

Pengelompokan Daerah di Indonesia Berdasarkan Indikator Penetapan Daerah Tertinggal Menggunakan Model Based Clustering

Nurwahyuni, Junaidi, Nurul Fiskia Gamayanti



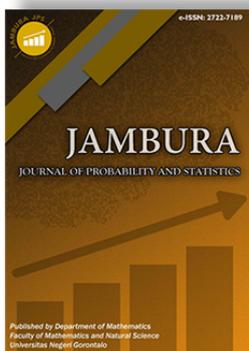
Volume 6, Issue 1, Pages 14–20, May 2025

Received 01 June 2024, Revised 29 December 2024, Accepted 28 April 2024, Published Online 31 Mei 2025

To Cite this Article : Nurwahyuni, Junaidi, N. F. Gamayanti, "Pengelompokan Daerah di Indonesia Berdasarkan Indikator Penetapan Daerah Tertinggal Menggunakan Model Based Clustering", *Jambura J. Probab. Stat.*, vol. 6, no. 1, pp. 14–20, 2025, <https://doi.org/10.34312/jjps.v4i1.25724>

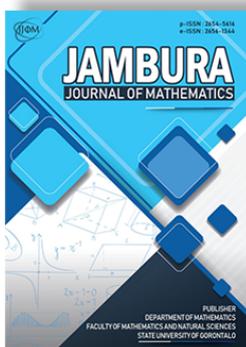
© 2025 by author(s)

JOURNAL INFO • JAMBURA JOURNAL OF PROBABILITY AND STATISTICS



	Homepage	: https://ejournal.ung.ac.id/index.php/jps/index
	Journal Abbreviation	: Jambura J. Probab. Stat.
	Frequency	: Biannual (May and November)
	Publication Language	: English (preferable), Indonesia
	DOI	: https://doi.org/10.34312/jjps
	Online ISSN	: 2722-7189
	Editor-in-Chief	: Ismail Djakaria
	Publisher	: Department of Mathematics, Universitas Negeri Gorontalo
	Country	: Indonesia
	OAI Address	: http://ejournal.ung.ac.id/index.php/jps/oai
	Google Scholar ID	: kWdujzMAAAJ
	Email	: redaksi.jjps@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS



Jambura Journal of Mathematics



Jambura Journal of Mathematics Education



Jambura Journal of Biomathematics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi



Pengelompokan Daerah di Indonesia Berdasarkan Indikator Penetapan Daerah Tertinggal Menggunakan Model Based Clustering

Nurwahyuni¹, Junaidi², Nurul Fiskia Gamayanti³

^{1,2,3}Program Studi Statistika, Jurusan Matematika, Program Studi Statistika, Fakultas MIPA, Universitas Tadulako

ARTICLE HISTORY

Received 01 June 2024

Revised 29 December 2024

Accepted 28 April 2024

Published 31 Mei 2025

KATA KUNCI

Daerah Tertinggal
Model Based Clustering
Finite Mixture

KEYWORDS

Underdeveloped Regions
Model-Based Clustering
Finite Mixture

ABSTRAK. Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional. Di Indonesia terdapat 62 daerah kabupaten yang tertinggal dan tersebar di berbagai provinsi. Penelitian ini bertujuan untuk mengelompokan 62 daerah kabupaten tersebut berdasarkan Indikator Penetapan Daerah Tertinggal yang terdiri dari Produk Domestik Regional Bruto (PDRB) per kapita (X_1), Persentase Pengeluaran Rumah Tangga Non Makanan (X_2), Angka Partisipasi SMP (APS SMP) (X_3), Angka Partisipasi SMA (APS SMA) (X_4), Desa Yang Punya Fasilitas Kesehatan (X_5), Desa Yang Punya Dokter (X_6), Desa Yang Punya SD (X_7), Desa Yang Punya SMP (X_8), Pendapatan Asli Daerah (PAD) Per Kapita (X_9). Adapun metode yang akan digunakan dalam penelitian ini adalah Model Based Clustering dengan menggunakan pendekatan distribusi t multivariat. Metode ini merupakan metode dengan model statistik sebagai fokus utamanya sehingga menempatkan metode ini pada asumsi Finite Mixture. Dalam Finite Mixture data diasumsikan bersumber dari sejumlah distribusi dan cluster yang tercipta akan mewakili masing-masing distribusi probabilitas tersebut. Penelitian ini menghasilkan model UUUC sebagai model terbaik dengan 2 cluster optimal dengan masing-masing karakteristik yang dihasilkan. Cluster 1 dengan tingkat ketertinggalan daerah yang rendah dan cluster 2 dengan tingkat ketertinggalan daerah yang tinggi. Kemudian, harapannya daerah kabupaten dengan tingkat ketertinggalan tertinggi tersebut dapat lebih diprioritaskan dalam mencapai pemerataan pembangunan secara cepat dan tepat sasaran.

ABSTRACT. Binary Underdeveloped regions are regencies whose areas and communities are less developed compared to other regions on a national scale. In Indonesia, there are 62 underdeveloped regencies scattered across various provinces. This study aims to classify these 62 regencies based on the Indicators for Determining Underdeveloped Regions, which include Gross Regional Domestic Product (GRDP) per capita (X_1), Percentage of Non-Food Household Expenditure (X_2), Junior High School Participation Rate (X_3), Senior High School Participation Rate (X_4), Villages with Health Facilities (X_5), Villages with Doctors (X_6), Villages with Elementary Schools (X_7), Villages with Junior High Schools (X_8), and Regional Original Revenue (PAD) per capita (X_9). The method used in this study is Model-Based Clustering using a multivariate t -distribution approach. This method focuses on a statistical model and is based on the Finite Mixture assumption. In the Finite Mixture framework, the data is assumed to originate from several distributions, and the resulting clusters represent these probability distributions. The study identifies the UUUC model as the best model, producing two optimal clusters with distinct characteristics: Cluster 1 with a low level of regional underdevelopment, and Cluster 2 with a high level of underdevelopment. It is hoped that regencies classified in the highly underdeveloped cluster can be prioritized to achieve equitable development more quickly and effectively.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License. Editorial of JJPS: Department of Statistics, Universitas Negeri Gorontalo, Jln. Prof. Dr. Ing. B. J. Habibie, Bone Bolango 96554, Indonesia.

1. Pendahuluan

Ketidakeimbangan pembangunan antar daerah merupakan suatu permasalahan yang kerap kali dialami oleh suatu negara termasuk Indonesia. Ketidakeimbangan pembangunan di Indonesia terjadi pada beberapa aspek seperti perekonomian masyarakat, sumber daya manusia, sarana, prasarana serta telekomunikasi dan informasi. Akibat dari permasalahan ketidakeimbangan pembangunan tersebut maka muncullah daerah yang terkategori sebagai daerah maju dan daerah tertinggal [1].

Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional [2]. Perlunya keseimbangan pembangunan dilakukan demi meminimalisir jumlah daerah tertinggal, karena masalah ketertinggalan pembangunan suatu daerah dapat meluas menjadi masalah sosial dan ekonomi masyarakat. Maka dari itu, pentingnya pengelompokan wilayah agar dapat menentukan prioritas dalam mencapai pemerataan pembangunan secara cepat dan tepat sasaran [1].

Metode pengelompokan merupakan metode yang memi-

*Corresponding Author.

liki fungsi untuk mengelompokan data-data ke dalam *cluster* yang serupa sesuai dengan sifat dari data tersebut [3]. Salah satu metode pengelompokan yang dapat digunakan untuk data multivariat adalah metode *Model Based Clustering*. *Model Based Clustering* adalah model pengelompokan yang berbasis pada probabilitas dengan pendekatan *finite mixture* dan setiap *cluster* yang dihasilkan merepresentasikan distribusi probabilitas tertentu yang dalam penelitian ini adalah distribusi *t* multivariat. Selanjutnya *marginal contour plot* digunakan untuk memvisualisasikan kepadatan probabilitas dalam model clustering berbasis probabilitas [4].

Pada distribusi *finite mixture*, distribusi normal multivariat lebih sering dipakai dalam penelitian *Model Based Clustering*. Hal ini disebabkan karena distribusi normal multivariat memiliki proses perhitungan yang jauh lebih sederhana. Namun pada kenyataannya *outlier* pada data observasi banyak dijumpai yang mengakibatkan tidak terpenuhinya asumsi dari distribusi normal. Kemudian masalah *over-estimate* pada jumlah cluster akan muncul jika kita tetap memaksakan asumsi distribusi normal pada data yang mengandung *outlier* [5]. Mengatasi masalah tersebut dalam penelitian [6] mereka menggunakan model campuran distribusi *t* multivariat dikarenakan distribusi ini memiliki ekor yang lebih panjang dibandingkan dengan distribusi normal sehingga lebih mampu dalam mengatasi data dengan *outlier*. Implementasi model campuran distribusi *t* multivariat biasanya menggunakan algoritma *Expectation-Maximization* (EM) untuk mengestimasi model. Pendekatan ini memberikan kerangka kerja yang lebih *robust* dalam analisis *clustering* terutama pada data yang mengandung *outliers*.

[7] juga melakukan *clustering* pada data yang mengandung *outlier*, peneliti menggunakan pendekatan *Model Based Clustering* dengan asumsi *finite mixture*. Metode ini mengasumsikan bahwa beberapa distribusi probabilitas menghasilkan data dan merepresentasikan setiap *cluster* yang telah dihasilkan. Peneliti juga menggunakan kriteria *Integrated Completed Likelihood* (ICL) dalam pemilihan model terbaik. Dari [7] dalam penelitiannya menggunakan *Model Based Clustering* dengan asumsi *mixture* berdistribusi *t* multivariat, penelitian dengan metode ini menunjukkan hasil pengelompokan yang jauh lebih baik daripada metode lainnya. Kemudian Pardede (2007) dalam penelitiannya melakukan perbandingan metode *Model Based Clustering* dengan metode *k-means* dalam analisis *cluster* hasil dari penelitian tersebut menunjukkan bahwa secara umum metode *Model Based Clustering* lebih efektif memisahkan *cluster* yang saling tumpang tindih dibandingkan dengan metode *k-means*.

Berdasarkan latar belakang tersebut maka tujuan yang ingin dicapai dalam penelitian ini adalah mendapatkan hasil pengelompokan daerah di Indonesia berdasarkan indikator penetapan daerah tertinggal menggunakan *Model Based Clustering*.

2. Metode Penelitian

2.1. Lokasi dan Tempat Penelitian

Lokasi pengambilan data bertempat di Badan Pusat Statistik (BPS) dan tempat penelitian bertempat di Laboratorium Statistika Terapan, Program Studi Statistika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Tadulako.

2.2. Populasi dan Sampel Penelitian

Populasi yang digunakan dalam penelitian ini merupakan daerah kabupaten yang telah dinyatakan tertinggal. Adapun sampel yang digunakan sebanyak 62 kabupaten tertinggal di Indonesia yang telah ditetapkan melalui Peraturan Presiden pada tahun 2021.

2.3. Prosedur Pengambilan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari website resmi Badan Pusat Statistik (BPS) pada tiap-tiap kabupaten tertinggal. Adapun variabel-variabel yang digunakan dalam penelitian ini adalah sebagai berikut:

Table 1. Variabel Penelitian

Variabel	Definisi Operasional	Satuan
X_1	Produk Domestik Regional Bruto (PDRB) Per kapita	Ribuan
X_2	Persentase Pengeluaran Rumah Tangga Non Makanan	Persentase
X_3	Angka Partisipasi SMP (APS SMP)	Persentase
X_4	Angka Partisipasi SMA (APS SMA)	Persentase
X_5	Desa yang Punya Fasilitas Kesehatan	Persentase
X_6	Desa yang Punya Dokter	Persentase
X_7	Desa yang Punya SD	Persentase
X_8	Desa yang Punya SMP	Persentase
X_9	Pendapatan Asli Daerah (PAD) Per Kapita	Ribuan

2.4. Analisis Data

Analisis data dalam penelitian ini dilakukan dengan menggunakan metode *Model Based Clustering*. Analisis data dalam penelitian ini menggunakan software R. Berikut adalah tahapan yang akan dilakukan untuk menganalisis data dalam penelitian ini:

- Mengumpulkan data.
- Membuat analisis statistika deskriptif terhadap variabel-variabel yang akan digunakan.
- Melakukan pemeriksaan keberadaan *outlier* multivariat terhadap data pada variabel dengan menggunakan plot jarak Mahalanobis terhadap jarak Robust.
- Mengelompokkan objek penelitian menggunakan metode *Model Based Clustering* dengan distribusi *t* multivariat.
- Penentuan model terbaik dan jumlah cluster optimal dengan kriteria *Integrated Completed Likelihood* (ICL).
- Melakukan interpretasi terhadap hasil clustering yang didapatkan.
- Menarik kesimpulan.

3. Hasil dan Pembahasan

3.1. Analisis Deskriptif

Pada penelitian ini analisis deskriptif dilakukan untuk memudahkan peneliti dalam mengetahui dan membaca gambaran umum karakteristik dari variabel yang digunakan.

Berdasarkan data yang tersaji pada table 2 beberapa kesimpulan yang dapat diambil perihal perekonomian, pengeluaran rumah tangga, pendidikan, kesehatan dan pendapatan asli daerah.

- Pada perekonomian, PDRB digunakan sebagai indikator kesejahteraan ekonomi masyarakat, tingkat produktivitas dan

Table 2. Analisis Deskriptif

Variabel	Min	Max
X_1	6181	91782
X_1	25,91%	75,36%
X_1	16,05%	136,19%
X_1	16,05%	138,16%
X_1	0,02%	71%
X_1	0,04%	68%
X_1	0,06%	4,58%
X_1	0,03%	2,88%
X_1	58,39	630998.31

pendapatan masyarakat. Dari 62 kabupaten tertinggal di Indonesia, Kabupaten Sorong menempati urutan pertama dengan PDRB per kapita (X_1) sebesar 91782 sedangkan Kabupaten Puncak Jaya menempati urutan terakhir dengan PDRB per kapita (X_1) sebesar 6181.

- Pada pengeluaran rumah tangga non makanan menunjukkan bahwa pola konsumsi dialokasikan lebih pada kebutuhan sekunder dan tersier. Dari 62 kabupaten tertinggal di Indonesia, Kabupaten Buru Selatan menempati urutan pertama dengan persentase pengeluaran non makanan sebesar 75,36 % sedangkan Kabupaten Mamberamo Raya menempati urutan terakhir dengan persentase pengeluaran non makanan sebesar 25,91 % menunjukkan bahwa rumah tangga pada kabupaten tersebut lebih banyak menghabiskan pendapatannya untuk kebutuhan primer seperti makanan.
- Pada pendidikan, angka partisipasi SMP (X_3) tertinggi dimiliki oleh Kabupaten Malaka sebesar 136,19 % sedangkan yang terendah dimiliki oleh Kabupaten Intan Jaya sebesar 16,05 %. Kemudian untuk tingkat SMA (X_4) angka partisipasi siswa SMA (X_4) tertinggi masih dimiliki oleh Kabupaten Malaka sebesar 138, 16 % dan terendah oleh Kabupaten Nduga sebesar 16,05 %.
- Pada kesehatan, dari 62 kabupaten tertinggal di Indonesia Kabupaten Pulau Taliabu memiliki persentase desa dengan fasilitas kesehatan (X_5) tertinggi sebesar 71 % dan yang terendah dimiliki oleh Kabupaten Mamberamo Raya sebesar 0,02 %. Kemudian persentase desa yang memiliki dokter tertinggi dimiliki oleh Kabupaten Puncak sebesar 68 % dan terendah dimiliki oleh Kabupaten Waropen sebesar 0,04 %.
- Pada Infrastruktur Pendidikan, dari 62 kabupaten tertinggal di Indonesia Kabupaten Lombok Utara memiliki persentase desa yang punya SD (X_7) tertinggi sebesar 4,58 % dan yang terendah dimiliki oleh Kabupaten Puncak Jaya sebesar 0,06 %. Kemudian, persentase desa yang punya SMP (X_8) tertinggi dimiliki oleh Kabupaten Lombok Utara juga sebesar 2,88 % dan yang terendah dimiliki oleh Kabupaten Nduga sebesar 0,03 %.
- Pada Pendapatan Asli Daerah (PAD) per kapita, dari 62 kabupaten tertinggal di Indonesia Kabupaten Kepulauan Mentawai memiliki PAD Per Kapita (X_9) tertinggi sebesar 630.998 dan yang terkecil dimiliki oleh Kabupaten Yahukimo sebesar 58,39. PAD merupakan sumber pendapatan asli daerah semakin tinggi PAD menunjukkan tingkat kemandirian finansial yang tinggi. PAD per kapita menunjukkan kontribusi setiap Individu di daerah tersebut.

3.2. Deteksi Outlier Multivariat

Pada penelitian deteksi outlier multivariat perlu dilakukan untuk membuktikan bahwa asumsi awal dari data indikator penetapan daerah tertinggal cenderung memiliki outlier, sehingga Model Based Clustering dengan distribusi finite mixture t multivariat tepat diterapkan pada penelitian ini agar mendapatkan hasil yang robust. Beberapa langkah yang akan dilakukan dalam proses pendeteksian ini yaitu melakukan perhitungan jarak Mahalanobis, perhitungan jarak Robust dan membuat diagnostic plot. Jarak Mahalanobis merupakan salah satu metode yang dapat digunakan untuk memeriksa keberadaan outlier pada data multivariat. Jarak mahalanobis merupakan jarak antara masing-masing vektor data dengan titik pusat data atau vektor rata-rata. Adapun perhitungan jarak Mahalanobis adalah sebagai berikut:

$$MD_i = \sqrt{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})}, i = 1, 2, \dots, n \quad (1)$$

Jarak mahalanobis pengamatan ke- i disimbolkan dengan MD_i , vector rata-rata dari sampel disimbolkan dengan \bar{x} sedangkan invers dari matriks varians kovarians sampel disimbolkan dengan S^{-1} . Nilai dari \bar{x} dan S^{-1} seringkali berubah-ubah karena kesensitifan dari jarak mahalanobis, hal ini dapat mengakibatkan data *non-outlier* berubah menjadi data *outlier*. Untuk memecahkan masalah tersebut maka penggunaan jarak *robust* adalah solusinya. Hal ini dikarenakan jarak *robust* dinilai lebih kuat dalam mengatasi *outlier*. Estimasi rata-rata dan matriks varians kovarians pada jarak *robust* akan dihitung menggunakan *Minimum Covariance Determinant* (MCD). Suatu ruang yang memiliki dimensi R^p yang di dalamnya terdapat sampel acak dari sejumlah n jika diasumsikan sebagai x_1, \dots, x_n . Oleh karenanya estimasi MCD bertujuan mendapatkan h_i subset sampel dari n objek observasi dengan ukuran maksimum $h_i = (\frac{n+p+1}{2})$ yang memiliki determinan matriks varians kovarians terkecil.

$$MCD = \min\{det(C)_1, det(C)_2, \dots, det(C)_n, \} \quad (2)$$

Melalui rumus berikut jarak Robust dapat diperoleh [8]:

$$RD_i = \sqrt{[(x_1 - T)^T C^{-1} (x_1 - T)]}, i = 1, 2, \dots, n \quad (3)$$

Dimana penaksir rata-rata sampel yang Robust berdasarkan MCD (rata-rata dari h subset sampel) disimbolkan dengan T [8]:

$$T_i = \frac{1}{h_i} \sum_{j=1}^{h_i} x_j, \quad (4)$$

Dimana penaksir matriks kovarians yang Robust dari h subset sampel disimbolkan dengan C [8]:

$$C_i = \frac{1}{h_i - 1} \sum_{j=1}^{h_i} (x_j - T_i)(x_j - T_i)^T \quad (5)$$

[8] memanfaatkan kuantil dari distribusi X_p^2 dan *Robust Distance* untuk melakukan pendeteksian outlier multivariat sebagai nilai *cut off*. Data ke dengan nilai $RD > \sqrt{X_{p;1-\frac{\alpha}{2}}^2}$ didefinisikan sebagai *outlier* multivariat. Kemudian pendeteksian dan klasifikasi *outlier* akan dilakukan secara visual dengan menggunakan *diagnostic plot*. Pembuatan *diagnostic plot* akan memanfaatkan fungsi dari *distance-distance plot* (dd plot) yang mana plot yang terbentuk adalah perpaduan garis antara *Mahalanobis Distance* terhadap *Robust Distance*. Berikut visualisasi untuk pendeteksian *outlier* multivariat untuk semua variabel.

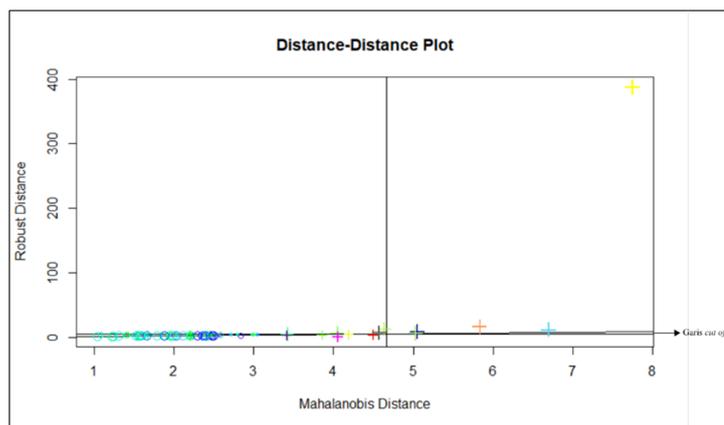


Figure 1. Plot Mahalanobis Distance terhadap Robust Distance

Berdasarkan Gambar 1 diketahui bahwa data outlier terindikasikan sebagai titik-titik yang terletak jauh dari garis lurus nilai cut off yang memiliki nilai sebesar 4,36. Melalui plot di atas terlihat bahwa terdapat delapan daerah yang terdeteksi sebagai outlier pada indikator daerah tertinggal. Kedelapan daerah tersebut dengan masing-masing nilai mahalanobis distance dan robust distance sebagai nilai x dan y adalah Kepulauan Mentawai (4,56; 7,18) ditandai dengan simbol plus (+) berwarna merah, Musi Rawas Utara (5,83; 16,54) ditandai dengan simbol plus (+) berwarna orange, Lombok Utara (6,69; 11,87) ditandai dengan simbol plus (+) berwarna biru muda, Kepulauan Aru (4,63; 12,91) ditandai dengan simbol plus (+) berwarna abu-abu, Buru Selatan (5,03; 6,25) ditandai dengan simbol plus (+) berwarna hijau, Pulau Taliabu (7,74; 387,54) ditandai dengan simbol plus (+) berwarna kuning, Supiori (4,04; 6,27) ditandai dengan simbol plus (+) berwarna pink, Mamberamo Raya (3,41; 4,66) ditandai dengan simbol plus (+) berwarna ungu dan Puncak (5,05 & 8,44) ditandai dengan simbol plus (+) berwarna biru tua.

3.3. Pengelompokan Menggunakan Model Based Clustering

Model Based Clustering mengasumsikan bahwa sebuah populasi terdiri dari subpopulasi dengan distribusi tertentu serta mempunyai perbedaan antar parameter. Keseluruhan subpopulasi mempunyai distribusi peluang *mixture* dengan proporsi yang berbeda untuk setiap subpopulasi. Asumsi ini mengarahkan pada model probabilitas matematika yaitu model *finite mixture*. Terdapat dua langkah utama yang penting untuk dilakukan dalam model *finite mixture* yaitu penaksiran parameter Model Based Clustering dan pemilihan model terbaik. Algoritma Expectation Maximization (EM) dapat digunakan untuk menaksir parameter dari Model Based Clustering [9].

Kemudian kriteria *Integrated Completed Likelihood* (ICL) akan digunakan dalam pemilihan model terbaik [10]. Pengembangan kerangka Model Based Clustering didasarkan pada dekomposisi *eigenvalue* dari matriks varians kovarians (Σ_g) sebagai berikut:

$$\Sigma_g = \lambda_g, D_g, A_g, v_g \tag{6}$$

dimana:

- λ_g : nilai scalar yang menunjukkan volume elips
- D_g : matriks orthogonal *eigenvector* yang merupakan orientasi dari komponen utama Σ_g
- A_g : diagonal matriks dengan elemen proporsional pada *eigen value* Σ_g dan menunjukkan kontur dari fungsi kepadatannya
- v_g : Derajat kebebasan

Model Based Clustering dengan distribusi *t* multivariat mampu mengidentifikasi sebanyak 28 model dengan 9 cluster maksimal [11]. Data Indikator Penetapan Daerah Tertinggal yang digunakan pada penelitian ini memiliki satuan yang berbeda-beda sehingga perlu dilakukan standarisasi. Fungsi *scale* pada *package* "teigen" akan secara default melakukan standarisasi pada data, sehingga peneliti tidak perlu melakukan standarisasi data secara terpisah.

3.4. Estimasi Parameter dengan Algoritma Expectation Maximization (EM)

Algoritma EM untuk pengelompokan dan klasifikasi membutuhkan inisialisasi parameter model (yang menggunakan perhitungan nilai yang diharapkan untuk menginisialisasi z_{ig}). Model Based Clustering menghasilkan cluster dengan bentuk yang berbeda-beda sehingga nilai parameter untuk masing-masing cluster juga menghasilkan nilai yang berbeda pula. Berikut nilai estimasi parameter untuk masing-masing matriks kovarians cluster: Nilai dari matriks varians-kovarian pada cluster 1 (Σ_1) diperoleh pada pers 7. Nilai scalar yang menunjukkan volume elips untuk cluster 1 (λ_1) adalah sebagai berikut 0,036. Nilai dari matriks orthogonal *eigenvector* untuk cluster 1 (D_1) dapat dilihat pada persamaan 8. Selanjutnya, Nilai dari diagonal matriks untuk cluster 1 (A_1) juga dapat dilihat pada persamaan 9. Nilai dari derajat kebebasan untuk cluster 2 (v_1) adalah 2,122. Nilai dari matriks varians-kovarian pada cluster 2 (Σ_2) diperoleh pada persamaan 10. Nilai scalar yang menunjukkan volume elips untuk cluster 2 (λ_2) 0,071. Nilai dari matriks orthogonal *eigenvector* untuk cluster 2 D_2 dapat dilihat pada persamaan 11. Nilai dari diagonal matriks untuk cluster 2 (A_2) dapat dilihat pada persamaan 12. Nilai dari derajat kebebasan untuk cluster 2 (v_2) adalah 2,122.

3.5. Pemilihan Model Terbaik dan Jumlah Cluster Optimal dengan Kriteria Integrated Completed Likelihood (ICL)

Pengelompokan dengan metode Model Based Clustering menghasilkan 28 model dengan jumlah cluster sebanyak 9. Berikut 28 model dan 9 cluster dengan masing-masing nilai ICL yang telah dihasilkan.

Tabel 3 menampilkan nilai ICL dari 28 model dan 9. Kemudian dari banyaknya model dan cluster yang telah dihasilkan akan dipilih salah satunya dengan melihat nilai ICL terbesar. Model dan jumlah cluster yang menghasilkan keterangan "-Inf" mengartikan bahwa iterasi estimasi parameter pada model dan jumlah cluster tersebut belum mencapai konvergen dan bernilai negatif. Maka model UUUC dengan jumlah cluster optimal sebanyak 2 cluster dengan nilai ICL sebesar -882,7315 terpilih menjadi model terbaik dan jumlah cluster optimal.

Berdasarkan Persamaan 6 diketahui model UUUC menunjukkan λ_g berlabel "U" nconstrained dengan makna bahwa kedua cluster mempunyai volume elips yang berbeda. D_g berlabel "U" nconstrained dengan makna bahwa orientasi dari kedua cluster berbeda. A_g berlabel "U" nconstrained dengan makna bahwa ke-

$$\Sigma_1 = \begin{bmatrix} 0,237 & 0,103 & -0,002 & 0,030 & 0,002443 & -0,001615 & 0,047 & -0,027 & 0,1101 \\ 0,103 & 0,278 & 0,053 & 0,093 & -0,0001498 & -0,0005446 & -0,073 & -0,060 & 0,09295 \\ -0,002 & 0,053 & 0,140 & 0,173 & -0,0006445 & -0,0002015 & -0,034 & -0,034 & 0,008769 \\ 0,030 & 0,093 & 0,173 & 0,380 & -0,001113 & -0,001566 & -0,128 & -0,099 & 0,08267 \\ 0,002 & -1,498 & -0,0006445 & -0,001 & 0,0006993 & 0,00003916 & 0,006 & 0,003 & 0,003683 \\ -0,001 & -0,0005446 & -0,0002015 & -0,001 & 0,000039 & 0,00007734 & 0,7678 & 0,9840 & -0,00004076 \\ 0,047 & -0,073 & -0,034 & -0,128 & 0,006069 & 0,0007678 & 0,365 & 0,248 & 0,001475 \\ -0,027 & -0,060 & -0,034 & -0,099 & 0,003340 & 0,0009840 & 0,248 & 0,229 & 0,001240 \\ 0,110 & 0,092 & -0,008 & 0,082 & 0,003683 & -0,00004076 & 0,014 & 0,001 & 0,4104 \end{bmatrix} \quad (7)$$

$$D_1 = \begin{bmatrix} 0,237 & 0,103 & -0,002 & 0,030 & 0,002443 & -0,001615 & 0,047 & -0,027 & 0,1101 \\ 0,103 & 0,278 & 0,053 & 0,093 & -0,0001498 & -0,0005446 & -0,073 & -0,060 & 0,09295 \\ -0,002 & 0,053 & 0,140 & 0,173 & -0,0006445 & -0,0002015 & -0,034 & -0,034 & 0,008769 \\ 0,030 & 0,093 & 0,173 & 0,380 & -0,001113 & -0,001566 & -0,128 & -0,099 & 0,08267 \\ 0,002 & -1,498 & -0,0006445 & -0,001 & 0,0006993 & 0,00003916 & 0,006 & 0,003 & 0,003683 \\ -0,001 & -0,0005446 & -0,0002015 & -0,001 & 0,000039 & 0,00007734 & 0,7678 & 0,9840 & -0,00004076 \\ 0,047 & -0,073 & -0,034 & -0,128 & 0,006069 & 0,0007678 & 0,365 & 0,248 & 0,001475 \\ -0,027 & -0,060 & -0,034 & -0,099 & 0,003340 & 0,0009840 & 0,248 & 0,229 & 0,001240 \\ 0,110 & 0,092 & -0,008 & 0,082 & 0,003683 & -0,00004076 & 0,014 & 0,001 & 0,4104 \end{bmatrix} \quad (8)$$

$$D_1 = \begin{bmatrix} 20,584 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 14,396 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8,578 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6,532 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4,005 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1,258 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0,550 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,015 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,002 \end{bmatrix} \quad (9)$$

$$D_1 = \begin{bmatrix} 0,849 & 0,325 & 0,541 & 0,456 & 0,007 & 0,262 & 0,157 & 0,110 & -0,031 \\ 0,325 & 0,792 & 0,522 & 0,428 & 0,005 & 0,072 & 0,153 & 0,092 & -0,042 \\ 0,541 & 0,522 & 1,533 & 1,208 & 0,008 & 0,159 & 0,110 & 0,110 & -0,008 \\ 0,456 & 0,428 & 1,208 & 1,053 & 0,006 & 0,113 & 0,095 & 0,098 & -0,020 \\ 0,007 & 0,005 & 0,008 & 0,006 & 4,753 \times 10^{-4} & 0,012 & 0,005 & 0,003 & 5,414 \times 10^{-4} \\ 0,262 & 0,072 & 0,159 & 0,113 & 0,012 & 0,530 & 0,179 & 0,092 & -0,008 \\ 0,157 & 0,153 & 0,110 & 0,095 & 0,005 & 0,179 & 0,123 & 0,067 & -0,017 \\ 0,110 & 0,092 & 0,110 & 0,098 & 0,003 & 0,092 & 0,067 & 0,050 & -0,006 \\ -0,031 & -0,042 & -0,008 & -0,020 & -5,414 \times 10^{-4} & -0,008 & -0,017 & -0,006 & 0,037 \end{bmatrix} \quad (10)$$

$$D_2 = \begin{bmatrix} -0,342 & 0,607 & -0,142 & -0,699 & -0,059 & 0,0307 & -0,011 & 0,014 & -0,002 \\ -0,315 & 0,217 & 0,886 & 0,178 & -0,098 & 0,154 & -0,045 & -0,030 & 6,431 \times 10^{-4} \\ -0,672 & -0,332 & -0,173 & 0,102 & -0,517 & -0,298 & 0,199 & -0,020 & 0,002 \\ -0,552 & -0,272 & -0,132 & 0,024 & 0,654 & 0,350 & -0,215 & 0,063 & -0,001 \\ -0,005 & 0,012 & -0,006 & 0,016 & 0,005 & -0,009 & 0,001 & -0,029 & -0,999 \\ -0,118 & 0,560 & -0,378 & 0,643 & -0,123 & 0,308 & 0,007 & -0,064 & 0,019 \\ -0,085 & 0,260 & 0,035 & 0,214 & 0,337 & -0,656 & 0,020 & 0,575 & -0,002 \\ -0,069 & 0,124 & 0,004 & 0,084 & 0,267 & -0,480 & -0,177 & -0,798 & 0,032 \\ 0,014 & -0,039 & -0,050 & 0,009 & -0,299 & -0,072 & -0,937 & 0,145 & -0,006 \end{bmatrix} \quad (11)$$

$$A_2 = \begin{bmatrix} 43,540 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 11,207 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7,750 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5,161 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,982 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,606 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0,433 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,107 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,001 \end{bmatrix} \tag{12}$$

Table 3. Nilai ICL Masing-masing Model dan Cluster

Model	G=1	G=2	G=3	...	G=9
UUUU	-1046,599	-1050,0494	-Inf	...	-Inf
UUUC	-Inf	-882,7315	-Inf	...	-Inf
CUCU	-Inf	-890,5805	-Inf	...	-Inf
CUCC	-Inf	-886,8637	-Inf	...	-Inf
CUUU	-Inf	-888,2703	-Inf	...	-Inf
CUUC	-Inf	-885,0892	-Inf	...	-Inf
CCCU	-Inf	-1042,1644	-Inf	...	-Inf
CCCC	-Inf	-1037,941	-Inf	...	-Inf
CIUU	-1190,838	-898,7658	-Inf	...	-Inf
CIUC	-Inf	-894,714	-Inf	...	-Inf
CICU	-Inf	-1098,5255	-Inf	...	-Inf
CICC	-Inf	-1101,8977	-Inf	...	-Inf
UIIU	-1448,515	-1345,6064	-Inf	...	-Inf
UIIC	-Inf	-1346,6623	-Inf	...	-Inf
CIIU	-Inf	-1351,2377	-Inf	...	-Inf
CIIC	-Inf	-1348,4408	-Inf	...	-Inf
UIUU	-Inf	-891,5645	-Inf	...	-Inf
UIUC	-Inf	-889,3525	-Inf	...	-Inf
UCCU	-Inf	-1045,1307	-Inf	...	-Inf
UCCC	-Inf	-1040,6941	-Inf	...	-Inf
UUUC	-Inf	-Inf	-Inf	...	-Inf
UUCC	-Inf	-1104,8631	-Inf	...	-Inf
UICU	-Inf	-1092,8594	-Inf	...	-Inf
UICC	-Inf	-1096,5636	-Inf	...	-Inf
UCUU	-Inf	-1040,1662	-Inf	...	-Inf
UCUC	-Inf	-1037,9542	-Inf	...	-Inf
CCUU	-Inf	-1047,3476	-Inf	...	-Inf
CCUC	-Inf	-1043,3144	-Inf	...	-Inf

Tolikara, Keerom, Waropen, Supiori, Mamberamo Raya, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Inyan Jaya, Deiyai

Berikut marginal contour plot yang merupakan visualisasi yang dapat merepresentasikan struktur matriks varians kovarians $\Sigma_g = \lambda D_g A_v$ pengelompokan daerah tertinggal di Indonesia.

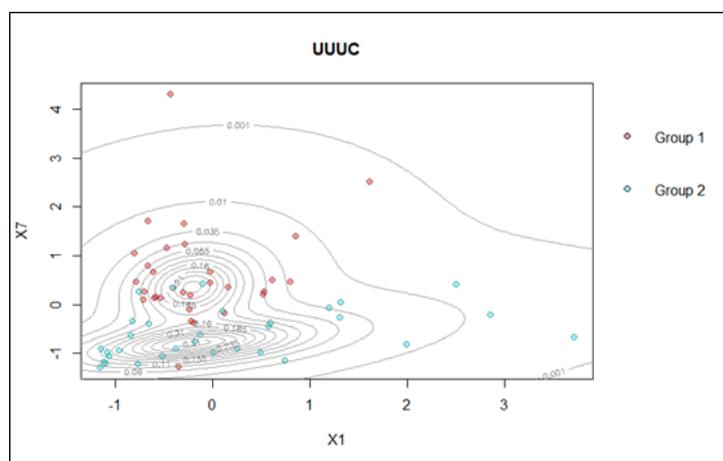


Figure 2. Marginal Contour Plot $X_1 X_7$

dua cluster mempunyai kontur yang berbeda. v berlabel “C” on-strained dengan makna bahwa kedua cluster mempunyai derajat bebas yang sama. Karena volume, orientasi dan kontur kedua cluster berbeda, maka kedua cluster mempunyai matriks kovarians (Σ_g) yang berbeda pula. Model UUUC merepresentasikan struktur matriks varians kovarians $\Sigma_g = \lambda D_g A_v$. Adapun hasil pembagian anggota masing-masing cluster adalah sebagai berikut.

1. Cluster 1 terdapat 30 kabupaten yaitu: Nias, Nias Selatan, Nias Utara, Nias Barat, Kepulauan Mentawai, Musi Rawas Utara, Pesisir Barat, Lombok Utara, Sumba Barat, Sumba Timur, Kupang, Timor Tengah Selatan, Belu, Alor, Lembata, Rote Ndao, Sumba Tengah, Sumba Barat Daya, Manggarai Timur, Sabu Raijua, Malaka, Donggala, Sigi, Tojo Una-Una, Maluku Tenggara Barat, Kepulauan Aru, Seram Bagian Barat, Seram Bagian Timur, Maluku Barat Daya, Kepulauan Sula.
2. Cluster 2 terdapat 32 kabupaten yaitu: Pulau Taliabu, Buru Selatan, Teluk Wondama, Teluk Bintuni, Sorong Selatan, Sorong, Tambrauw, Maybrat, Manokwari Selatan, Pegunungan Arfak, Jayawijaya, Nabire, Paniai, Puncak Jaya, Boven Digoel, Mappi, Asmat, Yahukimo, Pegunungan Bintang,

Gambar 2 menampilkan hasil visualisasi dari model terbaik yaitu UUUC dan jumlah cluster optimal sebanyak 2 cluster telah dihasilkan dari metode Model Based Clustering dan dipilih berdasarkan kriteria ICL. Kombinasi dari kedua variabel yang terpilih yaitu PDRB Per Kapita (X_1), Desa Yang Punya SD (X_7) pada Gambar 2 memberikan titik pusat jumlah kontur yang sesuai dengan banyaknya cluster serta menghasilkan bentuk kontur yang paling baik dan titik-titik yang mewakili anggota cluster tidak tumpang tindih. Kemudian perbedaan cluster pada Marginal Contour Plot terlihat melalui perbedaan warna pada titik-titik yang dihasilkan. Cluster pertama ditandai dengan titik-titik berwarna merah sedangkan cluster kedua ditandai dengan titik-titik berwarna biru.

3.6. Interpretasi Cluster Pada Model Based Clustering

Tahap selanjutnya setelah mendapatkan jumlah cluster dan anggotanya yaitu menginterpretasikan hasil cluster menggunakan nilai rata-rata dari setiap variabel pada masing-masing cluster. Adapun nilai rata-rata cluster yang didapatkan disajikan pada table 4 sebagai berikut: Nilai rata-rata yang ditampilkan dalam Tabel 4 memberikan gambaran awal mengenai perbedaan antara cluster 1 dan cluster 2. Terlihat bahwa nilai rata-rata terendah diberikan warna kuning sedangkan nilai rata-rata tertinggi diberikan warna

Table 4. Nilai Rata-rata Cluster

Variabel	Cluster 1	Cluster 2
X_1	23662	29056,8125
X_2	43,69666667	41,2046875
X_3	75,78366667	62,4228125
X_4	62,651	46,310625
X_5	0,541420371	2,538590718
X_6	0,424856724	15,84536504
X_7	1,617605009	0,625406146
X_8	0,650957297	0,180988044
X_9	28553375872	6836580590

hijau. Berdasarkan tabel 4 pula, maka karakteristik dari masing-masing cluster dapat diinterpretasikan sebagai berikut:

1. Cluster 1 memiliki PAD per kapita (X_9) yang tinggi tetapi PDRB per kapita (X_1) yang rendah hal ini mengartikan bahwa 30 kabupaten yang tergolong dalam cluster 1 merupakan daerah yang memiliki sumber pendapatan daerah yang lebih besar tetapi kegiatan ekonomi masyarakatnya kecil. Walaupun begitu 30 kabupaten tersebut lebih banyak mengalokasikan pengeluarannya ke kebutuhan non-makanan (X_2) yang mengartikan bahwa masyarakatnya cukup sejahtera. Kemudian Partisipasi Pendidikan yang lebih tinggi (X_3 & X_4) dan didukung dengan banyaknya desa yang memiliki SD (X_7) dan SMP (X_8) mengartikan bahwa 30 kabupaten tersebut memiliki akses pendidikan yang lebih baik dan sektor pendidikan yang lebih maju.
2. Cluster 2 memiliki PDRB per kapita (X_1) yang tinggi tetapi PAD per kapita (X_9) nya rendah, terbalik dengan cluster 1, hal ini mengartikan bahwa 32 kabupaten yang tergolong di dalamnya memiliki kegiatan ekonomi yang lebih tinggi tetapi sumber pendapatan daerah yang lebih rendah sehingga memiliki ketergantungan finansial atau kepada pemerintah pusat. Kemudian masyarakatnya lebih banyak menghabiskan pendapatannya untuk kebutuhan primer seperti makanan dibandingkan untuk kebutuhan sekunder seperti Pendidikan ditandai dengan rendahnya Persentase Pengeluaran Non-Makanan (X_2). Hal ini diperparah dengan Partisipasi SMP (X_3) dan SMA (X_4) yang rendah dan sedikitnya desa yang memiliki SD (X_7) dan SMP (X_8).

3.7. Visualisasi Cluster

Visualisasi dilakukan menggunakan peta dari 62 daerah di Indonesia, visualisasi dibuat menggunakan software R. Berikut hasil pemetaan dari clustering dengan metode Model Based Clustering.

Gambar 3 menampilkan gambaran visualisasi peta dari masing-masing anggota cluster 1 dan 2 yang telah dihasilkan melalui pengelompokan dengan metode Model Based Clustering. Identitas anggota masing-masing cluster ditandai dengan warna yang berbeda. Yaitu cluster 1 ditandai dengan daerah yang berwarna kuning, cluster 2 ditandai dengan daerah yang berwarna merah. Sedangkan daerah berwarna abu-abu bukan merupakan daerah tertinggal, daerah-daerah tersebut ditambahkan untuk membuat visualisasi peta Indonesia menjadi sempurna.

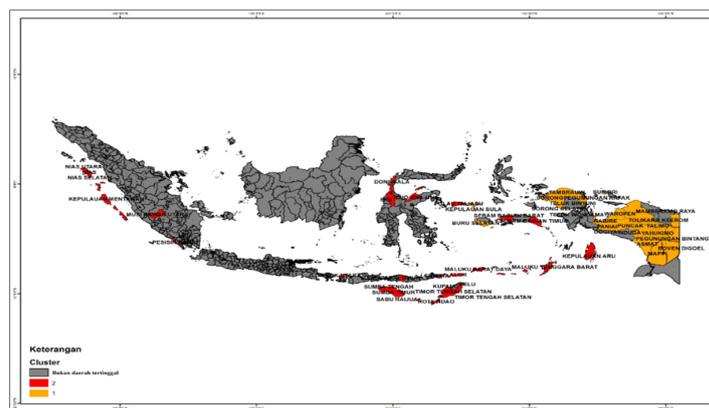


Figure 3. Peta Pengelompokan 62 Daerah Tertinggal Di Indonesia

4. Kesimpulan

Berdasarkan pada pembahasan bab sebelumnya, maka kesimpulan pada penelitian ini adalah pengelompokan daerah di Indonesia dengan berdasarkan Indikator Penetapan Daerah Tertinggal dengan menggunakan Model Based Clustering menggunakan kriteria ICL sebagai penentu model terbaik dan jumlah cluster optimal maka dihasilkan model UUUC dan 2 cluster optimal. Dimana cluster 1 mempunyai 30 kabupaten dengan tingkat ketertinggalan daerah yang rendah dan cluster 2 dengan 32 kabupaten dengan tingkat ketertinggalan daerah yang tinggi.

References

- [1] E. A. Sari, M. T. B. Saragih, I. A. Shariati, S. Sofyan, R. Al Baihaqi, and R. Nooraeni, "Klasifikasi kabupaten tertinggal di kawasan timur indonesia dengan support vector machine," *JIKO (Jurnal Informatika dan Komputer)*, vol. 3, no. 3, pp. 188–195, 2020.
- [2] Peraturan Presiden Republik Indonesia Nomor 63 Tahun 2020 Tentang Penetapan Daerah Tertinggal Tahun 2020-2024, 2020.
- [3] J. Han and M. Kamber, "Data mining: Concepts and techniques slides for textbook—chapter 7," *Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada, October*, vol. 15, 2011.
- [4] P. D. McNicholas, "Model-based clustering," *Journal of Classification*, vol. 33, pp. 331–373, 2016.
- [5] A. Cozzini, A. Jasra, and G. Montana, "Model-based clustering with gene ranking using penalized mixtures of heavy-tailed distributions," *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 03, p. 1341007, 2013.
- [6] G. McLachlan, "Finite mixture models," *A wiley-interscience publication*, 2000.
- [7] F. Damayanti, "Model based clustering mixture t-multivariat dengan kriteria integrated classification likelihood (pengelompokan provinsi di indonesia menurut capaian berkelanjutan tahun 2011)," *Institut Teknologi Sepuluh Nopember.Surabaya*, 2015.
- [8] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.
- [9] M. Agustini, "Model-based clustering dengan distribusi t multivariat menggunakan kriteria integrated completed likelihood dan minimum message length," *Institut Teknologi Sepuluh Nopember. Surabaya*, 2017.
- [10] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2002.
- [11] J. L. Andrews, J. R. Wickins, N. M. Boers, and P. D. McNicholas, "teigen: An r package for model-based clustering and classification via the multivariate t distribution," *Journal of Statistical Software*, vol. 83, pp. 1–32, 2018.