiri dari l kelompok $\pi_1, \ldots, \pi_l$ dengan masing-masing wilayah (*region*) $R_1, \ldots, R_l$. Suatu pengukuran terdiri dari p variabel prediktor, dilakukan pada l kelompok sebanyak n pengamatan, menghasilkan matrik data $\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right]^T$ dengan $\mathbf{x}_i = \left\{x_{i1}, ..., x_{ip}\right\}; i = 1, ..., n.$ Perbedaan l kelompok dapat diketahui dari bentuk densitasnya, $f_j(\mathbf{x})$ bila pengamatan berasal dari π_j ; j = 1, . . . , l dengan probabilitas prior p_j. Besarnya biaya yang harus dikeluarkan bila objek pengamatan yang berasal dari π_j dinyatakan sebagai π_k di mana $k \in \{j|\ j$ = 1, . . . , l } dinotasikan oleh $c(k|j)$ dengan probabilitas,

$$P(k \mid j) = P\left(\mathbf{X} \in R_k \mid \pi_j\right) = \int_{R_k} f_j(\mathbf{x})\, d\mathbf{x}. \tag{2.1}$$

Ekspektasi biaya salah pengelompokkan sebuah objek pengamatan $\mathbf{x}$ dari π_j dinyatakan sebagai π_k (*expected cost of misklassification)* disingkat ECM adalah:

$$ECM = \sum_{j=1}^{l} p_j \left(\sum_{\substack{k=1 \\ k \neq j}}^{l} P\left(k \mid j\right) c\left(k \mid j\right) \right). \tag{2.2}$$

Teorema 2.1 (Johnson dan Wichern, 2002: 614)

Diketahui suatu matrik data $\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right]^T$ terdiri dari p variabel dengan $\mathbf{x}_i = \left\{x_{i1}, ..., x_{ip}\right\}; i = 1, ..., n$ di mana p_j menyatakan probabilitas prior, $c(k|j)$ menyatakan biaya salah pengelompokkan dan

$f_j(\mathbf{x})$ menyatakan fungsi densitas kelompok ke-*j*. Jika $\boldsymbol{x} \in R^p$ dikelompokkan ke $\pi_k; k = 1,...,l$ dengan

$$\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j f_j(\mathbf{x}) c(k|j) \tag{2.3}$$

minimum maka *ECM* (2.2) akan minimum.

Bukti

ECM pada (2.2) dapat dituliskan sebagai

$$ECM = p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} P(j|k) c(j|k) + \sum_{\substack{j=1 \\ j\neq k}}^{l} p_j \sum_{\substack{k=1 \\ k\neq j}}^{l} P(k|j) c(k|j).$$

Substitusikan (2.1) ke (2.2) akan dihasilkan

$$ECM = p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) \int_{R_j} f_k(\mathbf{x}) d\mathbf{x} + \sum_{\substack{j=1 \\ j\neq k}}^{l} p_j \sum_{\substack{k=1 \\ k\neq j}}^{l} c(k|j) \int_{R_k} f_j(\mathbf{x}) d\mathbf{x}$$

Karena $\Omega = \bigcup_{j=1}^{l} R_j$, maka

$\int_{\Omega} f_k(\mathbf{x}) d\mathbf{x} = \sum_{j=1}^{l} \int_{R_j} f_k(\mathbf{x}) d\mathbf{x} = 1$. Sehingga $\sum_{\substack{j=1 \\ j\neq k}}^{l} \int_{R_j} f_k(\mathbf{x}) d\mathbf{x} = 1 - \int_{R_k} f_k(\mathbf{x}) d\mathbf{x}$. Akibatnya,

$$ECM = p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) \left[1 - \sum_{\substack{k=1 \\ k\neq j}}^{l} \int_{R_k} f_k(\mathbf{x}) d\mathbf{x} \right] + \sum_{\substack{j=1 \\ j\neq k}}^{l} p_j \sum_{\substack{k=1 \\ k\neq j}}^{l} c(k|j) \int_{R_k} f_j(\mathbf{x}) d\mathbf{x}$$

$$= p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) - p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) \sum_{\substack{k=1 \\ k\neq j}}^{l} \int_{R_k} f_k(\mathbf{x}) d\mathbf{x} + \sum_{\substack{j=1 \\ j\neq k}}^{l} p_j \sum_{\substack{k=1 \\ k\neq j}}^{l} c(k|j) \int_{R_k} f_j(\mathbf{x}) d\mathbf{x}$$

Dengan sifat penjumlahan integral diperoleh,

$$= p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) + \int_{R_k} \left[\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j \sum_{\substack{k=1 \\ k\neq j}}^{l} c(k|j) f_j(\mathbf{x}) - \sum_{\substack{j=1 \\ j\neq k}}^{l} p_k \sum_{\substack{k=1 \\ k\neq j}}^{l} c(j|k) f_k(\mathbf{x}) \right] d\mathbf{x}$$

$$= p_k \sum_{\substack{j=1 \\ j\neq k}}^{l} c(j|k) + \int_{R_k} \sum_{\substack{k=1 \\ k\neq j}}^{l} \left[\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j c(k|j) f_j(\mathbf{x}) - \sum_{\substack{j=1 \\ j\neq k}}^{l} p_k c(j|k) f_k(\mathbf{x}) \right] d\mathbf{x}$$

Karena probabilitas prior, biaya salah pengelompokkan dan nilai-nilai integral suatu fungsi densitas tidak pernah negatif maka *ECM* akan minimum apabila:

$$\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j c\left(k|j\right) f_j\left(\mathbf{x}\right) - \sum_{\substack{j=1 \\ j\neq k}}^{l} p_k c\left(j|k\right) f_k\left(\mathbf{x}\right) \leq 0$$

atau

$$\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j c\left(k|j\right) f_j\left(\mathbf{x}\right) \text{minimum.} \qquad \square$$

Anggap semua biaya salah pengelompokkan sama. Pengamatan $\boldsymbol{x} \in R^p$ akan dialokasikan ke dalam kelompok π_k jika

$$\sum_{\substack{j=1 \\ j\neq k}}^{l} p_j f_j\left(\mathbf{x}\right) \tag{2.4}$$

minimum. Sementara itu, (2.4) akan minimum jika $p_k f_k\left(\mathbf{x}\right)$ maksimum. Dengan kata lain, Pengamatan $\boldsymbol{x} \in R^p$ akan dialokasikan ke dalam kelompok π_k jika

$$p_k f_k\left(\mathbf{x}\right) > p_j f_j\left(\mathbf{x}\right) \text{untuk semua } j \neq k \tag{2.5}$$

atau

$$\ln p_k f_k\left(\mathbf{x}\right) > \ln p_j f_j\left(\mathbf{x}\right) \text{untuk semua } j \neq k. \tag{2.6}$$

Apabila **x** berdistribusi normal multivariat dengan vektor rata-rata $\boldsymbol{\mu}_j$ dan matrik kovariansi $\boldsymbol{\Sigma}_j$, maka bentuk fungsi densitas **x** dapat dinyatakan sebagai,

$$f_j\left(\mathbf{x}\right) = \left[\left(2\pi\right)^p \left|\boldsymbol{\Sigma}_j\right|\right]^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^t \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)\right], \tag{2.7}$$

dengan j = 1, . . . , l. Jika biaya misklasifikasi diasumsikan sama maka pengamatan $\boldsymbol{x} \in R^p$ akan dialokasikan ke dalam kelompok π_k jika

$$\ln p_k - \left(\frac{p}{2}\right)\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right) >$$

$$\ln p_j - \left(\frac{p}{2}\right)\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_j\right| - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_j\right) \text{untuk semua } j \neq k.$$

Karena komponen $\left(\frac{p}{2}\right)\ln\left(2\pi\right)$ konstan, maka Johnsosn dan Wichern (2002: 616) mendefinisikan skor diskriminan kuadratik untuk setiap pengamatan pada kelompok ke-j sebagai

$$d_j^Q(\mathbf{x}) = -\frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^t \Sigma_j^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_j\right) + \ln\left(p_j\right) \tag{2.8}$$

dengan $j = 1, \ldots, l$. Dengan menggunakan skor diskriminan kuadratik pada (2.8), maka $\mathbf{x} \in R^p$ akan dialokasikan ke dalam kelompok π_k jika

$$d_k^Q(\mathbf{x}) = \text{maks di antara } d_1^Q(\mathbf{x}),\ \ldots,\ d_l^Q(\mathbf{x}). \tag{2.9}$$

Menurut Joossens (2006: 32), meskipun (2.8) diturunkan dari densitas normal multivariat namun skor diskriminan kuadratik dapat diterapkan tanpa melalui asumsi distribusi tertentu.

Kenyataannya, p_j, vektor mean $\boldsymbol{\mu}_j$ dan matrik kovarian $\boldsymbol{\Sigma}_j$ tidak diketahui. Untuk mengestimasi $\boldsymbol{\mu}_j$ dan $\boldsymbol{\Sigma}_j$ digunakan penaksir tak bias untuk kedua parameter tersebut yaitu

$$\bar{\mathbf{x}}_j = \frac{1}{n_j}\sum_{h=1}^{n_j}\mathbf{x}_{hj} \text{ dan } \mathbf{S}_j = \frac{1}{\left(n_j-1\right)}\sum_{h=1}^{n_j}\left(\mathbf{x}_{hj}-\bar{\mathbf{x}}_j\right)\left(\mathbf{x}_{hj}-\bar{\mathbf{x}}_j\right)^t. \tag{2.10}$$

$\mathbf{x} \in R^p$ akan dialokasikan ke dalam kelompok π_k jika:

$\hat{d}_k^{CQ}(\mathbf{x}) > \hat{d}_j^{CQ}(\mathbf{x})$ untuk semua $j = 1, \ldots, l, j \neq k$ dengan

$$\hat{d}_j^{CQ}(\mathbf{x}) = -\frac{1}{2}\ln\left|\mathbf{S}_j\right| - \frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}_j\right)^t \mathbf{S}_j^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_j\right) + \ln\left(\hat{p}_j^C\right). \tag{2.11}$$

Untuk mengestimasi probabilita keanggotaan p_j dalam (2.11), bisa digunakan dua pendekatan (Hubert, 2004:302). Pertama p_j diasumsikan konstan, sehingga $\hat{p}_j^C = 1/l$ untuk setiap j. Pendekatan kedua, p_j dinyatakan sebagai frekuensi relatif pengamatan dalam kelompok, sehingga $\hat{p}_j^C = n_l/n$.

2.2 Pendeteksian dan Bentuk *Outlier*

Outlier merupakan suatu pengamatan yang menyimpang cukup jauh dari pengamatan lainnya sehingga menimbulkan kecurigaan bahwa pengamatan tersebut berasal dari distribusi data yang berbeda (Hawkins dalam Sujatmiko, 2005:4). Distribusi pertama disebut sebagai "distribusi dasar" (*basic distribution*) yang menghasilkan pengamatan "baik". Distribusi kedua disebut sebagai "distribusi kontaminan" (*contaminating distribution*) yang menghasilkan pengamatan "tidak baik". Jumlah

maksumum *outlier* dalam data yang diperbolehkan adalah 50 persen (Rousseeuw dan Leroy dalam Hubert dan Van Driessen, 2004: 303).

Outlier yang disebabkan oleh variabel prediktor dinamakan *leverage*. *Leverage* sangat sulit diketahui sejak awal karena:

1. Visualisasi seperti *scatter diagram* tidak mampu menggambarkan secara utuh dalam satu gambar.
2. Beberapa pencilan dalam data membentuk efek *masking*.

Identifikasi outlier pada data multivariat umumnya didasarkan pada jarak kuadrat mahalanobis. Sebuah pengamatan $\mathbf{x}_i$ diidentifikasi sebagai *outlier* jika jarak mahalanobis,

$$d_{MD}^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^t \mathbf{S}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) > \chi_{p,(1-\alpha)}^2. \tag{2.12}$$

Di sini $\bar{\mathbf{x}}$ dan $\mathbf{S}$ menyatakan vektor rata-rata dan matrik kovariansi. Dengan jarak mahalanobis, identifikasi outlier tidak maksimal bila data mangandung lebih dari satu pengamatan *outlier*. Hal ini muncul akibat adanya pengaruh *masking* dan *swamping*. *Masking* terjadi pada saat pengamatan *outlier* tidak terdeteksi karena adanya pengamatan *outlier* lain yang berdekatan. *Swamping* terjadi saat pengamatan baik teridentifikasi sebagai pengamatan outlier.

Baik *masking* maupun *swamping* keduanya dapat diatasi dengan menggunakan penaksir *robust* untuk vektor rata-rata dan matrik kovariansi sehingga dihasilkan jarak kuadrat mahalanobis *robust*. Salah satu penaksir *robust* yang mempunyai kemampuan mengukur jarak sekaligus mendeteksi titik *leverage* adalah MCD. Deteksi *outlier* melalui *Robust Distance* (Hubert dkk, 2007). Sebuah pengamatan $\mathbf{x}_i$ diidentifikasi sebagai *outlier* jika jarak mahalanobis *robust*,

$$d_{RD}^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}_{MCD}\right)^t \mathbf{S}_{MCD}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}_{MCD}\right) > \chi_{p,(1-\alpha)}^2, \tag{2.13}$$

di mana $\bar{\mathbf{x}}_{MCD}$ dan $\mathbf{S}_{MCD}$ menyatakan vektor rata-rata dan matrik kovariansi dari sebagian data $\boldsymbol{X}$ yang mempunyai determinan matrik kovariansi terkecil.

Berdasarkan pengaruh pengamatan *outlier* terhadap data maka *outlier* dapat dibedakan menjadi tiga. Pertama, *shift outlier*. *Shift outlier* mampu menggeser vektor rata-rata sehingga pusat data menjadi berubah. Pada data berdistribusi normal, pergeseran vektor rata-rata bisa melalui penambahan setiap elemen vektor rata-rata dengan satuan $Q_p = \sqrt{\chi_{p,0.001}^2 / p}$ di mana p menyatakan jumlah variabel dan $\chi_{p,0.001}^2$ menyatakan nilai *chi-square* dengan derajat bebas p dan level konvidensi (1 – 0.001) . Menurut Rocke dan Woodruft dalam Todorov dan Pires (2007) penambahan Q_p pada setiap elemen vektor

rata-rata dari data berdistribusi normal multivariat sudah mampu menggeser pusat ellipsoid sejauh Q_p. Data terkontaminasi *shift outlier* dapat dinyatakan sebagai

$$\pi_j = (1-\varepsilon) N_p\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \varepsilon N_p\left(\boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j\right); j = 1, \ldots, l . \quad (2.14)$$

di mana ε menyatakan proporsi *outlier* dalam data dan $\boldsymbol{\mu}_j^*$ menyatakan vektor rata-rata yang berfungsi sebagai *shift outlier*.

Jenis *Outlier* berikutnya adalah *scale outlier*. Jika *shift outlier* hanya mampu menggeser pusat ellipsoid, maka *scale outlier* mampu merubah bentuk ellipsoid. *Scale outlier* dapat dinyatakan dengan persamaan

$$\pi_j = (1-\varepsilon) N_p\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \varepsilon N_p\left(\boldsymbol{\mu}_j, \kappa\boldsymbol{\Sigma}_j\right); j = 1, \ldots, l . \quad (2.15)$$

di mana $\kappa\boldsymbol{\Sigma}_j$ menyatakan matrik kovariansi yang berfungsi sebagai *scale outlier*.

Jenis *outlier* ketiga merupakan gabungan dua *outlier* sebelumnya. Hubert dan Van Driessen (2004) menyebutnya dengan *radial outlier*. *Radial outlier* ini selain menggeser pusat ellipsoid juga merubah bentuk ellipsoid. Dalam distribusi hierarki, *radial outlier* dinyatakan dengan

$$\pi_j = (1-\varepsilon) N_p\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \varepsilon N_p\left(\boldsymbol{\mu}_j^*, \kappa\boldsymbol{\Sigma}_j\right); j = 1, \ldots, l \quad (2.16)$$

2.3 Penaksir *Robust* MCD

Penaksir *robust* MCD merupakan rata-rata dan kovariansi dari sebagian pengamatan yang meminimumkan determinan matrik kovariansi. Menurut Hubert (2007: 5), MCD memiliki sifat statistik yang baik karena memenuhi sifat *affine equivariant*. MCD juga tergolong penaksir *robust* dengan *breakdown point* tingkat tinggi karena memenuhi batas nilai maksimum *breakdown* 50 persen. Dari sudut pandang ketersediaan paket program, MCD telah terakomodir dalam S-PLUS, Matlab dan SAS dengan menggunakan algoritma Fast-MCD.

Definisi 2.1 MCD (Butler dkk, 1993: 1385).

Diketahui $\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right]^T$ merupakan himpunan data sejumlah n pengamatan terdiri dari p variabel di mana $n \geq p+1$. Penaksir MCD merupakan pasangan $\mathbf{t} \in \mathbb{R}^p$ dan $\mathbf{C} \in PDS(p)$ matrik definit positif

simetri berdimensi *pxp* dari suatu sub sampel berukuran *h* pengamatan di mana $(n+p+1)/2 \le h \le n$ dengan

$$\mathbf{t} = (1/h)\sum_{i=1}^{h}\mathbf{x}_i \text{ dan } \mathbf{C} := (1/h)\sum_{i=1}^{h}(\mathbf{x}_i - \mathbf{t}_1)(\mathbf{x}_i - \mathbf{t}_1)^T \quad (2.17)$$

yang meminimumkan $\det \mathbf{C}$.

Berdasarkan Definisi 2.1 di atas, metode MCD mencari himpunan bagian dari $\mathbf{X}$ sejumlah *h* elemen di mana *h* integer terkecil dari $(n+p+1)/2$. Misalkan himpunan bagian itu adalah $\mathbf{X}_h$. Terdapat ${}^nC_h$ kombinasi yang harus ditemukan untuk mendapatkan penaksir MCD. Untuk *n* kecil, penaksir MCD cepat ditemukan. Tetapi, jika *n* besar makan banyak sekali kombinasi sub sampel yang harus ditemukan untuk mendapatkan penaksir MCD.

Keterbatasan ini menghantarkan pada penemuan algoritma FAST-MCD oleh Rousseeuw dan Van Driessen (1999). Salah satu aspek terpenting dari algoritma FAST-MCD adalah teorema *C-Steps*.

Teorema 2.2 *C-Steps* (Rousseeuw dan Van Driessen, 1999: 214).

Diketahui $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$ merupakan himpunan data sejumlah *n* pengamatan terdiri dari *p* variabel. Misal $H_1 \subset \{1, ..., n\}$ dengan jumlah elemen H_1, $\#(H_1) = h$, tetapkan $\mathbf{t}_1 := (1/h)\sum_{i \in H_1}\mathbf{x}_i$ dan $\mathbf{C}_1 := (1/h)\sum_{i \in H_1}(\mathbf{x}_i - \mathbf{t}_1)(\mathbf{x}_i - \mathbf{t}_1)^T$. Jika $\det(\mathbf{C}_1) \ne 0$, definisikan jarak relatif

$d_1(i) = \sqrt{(\mathbf{x}_i - \mathbf{t}_1)^T \mathbf{C}_1^{-1}(\mathbf{x}_i - \mathbf{t}_1)}$ untuk $i = 1, ..., n$.

Selanjutnya ambil H_2 sedemikian sehingga $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n}, ..., (d_1)_{h:n}\}$, di mana $(d_1)_{1:n} \le (d_1)_{2:n} \le \cdots \le (d_1)_{n:n}$ menyatakan urutan jarak, dan hitung $\mathbf{t}_2$ dan $\mathbf{S}_2$ berdasarakan himpunan H_2. Maka

$\det(\mathbf{C}_2) \le \det(\mathbf{C}_1)$

dan akan sama jika dan hanya jika $\mathbf{t}_1 = \mathbf{t}_2$ dan $\mathbf{C}_2 = \mathbf{C}_1$.

Bukti.

Asumsikan $\det(\mathbf{C}_2) > 0$. Selanjutnya jarak relatif $d_2(i) = d_{(\mathbf{t}_2, \mathbf{C}_2)}(i)$ untuk semua $i = 1, ..., n$. Dengan menggunakan $\#(H_2) = h$ dan definisi $(\mathbf{t}_2, \mathbf{C}_2)$, diperoleh

$$\frac{1}{hp}\sum_{i\in H_2} d_2^2(i) = \frac{1}{hp}\operatorname{tr}\sum_{i\in H_2}(\mathbf{x}_i - \mathbf{t}_2)^T \mathbf{C}_2^{-1}(\mathbf{x}_i - \mathbf{t}_2)$$

$$= \frac{1}{hp}\operatorname{tr}\sum_{i\in H_2}\mathbf{C}_2^{-1}(\mathbf{x}_i - \mathbf{t}_2)(\mathbf{x}_i - \mathbf{t}_2)^T$$

$$= \frac{1}{p}\operatorname{tr}\mathbf{C}_2^{-1}\mathbf{C}_2 = \frac{1}{p}\operatorname{tr}(\mathbf{I}) = 1. \qquad (2.18)$$

Selanjutnya,

$$\lambda := \frac{1}{hp}\sum_{i\in H_2} d_1^2(i) = \frac{1}{hp}\sum_{i=1}^{h}\left(d_1^2\right)_{i:n} \le \frac{1}{hp}\sum_{j\in H_1} d_1^2(j) = 1, \qquad (2.19)$$

di mana $\lambda > 0$.

Dengan menggabungkan (2.18) dan (2.19) dihasilkan

$$\frac{1}{hp}\sum_{i\in H_2} d^2_{(\mathbf{t}_1, \lambda\mathbf{C}_1)}(i) = \frac{1}{hp}\sum_{i\in H_2}(\mathbf{x}_i - \mathbf{t}_i)^T \frac{1}{\lambda}\mathbf{C}_1^{-1}(\mathbf{x}_i - \mathbf{t}_1)$$

$$= \frac{1}{\lambda hp}\sum_{i\in H_2} d_1^2(i) = \frac{\lambda}{\lambda} = 1.$$

Akibatnya $\det(\lambda\mathbf{C}_1) \le \det(\mathbf{C}_1)$. Sementara dari (2.19) diperoleh pertidaksamaan $\det(\mathbf{C}_2) \le \det(\lambda\mathbf{C}_1)$. Sehingga

$$\det(\mathbf{C}_2) \le \det(\lambda\mathbf{C}_1) \le \det(\mathbf{C}_1). \qquad (2.20)$$

Lebih lanjut, $\det(\mathbf{C}_2) = \det(\mathbf{C}_1)$ jika dan hanya jika (2.20) menjadi bentuk persamaan. Pertama, $\det(\mathbf{C}_2) = \det(\lambda\mathbf{C}_1)$ jika dan hanya jika $(\mathbf{t}_2, \mathbf{C}_2) = (\mathbf{t}_1, \lambda\mathbf{C}_1)$. Kedua, $\det(\lambda\mathbf{C}_1) = \det(\mathbf{C}_1)$ jika dan hanya jika $\lambda = 1$. Akibatnya, $(\mathbf{t}_2, \mathbf{C}_2) = (\mathbf{t}_1, \mathbf{C}_1)$. □

Bentuk lain penaksir MCD adalah dengan menggunakan pembobot. Pengamatan yang tidak disertakan dalam penghitungan penaksir rata-rata dan kovariansi MCD diberi bobot nol, lainnya diberi bobot sama dengan satu. Penaksir MCD dihitung dengan

$$\mathbf{t}_{MCD} = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i} \tag{2.21}$$

dan

$$\mathbf{C}_{MCD} = \frac{\sum_{i=1}^{n} w_i \left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)\left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)^T}{\sum_{i=1}^{n} w_i - 1} \tag{2.22}$$

dengan

$$w_i = \begin{cases} 1 \text{ jika } \left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)^T \mathbf{C}_{MCD}^{-1} \left(\mathbf{x}_i - \mathbf{t}_{MCD}\right) \leq \chi^2_{p,0.975} \\ 0 \text{ lainnya} \end{cases}$$

Definisi 2.2 ***Affine Equivariant*** **(Lopuhaa dan Rousseeuw, 1991: 230).**

Diketahui $\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right]^T$ merupakan himpunan data sejumlah *n* pengamatan dalam ruang $\mathbb{R}^p$ dan $\mathbf{t} \in \mathbb{R}^p$ adalah penakasir parameter lokasi berdasarkan **X**. Penaksir $\mathbf{t}(\mathbf{X})$ dikatakan *affine equivariant* jika $\mathbf{t}\left(\mathbf{XA} + \mathbf{jv}^T\right) = \mathbf{t}(\mathbf{X})\mathbf{A} + \mathbf{v}$ untuk semua **A** matrik nonsingular berdimensi *pxp* semua $\mathbf{v} \in \mathbb{R}^p$ di mana **v** vektor *px*1 dan $\mathbf{j} = \left[1, 1, ..., 1\right]^T$ vektor *nx*1.

Penaksir rata-rata MCD $\mathbf{t}_{MCD}$ bersifat *affine equivariant.* Misal,

$$\mathbf{t}_n(\mathbf{X}) = \mathbf{t}_{MCD} = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i} = \frac{\mathbf{w}^T \mathbf{X}}{\mathbf{w}^T \mathbf{j}},$$

dengan **w** vektor yang memuat elemen pembobot masing-masing pengamatan atau $\mathbf{w} = \left[w_1, ..., w_n\right]^T$. Maka,

$$\mathbf{t}_n\left(\mathbf{XA} + \mathbf{jv}^T\right) = \frac{\mathbf{w}^T\left(\mathbf{XA} + \mathbf{jv}^T\right)}{\mathbf{w}^T \mathbf{j}}$$

$$= \frac{\mathbf{w}^T \mathbf{XA} + \mathbf{w}^T \mathbf{jv}^T}{\mathbf{w}^T \mathbf{j}}$$

$$= \frac{\mathbf{w}^T\mathbf{XA}}{\mathbf{w}^T\mathbf{j}} + \frac{\mathbf{w}^T\mathbf{j}\mathbf{v}^T}{\mathbf{w}^T\mathbf{j}} = \mathbf{t}(\mathbf{X})\mathbf{A} + \mathbf{v}$$

Penaksir kovariansi MCD, $\mathbf{C}_{MCD}$, juga bersifat *affine equivariant*. Suatu penaksir parameter sebaran data $\mathbf{C}(\mathbf{X})$ bersifat *affine equivariant* jika

$$\mathbf{C}\left(\mathbf{XA} + \mathbf{j}\mathbf{v}^T\right) = \mathbf{A}^T\mathbf{C}(\mathbf{X})\mathbf{A}$$

Dengan mengambil $\mathbf{C}_{MCD} = \dfrac{\sum_{i=1}^{n} w_i\left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)\left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)^T}{\sum_{i=1}^{n} w_i - 1}$ maka,

$$\mathbf{C}_n\left(\mathbf{XA} + \mathbf{j}\mathbf{v}^T\right) = \frac{\sum_{i=1}^{n} w_i\left(\mathbf{x}_i\mathbf{A} + \mathbf{v} - \left(\mathbf{t}_{MCD}\mathbf{A} + \mathbf{v}\right)\right)\left(\left(\mathbf{x}_i\mathbf{A} + \mathbf{v}\right) - \left(\mathbf{t}_{MCD}\mathbf{A} + \mathbf{v}\right)\right)^T}{\sum_{i=1}^{n} w_i - 1}$$

$$= \frac{\sum_{i=1}^{n} w_i\left(\mathbf{x}_i\mathbf{A} - \mathbf{t}_{MCD}\mathbf{A}\right)\left(\mathbf{x}_i\mathbf{A} - \mathbf{t}_{MCD}\mathbf{A}\right)^T}{\sum_{i=1}^{n} w_i - 1}$$

$$= \frac{\mathbf{A}^T\sum_{i=1}^{n} w_i\left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)\left(\mathbf{x}_i - \mathbf{t}_{MCD}\right)^T\mathbf{A}}{\sum_{i=1}^{n} w_i - 1} = \mathbf{A}^T\mathbf{C}_n(\mathbf{X})\mathbf{A}$$

Ukuran ke-*robust*-an yang sangat bermanfaat dari suatu penaksir adalah *breakdown point*. *Breakdown point* adalah jumlah pengamatan minimal yang dapat menggantikan sejumlah pengamatan mula-mula yang berakibat pada nilai taksiran yang dihasilkan sangat berbeda dari taksiran sebenarnya. *Breakdown point* dari penaksir parameter lokasi adalah $\mathbf{t}(\mathbf{X})$ adalah proporsi *outlier* terkecil m/n yang mengakibatkan nilai taksiran menjadi takterhingga:

$$\varepsilon_n^*(\mathbf{t}, \mathbf{X}) = \min_{1\le m\le n}\left\{\frac{m}{n}; \sup_{\mathbf{Y}_m}\left\|\mathbf{t}(\mathbf{Y}_m) - \mathbf{t}(\mathbf{X})\right\| = \infty\right\}$$

di mana suprimum diperoleh untuk semua kemungkinan himpunan yang terkontaminasi *outlier* $\mathbf{Y}_m$. $\mathbf{Y}_m$ diperoleh dengan menggantikan m elemen dari himpunan data $\mathbf{X}$ dengan nilai-nilai sembarang. *Breakdown point* penaksir

kovariansi $\mathbf{C}(\mathbf{X})$ didefinisikan sebagai proporsi *outlier* terkecil m/n yang mengakibatkan nilai eigen terbesar $\lambda_1(\mathbf{C})$ mencapai tak berhingga atau nilai eigen terkecil $\lambda_p(\mathbf{C})$ mendekati nol:

$$\varepsilon_n^*(\mathbf{C},\mathbf{X}) = \min_{1\le m\le n}\left\{\frac{m}{n}; \sup_{\mathbf{Y}_m} D\left(\mathbf{C}(\mathbf{Y}_m), \mathbf{t}(\mathbf{X})\right) = \infty\right\}$$

di mana suprimum diperoleh untuk semua kemungkinan himpunan yang terkontaminasi *outlier* $\mathbf{Y}_m$ di mana

$D(\mathbf{A},\mathbf{B}) = \max\left\{\left|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})\right|, \left|\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}\right|\right\}$ dengan

$\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_p(\mathbf{A})$ nilai-nilai egien dari matrik **A** berdimensi *pxp*.

Teorema 2.3 *Breakdown Point* Penaksir MCD (Lopuhaa dan Rousseeuw, 1991: 235).

Diketahui $\mathbf{X}_n = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$ merupakan himpunan data sejumlah n pengamatan dalam ruang $\square^p$ dengan $n \ge p+1$, dan $\mathbf{t}$ dan $\mathbf{C}$ penaksir MCD untuk rata-rata dan kovariansi. Jika $p = 1$ maka $\varepsilon_n^*(\mathbf{t},\mathbf{X}) = \left[(n+1)/2\right]/n$ dan $\varepsilon_n^*(\mathbf{C},\mathbf{X}) = \left[n/2\right]/n$. Pada saat $p \ge 2$, maka
$\varepsilon_n^*(\mathbf{t},\mathbf{X}) = \varepsilon_n^*(\mathbf{C},\mathbf{X}) = \left[(n-p+1)/2\right]/n$.

Bukti

Pada saat p = 1, $\mathbf{t}$ merupakan titik tengah dari interval terpendek yang meliputi sedikitnya $\left[n/2\right]+1$ pengamatan, dan $\mathbf{C}$ menyatakan proporsi panjang intervalnya. Dengan mengganti sedikitnya $\left[(n+1)/2\right]$ pengamatan mengakibatkan $\|\mathbf{t}\|$ menuju tak berhingga. Dengan mengganti sedikitnya $\left[n/2\right]$ pengamatan mengakibatkan $\mathbf{C}$ menjadi **0**.

Untuk $p \ge 2$ akan dibuktikan bahwa $\varepsilon_n^*(\mathbf{t},\mathbf{X})$ dan $\varepsilon_n^*(\mathbf{C},\mathbf{X})$ sedikitnya $\left[(n-p+1)/2\right]/n$. Pada saat h mengambil jumlah observasi minimal untuk mendapatkan taksiran $(\mathbf{t},\mathbf{C})$. Terdapat $n - h$ pengamatan yang tidak disertakan dalam penghitungan $(\mathbf{t},\mathbf{C})$. Selama penggantian pengamatan hanya sejumlah $n -$

h maka $(\mathbf{t},\mathbf{C})$ tidak mengalami banyak perubahan. Tetapi pada saat sejumlah $n-h+1$ pengamatan diganti dengan nilai ekstrim, maka $(\mathbf{t},\mathbf{C})$ akan berubah secara nyata. Misal $\mathbf{Y}_m$ himpunan data yang diperoleh dari himpunan data $\mathbf{X}$ dengan menggantikan sedikitnya $m = n-h+1 = n-\frac{n+p+1}{2}+1 = \frac{n-p+1}{2}$ pengamatan dengan menambahkan v, $\mathbf{Y}_m = \left[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{h-1}, \mathbf{x}_h + \mathbf{v}, \mathbf{x}_{h+1} + \mathbf{v}, ..., \mathbf{x}_n + \mathbf{v}\right]^T$.

Maka

$$\mathbf{t}^m = \frac{1}{h}\left(\sum_{i=1}^{h-1}\mathbf{x}_i + \mathbf{x}_h + \mathbf{v}\right) = \frac{1}{h}\left(\sum_{i=1}^{h}\mathbf{x}_i + \mathbf{v}\right)$$

$$= \bar{\mathbf{x}} + v = \left[\bar{x}_1 + v, \bar{x}_2 + v, ..., \bar{x}_p + v\right]^T.$$

Jika $v \to 0$ maka $\mathbf{t}^m \to \mathbf{t}$, tetapi jika $v \to \infty$ maka $\mathbf{t}^m \to \infty$..

Selanjutnya, misal $\mathbf{Y}_m$ himpunan data yang diperoleh dari himpunan data $\mathbf{X}$ dengan menggantikan sedikitnya $m = n-h+1 = \frac{n-p+1}{2}$ pengamatan dengan mengalikan konstanta a, $\mathbf{Y}_m = \left[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{h-1}, a\mathbf{x}_h, a\mathbf{x}_{h+1}, ..., a\mathbf{x}_n\right]^T$.

Maka $\mathbf{C}_{MCD}^m = a\mathbf{C}_{MCD}$. Jika $a \to \infty$ maka $\lambda_1(\mathbf{Y}_m) = a\lambda_1(\mathbf{X}) \to \infty$. □

2.4 Penaksir *Robust* MWCD

Diketahui $\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right]^T$ himpunan data. Parameter $\boldsymbol{\mu}$ diestimasi dengan meminimalkan jumlah pembobot dari jarak kuadrat Mahalanobis di mana pembobot bergantung pada urutan jaraknya. Fungsi pembobot yang digunakan adalah $a_n(i) = h^+\left(i/(n+1)\right), i = 1, ..., n$ di mana $h^+ : (0,1) \to [0,\infty)$ sehingga

$$\sup\left\{u; h^+(u) > 0\right\} = 1-\alpha,$$

Dengan $0 \le \alpha \le \frac{1}{2}$ dan $h^+(u) > 0$ untuk setiap $u \in (0, 1-\alpha]$. Oleh karena proporsi α dari pengamatan x_i diberi bobot 0, maka diperoleh penaksir *robust*.

Definisi 2.3 Penaksir MWCD (Roelant dkk. 2007: 2).

Penaksir MWCD adalah setiap penyelesaian
$\left(\hat{\boldsymbol{\mu}}_{MWCD}\left(X_n\right), \hat{\boldsymbol{\Sigma}}_{MWCD}\left(X_n\right)\right) = \underset{\mathbf{m},\mathbf{C},\det\mathbf{C}=1}{\arg\min}\, D_n\left(\mathbf{m},\mathbf{C}\right)$
di antara semua $\left(\mathbf{m},\mathbf{C}\right) \in \mathbb{R}^p \times \mathrm{PDS}\left(p\right)$ di mana $\mathrm{PDS}\left(p\right)$ adalah kelas matrik positif definit simetri berdimensi *p*. Fungsi objektif D_n didefinisikan sebagai

$$D_n\left(\mathbf{m},\mathbf{C}\right) = \sum_{i=1}^{n} a_n\left(R_i\right) d_i^2\left(\mathbf{m},\mathbf{C}\right)$$

Dengan $d_i^2\left(\mathbf{m},\mathbf{C}\right) = \left(\mathbf{x}_i - \mathbf{m}\right)^T C^{-1}\left(\mathbf{x}_i - \mathbf{m}\right)$ dan R_i menyatakan peringkat dari $d_i^2\left(\mathbf{m},\mathbf{C}\right)$ diantara $d_1^2\left(\mathbf{m},\mathbf{C}\right),...,d_n^2\left(\mathbf{m},\mathbf{C}\right)$.

Jika ada beberapa penyelesaian masalah minimalisasi fungsi objektif, hanya satu yang dipilih sebagai penaksir MWCD. Syarat $\det\mathbf{C} = 1$ mempunyai implikasi pada $\hat{\mathbf{V}}_{MWCD}$ dapat dianggap sebagai penaksir kovariansi. Formula lain yang equivalent dengan penaksir MWCD diperoleh sebagai berikut.

Diketahui $\Re$ merupakan himpunan semua permutasi dari $\left\{1,...,n\right\}$. Selanjutnya, untuk sembarang vektor $R = \left\{R_1,...,R_n\right\} \in \Re$, maka

$$\hat{\boldsymbol{\mu}}\left(\mathbf{R}\right) = \frac{\sum_{i=1}^{n} a_n\left(R_i\right)\mathbf{x}_i}{\sum_{i=1}^{n} a_n\left(R_i\right)}$$

$$\hat{\boldsymbol{\Sigma}}\left(\mathbf{R}\right) = \frac{\sum_{i=1}^{n} a_n\left(R_i\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\left(R_i\right)\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\left(R_i\right)\right)^T}{\sum_{i=1}^{n} a_n\left(R_i\right)}$$

Sebagaimana MCD, algoritma MWCD juga memanfaatkan teorema *C-Steps*. Dasar algoritma MWCD yang dikembangkan oleh Roelant dkk (2006) merupakan bentuk generalisasi dari Teorema *C-Steps*. Adapun algoritma MWCD dinyatakan dalam Teorema 2.4 sebagai berikut.

Teorema 2.4 Algoritma MWCD (Roelant dkk, 2006: 4).

Diketahui suatu himpunan data $\mathbf{X}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^p$ dan fungsi pembobot tidak naik a_n. Diketahui $Q_1 = \sum_{j=1}^{n} a_n(R_{1j}) d_1^2(j)$ dengan $R_1 = \{R_{11}, ..., R_{1n}\}$ vektor yang menyatakan urutan jarak $d_1^2(j) = (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1)^T \hat{\mathbf{V}}_1^{-1} (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1), j = 1, ..., n$ di mana $\hat{\boldsymbol{\mu}}_1 \in \mathbb{R}^p$ dan $\hat{\boldsymbol{\Sigma}}_1 \in \mathbb{R}^{pxp}$ dengan $\det \hat{\mathbf{V}}_1 = 1$. Hitung $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}(R_1)$ dan $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}(R_1)$. Hitung

$\hat{\mathbf{V}}_2 = (\det \hat{\boldsymbol{\Sigma}}_2)^{-1/p} \hat{\boldsymbol{\Sigma}}_2$

dan

$d_2^2(j) = (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)^T \hat{\mathbf{V}}_2^{-1} (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2), j = 1, ..., n$

dengan vektor ururtan jaraknya R_2. Jika $Q_2 = \sum_{j=1}^{n} a_n(R_{2j}) d_2^2(j)$ maka $Q_2 \le Q_1$.

Bukti.

$Q_2 = \sum_{j=1}^{n} a_n(R_{2i}) d_2^2(j) \le \sum_{j=1}^{n} a_n(R_{1j}) d_2^2(j)$ karena R_{2j} adalah vektor berdasarkan $d_2^2(j)$ dan a_n fungsi yang tidak naik. Fungsi a_n memberikan bobot paling besar untuk jarak paling kecil. Akibatnya, jumlahan yang dihasilkan lebih kecil daripada kombinasi pembobot dan jarak lainnya. Selanjutnya,

$$\sum_{j=1}^{n} a_n(R_{1j}) d_2^2(j) \le \sum_{j=1}^{n} a_n(R_{1j}) d_1^2(j) = Q_1$$

karena $\hat{\boldsymbol{\mu}}_2$ dan $\hat{\mathbf{V}}_2$ meminimumkan $\sum_{j=1}^{n} a_n(R_{1j}) d_j^2(\mathbf{m}, \mathbf{C})$. Anggap bahwa terdapat beberapa $m \in \mathbb{R}^p$ dan $C \in PDS(p)$ dengan $\det \mathbf{C} = 1$ sedemikian sehingga

$$\sum_{j=1}^{n} a_n(R_{1j}) d_j^2(\mathbf{m}, \mathbf{C}) < \sum_{j=1}^{n} a_n(R_{1j}) d_j^2(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}).$$

Implikasi dari *Preposition* 1 (Roelant dkk., 2007: 3) adalah

$$\frac{1}{N} \sum_{j=1}^{n} a_n(R_{1j}) d_j^2(\mathbf{m}, \det \hat{\boldsymbol{\Sigma}}_2^{1/p} \mathbf{C}) < p.$$

Sehingga terdapat sebuah konstanta $0 < c < 1$ sedemikian sehingga $\frac{1}{N}\sum_{j=1}^{n} a_n\left(R_{1j}\right) d_j^2\left(\mathbf{m}, c \det \hat{\mathbf{\Sigma}}_2^{1/p}\mathbf{C}\right) = p.$ Berdasarkan Lemma 1 (Roelant dkk., 2007: 16) maka

$$\det \hat{\mathbf{\Sigma}}_2 < \det\left(c \det \hat{\mathbf{\Sigma}}_2^{1/p} C\right) = c^p \det \hat{\mathbf{\Sigma}}_2$$

tampak kontradiktif sehingga

$$\sum_{j=1}^{n} a_n\left(R_{1j}\right) d_j^2\left(\mathbf{m}, \mathbf{C}\right) \geq \sum_{j=1}^{n} a_n\left(R_{1j}\right) d_j^2\left(\hat{\boldsymbol{\mu}}_2, \hat{\mathbf{V}}_2\right)$$

atau $Q_2 \leq Q_1$. □

Penaksir *robust* MWCD bersifat *affine equivariant* dan tergolong penaksir *robust* dengan *breakdown point* tinggi. *Breakdown point* untuk penaksir MWCD sama seperti MCD. Hal ini disebabkan jumlah pengamatan yang diberikan bobot nol pada penaksir MWCD sama dengan MCD. Akibatnya jumlah *breakdown point* yang dihasilkan MWCD juga sama dengan MCD.

2.5 Penaksir *Robust* dalam Analisis Diskriminan Kuadratik

Analisis diskriminan kuadratik *robust* diperoleh dengan mengganti penaksir vektor rata-rata dan matrik kovariansi sampel dengan MCD dan MWCD. Skor diskriminan kuadratik *robust* untuk penaksir MCD dan MWCD didefinisikan sebagai:

$$\hat{d}_j^{QMCD}\left(\mathbf{x}\right) = -\frac{1}{2}\ln\left|\mathbf{S}_{MCDj}\right| - \frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_{MCDj}\right)^t \mathbf{S}_{MCDj}^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}_{MCDj}\right) + \ln\left(\hat{p}_j^C\right) \tag{2.23}$$

dan

$$\hat{d}_j^{QMWCD}\left(\mathbf{x}\right) = -\frac{1}{2}\ln\left|\mathbf{S}_{MWCDj}\right| - \frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_{MWCDj}\right)^t \mathbf{S}_{MWCDj}^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}_{MWCDj}\right) + \ln\left(\hat{p}_j^C\right) \tag{2.24}$$

Selanjutnya alokasikan **x** sebagai kelompok π_k jika

$$\hat{d}_k^{QMCD}\left(\mathbf{x}\right) \leq \hat{d}_j^{QMCD}\left(\mathbf{x}\right); k \neq j. \tag{2.25}$$

Demikian juga untuk penaksir MWCD, alokasikan **x** sebagai kelompok π_k jika

$$d_k^{QMWCD}\left(\mathbf{x}\right) \geq d_j^{QMWCD}\left(\mathbf{x}\right); k \neq j. \tag{2.26}$$

Untuk mengukur seberapa baik aturan diskriminan yang dihasilkan dapat digunakan beberapa metode. Johnson dan Wichern (2002) mengemukakan setidaknya tuhuh metode evaluasi fungsi diskriminan. Ketujuh metode tersebut adalah:

(i) *Expected Cost of Misclassification* (ECM).

(ii) *Total Probability of Misclassification* (TPM).

(iii) *Optimum Error Rate* (OER).

(iv) *Actual Error Rate* (AER).

(v) *Apparent Error Rate* (APER).

(vi) *Error-rate Estimate*.

(vii) *Holdout Procedure*.

ECM dibangun di atas tiga komponen, yaitu probabilita prior p_j, biaya misklasifikasi $c(k|j)$ dengan probabilita misklasifikasi $P(k|j)$. Secara singkat, ECM diformulasikan sebagai

$$\text{ECM} = \sum_{j=1}^{l} p_j \left(\sum_{\substack{k=1 \\ k \neq j}}^{l} P\left(k \middle| j\right) c\left(k \middle| j\right) \right) \tag{27}$$

Bila faktor biaya misklasifikasi $c(k|j)$ diabaikan atau diasumsikan sama untuk setiap kelompok, maka dari persamaan (27) dihasilkan rumusan TPM yang dinyatakan sebagai

$$\text{TPM} = \sum_{j=1}^{l} p_j \left(\int_{\substack{R_k \\ k \neq j}} f_j\left(\mathbf{x}\right) d\mathbf{x} \right) \tag{28}$$

dengan

$$\int_{\substack{R_k \\ k \neq j}} f_j\left(\mathbf{x}\right) d\mathbf{x} = P\left(k \middle| j\right)$$

Nilai minimum dari TPM menyatakan OER. OER diformulasikan sebagai

$$\text{OER} = \sum_{j=1}^{l} p_j \left(\int_{\substack{R_k \\ k \neq j}} f_j\left(\mathbf{x}\right) d\mathbf{x} \right) \tag{29}$$

Di mana R_k ditentukan oleh persamaan (27).

Baik ECM, TPM, maupun OER ketiganya dapat dihitung apabila fungsi densitas populasi diketahui. Dalam prakteknya, parameter populasi yang muncul dalam aturan diskriminan diestimasi dari sampel (Johnson dan Wichern, 2002). Oleh karena itu, evaluasi kinerja aturan diskriminan dari data sampel menggunakan AER yang difornulasikan sebagai

$$\text{AER} = \sum_{j=1}^{l} p_j \left(\int_{\substack{\hat{R}_k \\ k \neq j}} f_j(\mathbf{x}) d\mathbf{x} \right) \tag{30}$$

Tampak bahwa AER baik untuk sampel yang akan diklasifikasikan kemudian karena (30) ternyata masih sulit dihitung akibat dari fungsi densitas kelompok yang juga masih tidak diketahui. Dalam tataran sampel, kesulitan ini diatasi dengan memfungsikan data menjadi dua. Fungsi pertama disebut sebagai data *training*. Fungsi data *training* ini untuk menentukan aturan diskriminan. Fungsi kedua dinamakan sebagai evaluasi. Dengan data evaluasi ini kinerja aturan diskriminan diukur dengan

$$APER = \frac{\sum_{j=1}^{l} n_{jM}}{\sum_{j=1}^{l} n_j} \tag{31}$$

Singkatnya, APER adalah persentase pengamatan yang dikelompokkan salah. Kelemahan APER terletak pada hasil yang terlalu optimis (Hubert dan Van Driessens, 2004; Johnson dan Wichern, 2002) karena data yang digunakan sebagai training juga digunakan sebagai data evaluasi.

Melalui langkah serupa APER tetapi dengan membedakan data menjadi dua, yaitu data *training* dan data evaluasi dihasilkan *Error-rate Estimate*. Dengan membagi data menjadi dua, memudahkan dalam penentuan aturan diskriminan dan mengukur kinerja. Kelemahannya adalah hasil estimasi terhadap AER lebih rendah (*underestimate*). Kelemahan ini dapat diatasi dengan memperbesar jumlah sampel pada masing-masing kelompok.

Kelemahan *Error-rate Estimate* yang hanya melalui satu kali pengukuran dapat diatasi dengan metode *Holdout.* Hubert dan Van Driessens (2004) menyebutnya sebagai metode *Cross Validation* dan Karsen (1972) menamainya

sebagai metode *Jacknife*. Menurut Johnson dan Wichern (2002) metode *Holdout* ini baik karena menyertakan seluruh pengamatan dalam mengukur kinerja aturan diskriminan akan tetapi Hubert dan Van Driessens (2004) menyatakan pada data yang besar prosedur ini membutuhkan waktu yang relatif lama.

2.6 Penentuan Rumah Tangga Miskin

Pada bagian ini dibahas beberapa penentuan rumah tangga miskin. Metode penentuan rumah tangga miskin di Indonesia selama ini mengacu pada metode penentuan rumah tangga miskin BPS. Metode penentuan rumah tangga miskin yang paling populer adalah garis kemiskinan makanan dan non makanan biasa disebut garis kemiskinan. Garis kemiskinan makanan sendiri merupakan konversi minimum kalori yang harus tersedia setiap hari agar seseorang dapat melakukan kegiatan sehari-hari ke dalam bentuk rupiah.

Menurut Ritonga (2004) penggunaan garis kemiskinan menemui beberapa kendala. Garis kemiskinan didasarkan pada SUSENAS Modul Konsumsi yang diselenggarakan setiap tiga tahun sekali. Jumlah sampel rumah tangga terpilih tidak sebesar SUSENAS Kor. Representasi garis kemiskinan hanya terbatas pada tingkat propinsi.

Pada tahun 2000, BPS mencoba memperkenalkan konsep mikro penentuan rumah tangga miskin. Berdasarkan studi tersebut diperoleh delapan variabel yang layak dan operasional untuk penentuan rumahtangga miskin di lapangan (BPS, 2000), yaitu: luas lantai perkapita (lebih kecil atau lebih besar dari 8m2), jenis lantai (tanah atau bukan tanah), ketersediaan air bersih (tidak terlindung atau terlindung), keberadaan jamban (tidak ada atau ada), kepemilikan asset (tidak punya atau punya), variasi konsumsi lauk pauk (tidak bervariasi dan bervariasi), pembelian pakaian (tidak pernah membeli minimal satu stel pakaian dalam setahun atau pernah), kehadiran dalam kegiatan sosial (tidak pernah hadir atau pernah). Kedelapan variabel tersebut telah mencakup aspek sosial dan ekonomi penduduk/rumahtangga diantaranya aspek sandang, pangan, perumahan, kepemilikan asset dan aktivitas sosial dan telah disertakan dalam SUSENAS tahun 2002.