UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 3 Nomor 3 Tahun 2025 Halaman 249-255.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


# Application of Partial Least Squares and Robust Approaches in Discriminant Analysis for High-Dimensional Data

**Rahmadina Adityana, Dodi Vionanda*, Dony Permana, Fadhilah Fitri**
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

**Submitted :** 25 Juni 2025
**Revised :** 04 Agustus 2025
**Accepted :** 06 Agustus 2025

## ABSTRACT

*Classical discriminant analysis, namely linear discriminant analysis and quadratic discriminant analysis, is generally known to suffer from singularity problems when exprerienced with high-dimensional data and is not robust to outliers that make the data not multivariate normally distributed. This research focuses on investigating the classification performance of discriminant analysis on high-dimensional data by applying two approaches, namely the Partial Least Square (PLS) dimension reduction approach as a solution to high-dimensional data and a robust approach with the Minimum Covariance Determinant (MCD) estimator technique that is robust to outliers. The data used for this study is Lee Silverman Voice Treatment (LSVT) data. PLS forms five optimal latent variables that represent predictor variable information. Based on the assumption test of covariance homogeneity between groups, the test statistic value is greater than the chi-square table or the p-value is smaller than the significance level, which means that the assumption is unfulfilled, so quadratic discriminant analysis is applied. The evaluation results showed that the quadratic discriminant analysis model with the MCD approach on the PLS transformed data was able to achieve 81% accuracy, 71% precision, 86% recall, and 77% F1-score. These values indicate that both approaches are able to maintain the efficiency of discriminant analysis classification performance on high-dimensional and multivariate non-normally distributed data.*

**Keywords:** *Discriminant Analysis, High-Dimensional Data, Minimum Covariance Determinant (MCD), Partial Least Square (PLS), Robust Estimation.*



## I. PENDAHULUAN

Analisis multivariat merupakan salah satu teknik dalam statistika yang dirancang untuk menganalisis data dengan lebih dari dua variabel secara bersamaan. Salah satu metode dalam analisis multivariat adalah analisis diskriminan. Johnson dan Wichern (2002) menyebut bahwa analisis diskriminan melakukan pemisahan objek sebisa mungkin sehingga membantu pengalokasian objek baru berdasarkan karakteristik variabel independen. Dalam pendekatan klasik, analisis diskriminan terdiri dari dua jenis yaitu analisis diskriminan linier dan analisis diskriminan kuadratik. Analisis diskriminan linier memerlukan asumsi normal multivariat pada variabel independen dan kehomogenan kovarians antar kelompok objek. Ketika asumsi kehomogenan kovarians tidak terpenuhi, analisis diskriminan kuadratik menjadi alternatif yang tepat dengan asumsi kenormalan multivariat tetap terpenuhi.

Namun, ketika data berdimensi tinggi dimana jumlah variabel yang jauh lebih besar daripada jumlah sampel, penerapan analisis diskriminan menghadapi tantangan besar. Varmuza dan Filzmiser (2009) menyebut bahwa kondisi tersebut dapat menyebabkan masalah singularitas dimana matriks tidak dapat diinvers, sedangkan hal tersebut termasuk prosedur dalam analisis diskriminan. Oleh karena itu, salah satu cara untuk menghadapi masalah ini adalah menggunakan metode reduksi dimensi sebelum analisis diskriminan dapat dilakukan secara efektif.

Metode reduksi yang sering digunakan pada analisis multivariat adalah *Partial Least Squares* (PLS) dan *Principal Component Analysis* (PCA). Namun, ketika diskriminasi dan klasifikasi menjadi tujuan maka *Partial Least Squares* (PLS) menjadi metode yang paling tepat digunakan untuk reduksi dimensi data (Barker dan Rayens, 2003). PLS memperhatikan varians variabel prediktor sekaligus mempertimbangkan kovariansnya dengan variabel respon, sedangkan PCA hanya memperhatikan variabilitas variabel prediktor.


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


Algoritma PLS yang digunakan pada penelitian ini adalah algoritma *Statistically Inspired Modification of Partial Least Squares* (SIMPLS). Algoritma ini memberikan efisiensi komputasi yang lebih tinggi dan tidak memerlukan inversi matriks dalam perhitungan koefisien regresi sehingga menjadikannya lebih stabil dan efisien dalam menangani dataset kompleks dengan banyak variabel (Varmuza dan Filzmiser, 2009).

Setelah data tereduksi, masalah lain yang mungkin terjadi adalah pelanggaran asumsi normal multivariat pada analisis diskriminan yang disebabkan oleh adanya pencilan. Hubert dkk. (2024) menyebut bahwa pencilan merupakan hal yang sensitif bagi analisis diskriminan. Oleh karena itu, penting untuk mempertimbangkan pendekatan yang lebih kuat dalam analisis diskriminan untuk mengatasi masalah ini yang dikenal dengan metode *robust*. Todorov dan Pires (2007) membahas beberapa metode *robust* yang pernah diterapkan pada analisis diskriminan untuk mengatasi masalah kelemahan estimasi dalam analisis diskriminan, yaitu *Minimum Covariance Determinant* (MCD), *Minimum Volume Ellipsoid* (MVE), *Minimum Within Covariance Determinant* (MWCD), *S-estimator*, dan *M-estimator*.

Teknik penduga MCD akan diterapkan pada penelitian ini untuk menduga parameter yang digunakan pada analisis diskriminan klasik. Todorov dan Pires (2007) menyatakan penduga MCD menjadi metode *robust* yang dominan dalam praktik statistik karena didukung oleh algoritma yang efisien serta tersedia secara luas dalam perangkat lunak statistik popular, seperti R, S-Plus, SAS, dan Matlab. Namun, estimator ini tidak efisien pada model normal. Berbagai penelitian telah menunjukkan efektivitas penduga MCD dalam analisis diskriminan. Menurut Hubert dkk. (2024), penggunaan penduga MCD untuk lokasi dan sebaran dalam perhitungan skor diskriminan menghasilkan prosedur yang bersifat kokoh terhadap pencilan. Studi simulasi yang dilakukan Hubert dan Van (2004) juga mendemonstrasikan bagaimana pendekatan *robust* berbasis MCD tidak terpengaruh oleh pencilan, berbeda dengan aturan klasik yang sangat terganggu oleh kehadiran beberapa pencilan. Selain itu, penelitian yang menggunakan MCD pada analisis diskriminan juga dilakukan oleh Budyanra (2016). Penelitian ini membandingkan keefektifan klasifikasi metode analisis diskriminan klasik dengan analisis diskriminan yang menggunakan penduga MCD ketika data simulasi mengandung pencilan. Hasil penelitian tersebut menyebut bahwa penduga MCD mampu mengurangi kesalahan dalam klasifikasi pada analisis diskriminan karena memiliki rata-rata salah pengklasifikasian sebesar 11%, sedangkan analisis tanpa menggunakan penduga MCD memberikan nilai sebesar 22%. Penelitian lain juga dilakukan oleh Ariyandi (2021) yang membandingkan penduga *robust* MCD dan MVE pada data bangkitan yang diatur mengandung pencilan. Penelitian tersebut menyatakan bahwa peduga *robust* MCD lebih baik dibanding MVE karena menghasilkan *Apparent Error Rate* (APER) yang lebih kecil.

Data berdimensi tinggi yang kompleks dan tantangan yang ditimbulkan terhadap asumsi klasik dipertimbangkan dalam analisis diskriminan. Oleh sebab itu, tujuan penelitian berfokus pada penerapan teknik reduksi dimensi PLS dan pendekatan *robust* MCD untuk menjaga kemampuan analisis diskriminan tetap efisien dalam klasifikasi. Penelitian ini menggunakan data penilaian *Lee Silverman voice treatment* (LSVT) pada pasien penyakit Parkinson. Penyakit Parkinson merupakan kondisi rusaknya sistem saraf pada otak secara bertahap dan LSVT menjadi metode terapi wicara yang terbukti efektif bagi penderita penyakit Parkinson yang sebagian besar mengalami gangguan vokal (Tsanas, dkk., 2013). Data ini merupakan data berdimensi tinggi karena memiliki jumlah variabel prediktor yang lebih besar dari jumlah sampel.

## II. METODE PENELITIAN

### A. Jenis Penelitian dan Sumber Data

Penelitian ini merupakan jenis penelitian kuantitatif dengan menggunakan data sekunder yang diperoleh dari repositori *online* UCI *Machine Learning*. Data yang digunakan merupakan hasil rekaman fonasi vokal /a/ yang berkelanjutan dari pasien penyakit Parkinson yang dilakukan pada ruangan kedap suara di *National Center for Voice and Speech-Denver* (NCVS), sebuah Lembaga afiliasi dari Universitas Colorado-Boulder (Tsanas, dkk., 2013). Data terdiri dari 126 sampel dengan 310 pengukuran sebagai variabel prediktor bertipe numerik dan satu fitur sebagai variabel respon dengan kategori biner dimana “1” berarti suara dianggap memenuhi standar kualitas yang diharapkan sehingga pasien dapat melanjutkan latihan suara selama sesi terapi tatap muka dan “2” berarti suara dianggap tidak memenuhi standar kualitas yang diharapkan, sehingga seorang ahli tidak akan membiarkan pasien melanjutkan terapi dengan suara tersebut.

### B. Teknik Analisis Data

Penelitian ini melakukan analisis dengan langkah-langkah berikut:

1. Mempersiapkan data yang akan dianalisis.
2. Melakukan eksplorasi data untuk mengidentifikasi data hilang dan keseimbangan data.


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


3. Melakukan standardisasi data untuk menyamakan skala variabel yang memiliki rentang nilai yang berbeda. Hal ini dilakukan untuk mengurangi bias dalam proses analisis. Setiap amatan distandardisasi sehingga memiliki nilai rata-rata nol dan varians satu. Menurut Varmuza dan Filzmiser (2009), persamaan yang digunakan dalam standardisasi data dapat ditulis sebagai berikut,

$$x_{jk}(autoscaled) = \frac{x_{jk}(original) - \bar{x}_k}{s_k} \quad (1)$$

dimana $x_{jk}(autoscaled)$ merupakan nilai amatan ke-$j$ variabel ke-$k$ yang telah distandardisasi, $x_{jk}(original)$ merupakan nilai amatan ke-$j$ variabel ke-$k$ pada data asli, $\bar{x}_k$ merupakan rata-rata variabel ke-$k$, dan $s_k$ merupakan standar deviasi variabel ke-$k$.
4. Membagi data secara acak menjadi dua bagian, yaitu 70% sebagai data latih dan 30% sebagai data uji.
5. Menerapkan metode reduksi data dengan pendekatan PLS pada data latih untuk memperoleh variabel laten PLS dan memvalidasi model PLS dengan teknik *k-fold cross validation* dimana $k = 5$.
6. Memilih jumlah variabel laten optimal dengan cara menguji model PLS dari sejumlah variabel laten yang nanti akan digunakan pada analisis diskriminan.
7. Menguji dua asumsi pada analisis diskriminan.
   a. Distribusi normal multivariat
      Pengecekan asumsi distribusi normal multivariat dilakukan menggunakan *Q-Q plot*. Ketika titik-titik objek mengikuti garis diagonal dan tidak ada titik yang menyimpang cukup jauh maka data dinyatakan mengikuti distribusi normal multivariat.
   b. Homogenitas kovarians antar kelompok
      Dalam pengujian homogenitas matriks kovarians antar dua kelompok, hipotesis yang ditetapkan adalah

$$H_0: \Sigma_1 = \Sigma_2$$
$$H_1: \Sigma_1 \neq \Sigma_2$$

Pengujian ini dilakukan menggunakan uji *Box's M* dengan pendekatan *chi-square* $\chi^2$. Tolak $H_0$ jika nilai statistik $u > \chi^2_{\frac{1}{2}(p(p+1))}$ dimana $p$ merupakan jumlah variabel prediktor dan

$$u = -1(1-c)\ln M \quad (2)$$

dengan

$$c = \left[\sum_{i=1}^{2} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{2} \frac{1}{n_i - 1}}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)}\right] \quad (3)$$
$$\ln M = \frac{1}{2}\sum_{i=1}^{2}(n_i - 1)\ln |\mathbf{S}_i| - \frac{1}{2}\left(\sum_{i=1}^{2}(n_i - 1)\right)\ln |\mathbf{S}_{\text{pl}}|$$

8. Melakukan analisis diskriminan

Pada analisis diskriminan, menurut Johnson dan Wichern (2002), objek dipisahkan semaksimal mungkin agar mampu mengalokasikan objek baru dengan tepat berdasarkan karakteristik variabel prediktor dengan membentuk fungsi diskriminan. Bentuk fungsi diskriminan didasarkan pada asumsi data. Ketika data mengikuti distribusi normal multivariat dan kovarians antar kelompok sama maka fungsi dibentuk berdasarkan analisis diskriminan linier dengan persamaan sebagai berikut,

$$y = \mathbf{a}^T\mathbf{x} = ((\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\, \mathbf{S}_{\text{pl}}^{-1})^T\mathbf{x} \quad (4)$$

dimana $\mathbf{a}$ sebagai koefisien yang memproyeksikan titik-titik $y$ ke garis yang memaksimalkan jarak kedua kelompok, $\bar{\mathbf{x}}_1$ dan $\bar{\mathbf{x}}_2$ secara berurut merupakan rata-rata kelompok 1 dan 2, $\mathbf{S}_{\text{pl}}$ merupakan gabungan matriks kovarians kedua kelompok yang ditulis dalam Persamaan (5), dan $\mathbf{x}$ merupakan vektor acak variabel prediktor.

$$\mathbf{S}_{\text{pl}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \quad (5)$$
$$\underset{(p \times p)}{\mathbf{S_1}} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^T \quad (6)$$
$$\underset{(p \times p)}{\mathbf{S_2}} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^T \quad (7)$$

Namun, ketika kovarians antar kelompok berbeda maka fungsi dibentuk berdasarkan analisis diskriminan kuadratik yang ditulis dalam persamaan berikut,


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


$$d_1^Q = -\frac{1}{2}\ln \mathbf{S}_1 - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_1)^T\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) + \ln p_1 \quad (8)$$

$$d_2^Q = -\frac{1}{2}\ln \mathbf{S}_2 - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_2)^T\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) + \ln p_2 \quad (9)$$

dengan $\mathbf{S}_1$ dan $\mathbf{S}_2$ secara berurut merupakan matriks kovarians sampel kelompok 1 dan 2, lalu $p_1$ dan $p_2$ secara berurut merupakan proporsi awal dari kelompok 1 dan 2.

Apabila asumsi normalitas tidak terpenuhi, metode *robust* dengan teknik MCD digunakan untuk menduga rata-rata dan matriks kovarians kedua kelompok yang lebih kokoh terhadap pencilan. Berikut adalah tahapan-tahapan yang dilakukan untuk memperoleh nilai duga tersebut.

a. Pilih secara acak subhimpunan sebanyak $h$ observasi dari data $X_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ dengan $p$ variabel. Nilai bawaan (*default*) $h$ adalah $\frac{(n+p+1)}{2}$, tetapi besar $h$ juga dapat ditentukan dengan $\frac{(n+p+1)}{2} \le h \le n$.

b. Hitung rata-rata $\bar{\mathbf{x}}_{MCD}$ dan matriks kovarinsnya $\mathbf{S}_{MCD}$ dari subhimpunan tersebut menggunakan Persamaan (10) dan (11).

$$\bar{\mathbf{x}}_{MCD} = \frac{1}{h}\sum_{j=1}^{h}\mathbf{x}_j \quad (10)$$

$$\mathbf{S}_{MCD} = \frac{1}{h-1}\sum_{j=1}^{h_1}(\mathbf{x}_j-\bar{\mathbf{x}}_1)(\mathbf{x}_j-\bar{\mathbf{x}}_1)^T \quad (11)$$

c. Gunakan nilai rata-rata $\bar{\mathbf{x}}_{\mathrm{MCD}}$ dan matriks kovarians $\mathbf{S}_{MCD}$ untuk menghitung jarak Mahalanobis setiap sampel himpunan data latih yang telah direduksi menggunakan Persamaan (12) dan urutkan dari terkecil hingga terbesar serta hitung determinan matriks kovarians $det\ (\mathbf{S}_{MCD})$**.** Lalu, pilih $h$ observasi dengan nilai $\Delta_j$ terkecil untuk membentuk subhimpunan baru.

$$\Delta_j^2 = \left(\mathbf{x}_j - \bar{\mathbf{x}}\right)^T\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}), \qquad j = 1, 2, \dots, n \quad (12)$$

dimana $n$ merupakan banyak sampel pada data latih.

d. Lakukan langkah b dan c hingga ditemukan $det\ (\mathbf{S}_{MCD})$ terkecil dimana perbedaan antara estimasi $det\ (\mathbf{S}_{MCD})$ baru dan sebelumnya cukup kecil atau sama yakni $det\ (\boldsymbol{S}_{MCD_{baru}}) \le \det\ (\boldsymbol{S}_{MCD_{lama}})$.

e. Hitung rata-rata dan matriks kovarians kedua kelompok dari subhimpunan terakhir yang memiliki determinan kovarians terkecil.

Ketika matriks kovarians kedua kelompok sama maka gunakan nilai rata-rata dan matriks kovarians kedua kelompok dalam perhitungan fungsi diskriminan linier. Namun, jika berbeda gunakan nilai-nilai tersebut dalam perhitungan fungsi diskriminan kuadratik.

9. Menguji model analisis diskriminan menggunakan data uji dan menghitung kinerja model berdasarkan nilai *accuracy*, *precision, recall*, dan *F1-score* secara berurutan menggunakan Persamaan (13) hingga (16).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

10. Melakukan interpretasi dan penarikan kesimpulan terkait kemampuan analisis diskriminan dalam klasifikasi data dimensi tinggi dengan pendekatan *Partial Least Square* (PLS) sebagai reduksi dimensi dan metode *robust*.

## III. HASIL DAN PEMBAHASAN

### A. Eksplorasi Data

Dalam proses eksplorasi data LSVT, tidak ditemukan adanya nilai yang hilang (*missing value*) pada seluruh variabel yang menandakan data penelitian ini lengkap. Namun, jumlah sampel dari kedua kelompok tidak seimbang dengan mayoritas 67% dikelompokkan pada pasien yang menemukan suara sesuai standar, sedangkan sisanya 33% dari kelompok dengan suara yang tidak memenuhi standar. Perbandingan proporsi kedua kelompok dapat dilihat pada Gambar 1.


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



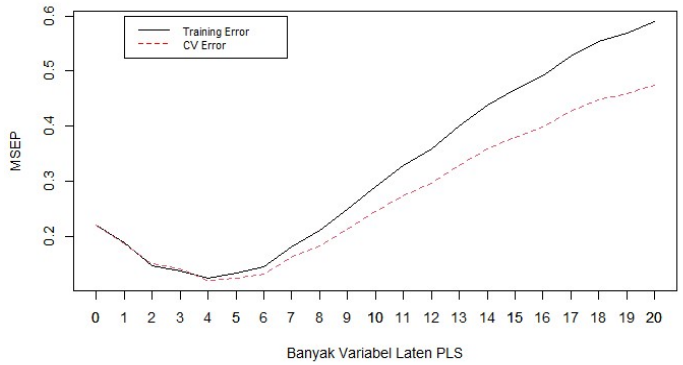
**Gambar 1**. Proporsi Sampel Kelompok Pasien Parkinson Data *Lee Silverman Voice Treatment*

## B. Standarisasi Data

Data LSVT memiliki nilai amatan antar variabel dengan skala yang beragam dan perbedaan yang besar. Data distandardisasi agar skala antar variabel seragam sehingga setiap variabel memiliki kontribusi dalam analisis dan mengurangi bias.

## C. Reduksi Dimensi dengan metode PLS

Dalam proses reduksi dimensi dengan metode PLS pada data latih, model PLS dilatih dan divalidasi menggunakan teknik *k-fold cross validation* dengan 5 subset data. Tidak semua model variabel laten divalidasi karena semakin besar jumlah variabel laten maka *Mean Square Error of Prediction* (MSEP) model akan semakin konstan atau tetap tidak berubah. Model dengan nol hingga dua puluh variabel laten divalidasi dan diperoleh kesimpulan bahwa model dengan empat variabel laten memberikan MSEP terendah seperti yang ditunjukkan oleh Gambar 1.




**Gambar 1**. *Mean Square Error of Prediction* dari Sejumlah Variabel Laten PLS yang Berbeda

Berdasarkan Gambar 1, peningkatan jumlah variabel laten setelah variabel laten dengan MSEP minumum menghasilkan model yang semakin *overfitting*. Oleh karena itu, performa klasifikasi model dengan empat variabel laten dan dua model terdekat dari model tersebut yaitu dua, tiga, lima, dan enam variabel laten akan diuji berdasarkan metrik evaluasi akurasi untuk memilih variabel laten optimal.

**Tabel 1**. Nilai Akurasi Model PLS dengan Lima Variabel Laten yang Berbeda

| Jumlah variabel Laten | Nilai Akurasi (%) |
|---|---|
| 2 | 78,38 |
| 3 | 81,08 |
| 4 | 78,38 |
| 5 | 81,08 |
| 6 | 78,38 |


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

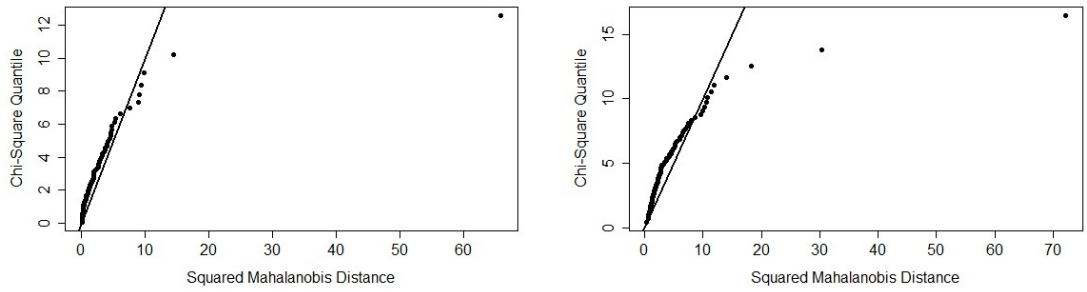
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


Berdasarkan Tabel 1, hasil evaluasi akurasi tertinggi dimiliki oleh model PLS dengan tiga dan lima variabel laten yang bernilai sama. Oleh karena itu, kedua jumlah variabel laten tersebut diuji lebih lanjut dalam analisis diskriminan untuk mengidentifikasi jumlah variabel optimal yang memberikan kinerja klasifikasi terbaik.

## D. Uji Asumsi Analisis Diskriminan

1. Distribusi Normal Multivariat

*Q-Q plot* dari data LSVT yang telah direduksi dengan PLS disajikan pada Gambar 2.




(a) (b)

**Gambar 2**. (a) *Q-Q Plot* Data LSVT yang Direduksi Tiga Variabel Laten PLS dan (b) *Q-Q Plot* Data LSVT yang Direduksi Lima Variabel Laten PLS

Berdasarkan Gambar 2, sebagian besar titik pada kedua opsi data mengikuti garis diagonal yang merupakan pola distribusi normal multivariat. Akan tetapi, terdapat beberapa titik yang menyimpang jauh dari garis diagonal sehingga kedua data dinyatakan tidak berdistribusi normal multivariat. Titik yang cukup jauh dari garis diagonal diindikasi merupakan pencilan.

2. Homogenitas Kovarians Antar Kelompok

Asumsi homogenitas kovarians antar kelompok pada kedua opsi data tidak terpenuhi. Berdasarkan hasil uji *Box's M* yang disajikan oleh Tabel 2, pada tingkat signifikan 5%, nilai statistik uji pada kedua kondisi data lebih besar dari nilai *chi-square table* atau nilai $p$ lebih kecil dari taraf signifikan yang menandakan hipotesis awal ditolak.

**Tabel 2**. Hasil Uji *Box's M* Data LSVT yang Direduksi Tiga dan Lima Variabel Laten PLS

| ***Output*** | **Tiga Variabel Laten PLS** | **Lima Variabel Laten PLS** |
|---|---|---|
| Statistik Uji ($u$) | 112,46 | 191,76 |
| Derajat Bebas | 6 | 15 |
| *p-value* | 2,2E-16 | 2,2E-16 |
| *Chi-Square Table* | 12,592 | 24,996 |

## E. Analisis Diskriminan

Berdasarkan hasil pengujian asumsi pada analisis diskriminan, penelitian ini melakukan analisis diskriminan kuadratik *robust*. Kedua opsi data tidak mengikuti distribusi normal multivariat karena adanya pencilan dan kovarians antar kelompok tidak homogen. Dengan demikian, metode *robust* dengan teknik MCD digunakan untuk menduga lokasi dan sebaran pada analisis diskriminan serta teknik kuadratik diterapkan.

Model dari analisis diskriminan kuadratik *robust* yang dibentuk dengan opsi lima variabel laten PLS menghasilkan kinerja yang lebih baik dibandingkan dengan model yang dibentuk dengan opsi tiga variabel laten PLS. Hal ini didasarkan dari penilaian metrik evaluasi klasifikasi yang disajikan pada Tabel 3. Dengan demikian, data LSVT lebih optimal bekerja ketika direduksi dengan lima variabel laten PLS.

**Tabel 3**. Evaluasi Klasifikasi Model Analisis Diskriminan Kuadratik

| **Metrik Evaluasi Klasifikasi** | **Model Analisis Diskriminan Kuadratik** | |
|---|---|---|
| | **Data LSVT dengan Tiga Variabel Laten PLS (%)** | **Data LSVT dengan Lima Variabel Laten PLS (%)** |
| *Accuracy* | 70 | 81 |
| *Precision* | 57 | 71 |


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X


| | | |
|---|---|---|
| *Recall* | 86 | 86 |
| *F1-score* | 69 | 77 |

Dari Tabel 3, model analisis diskriminan kuadratik *robust* dari lima variabel laten PLS memiliki kinerja klasifikasi dengan *accuracy* sebesar 81% yang menunjukkan model bekerja dengan cukup baik secara keseluruhan dalam klasifikasi. Dalam ketepatan prediksi pasien kelompok 1, nilai *precision* menunjukkan dari seluruh pasien yang diprediksi suara memenuhi standar, 71% mampu diprediksi dengan benar, sedangkan dalam kemampuan mengenali kelompok 1, nilai *recall* menunjukkan dari seluruh pasien yang memiliki suara standar, 86% berhasil diidentifikasi dengan benar oleh model. Model cukup seimbang dalam mengenali dan memprediksi pasien yang memiliki suara sesuai standar dimana ditunjukkan oleh *F1-score* sebesar 77%. Nilai-nilai ini menunjukkan kemampuan analisis diskriminan dalam klasifikasi ketika dihadapkan dengan data berdimensi tinggi dan pelanggaran asumsi dengan menerapkan PLS dan MCD pada kasus data LSVT.

## IV. KESIMPULAN

Berdasarkan hasil dari proses reduksi dimensi dengan PLS, data LSVT secara optimal direduksi menjadi lima variabel laten sehingga mampu diterapkan pada analisis diskriminan. Selain itu, metode *robust* MCD mampu menjaga efisiensi klasifikasi model dari analisis diskriminan ketika data tidak mengikuti distribusi normal multivariat. Kedua pendekatan ini mampu bekerja bagi analisis diskriminan ketika berhadapan dengan data LSVT yang berdimensi tinggi dan pelanggaran asumsi normal multivariat.

Bagi peneliti selanjutnya, disarankan menggunakan algoritma PLS lainnya dan pendekatan lain untuk reduksi dimensi serta teknik pada metode *robust* lainnya agar dapat mengetahui algoritma dan pendekatan yang lebih optimal dalam meningkatkan kinerja model analisis diskriminan saat berhadapan dengan data berdimensi dan pelanggaran asumsi normalitas multivariat.

## DAFTAR PUSTAKA

Ariyandy, Z. (2021). *Perbandingan Penggunaan Penduga Robust Minimum Volume Ellipsoid (Mve) Dan Minimum Covariance Determinant (Mcd) Pada Analisis Diskriminan Kuadratik* (Doctoral dissertation, Universitas Brawijaya).

Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society, 17*(3), 166-173.

Budyanra. (2016). Ketepatan Pengklasifikasian Fungsi Diskriminan Linier Robust Dua Kelompok Dengan Metode Fast Minimum Covariate Determinant (Fast–Mcd). *Jurnal Statistika, 4*(2), 15-19.

Hubert, M., Raymaekers, J., and Rousseeuw, P. J. (2024). Robust discriminant analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 16*(5), e70003.

Hubert, M., and Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis, 45*(2), 301-320.

Johnson, R. A., and Wichern, D. W. (2002). *Applied multivarite statistical analysis (5th ed.)*. New Jersey: Prentice Hall.

Todorov, V., and Pires, A. M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT-Statistical Journal, 5*(1), 63-83.

Tsanas, A., Little, M. A., Fox, C., and Ramig, L. O. (2013). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 22*(1), 181-190.

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC press.


DOI: https://doi.org/10.24036/ujsds/vol3-iss3/396