PENERAPAN ALGORITMA *EXPECTATION-MAXIMIZATION* PADA PEMODELAN NORMAL MIXTURE NILAI TAMBAH PRODUK DOMESTIK BRUTO DI SETIAP NEGARA

Melva Hilda Stephanie Situmorang[1)], Irwan Susanto[2)] Sri Sulistijowati Handajani[3)]
[1]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
email: stephaniemelva@student.uns.ac.id
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
email: irwansusanto@staff.uns.ac.id
[3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
email: rr_ssh@staff.uns.ac.id

ABSTRAK

Pemodelan finite mixture merupakan salah satu metode klaster yang didasarkan pada representasi fungsi distribusi probabilitas dan berlaku pada distribusi diskrit maupun kontinu Penggunaan model finite mixture dapat digunakan pada data yang memiliki pola multimodal yang diindikasikan sebagai adanya klaster yang berbeda pada data dengan jumlah puncak pada histogram lebih dari satu dan tidak terpenuhinya uji signifikasi pola unimodal. Dalam paper ini, model finite mixture digunakan untuk memodelkan data nilai tambah produk domestik bruto di setiap negara pada tahun 2017. Parameter dari hasil model finite mixture diestimasi menggunakan algoritma Expectation-Maximization (EM) dengan inisiasi awal menggunakan K-Means. Model finite harus diuji signifikansi menggunakan bootstrap likelihood ratio statistics untuk mengetahui model finite mixture yang sesuai, kemudian menggunakan nilai Akaike Information Criterion (AIC) dan Bayesian Information Criterion (BIC) untuk mengetahui jumlah klaster terbaik. Berdasarkan hasil analisis, model normal mixture dengan dua komponen dapat digunakan untuk memodelkan data nilai tambah produk domestik bruto (PDB) di setiap negara pada tahun 2017.

Kata Kunci: *Mixture Model, Algoritma EM, Maximum Likelihood, nilai tambah PDB*

1. PENDAHULUAN

Produk Domestik Bruto (PDB) merupakan jumlah nilai tambah yang dihasilkan oleh seluruh unit usaha dalam suatu negara atau merupakan jumlah nilai barang dan jasa akhir yang dihasilkan oleh seluruh unit ekonomi dan berfungsi sebagai indikator kesehatan ekonomi suatu negara. Nilai tambah merupakan selisih antara nilai produksi (output) dan biaya yang habis dipakai selama proses produksi dari suatu produk, baik barang maupun jasa. Mengukur perdagangan dalam nilai tambah adalah cara memperkirakan berbagai sumber (berdasarkan negara dan industri) yang berkontribusi pada nilai tambah di sepanjang rantai pasokan internasional.

Nilai tambah suatu industri adalah kontribusi industri swasta atau sektor pemerintah terhadap PDB keseluruhan. Komponen nilai tambah terdiri dari kompensasi karyawan, pajak atas produksi dan impor dikurangi subsidi, dan surplus operasi bruto.

Thrun (2019) telah melakukan penelitian sebelumnya mengenai analisis klaster PDB per kapita berbasis jarak dalam peta topografi menggunakan *Databionic Swarm* (DBS) yang teridentifikasi dua klaster dalam peta topografi.

2. KAJIAN LITERATUR

2.1. Model *Finite Mixture*

Model *finite mixture* didasarkan pada representasi fungsi distribusi probabilitas (kumulatif) dan berlaku pada distribusi diskrit maupun kontinu. Vektor variabel random $x = [x_1, x_2, ..., x_n]^T$ yang bertipe diskrit maupun kontinu berasal dari distribusi *finite mixture* jika fungsi

densitas probabilitas $g(x_i)$ didefinisikan sebagai:

$$g(x_i) = w_1 f_1(x_1) + \ldots + w_k f_k(x_k)$$

Untuk $x_i, i = 1,2,\ldots,n$ dengan $f_k(x_i)$ adalah fungsi densitas probabilitas untuk semua $k = 1,2,\ldots,K$, dengan k adalah banyaknya komponen *mixture*. Vektor $\mathrm{w} = [w_1, w_2, \ldots, w_k]^T$ disebut vektor parameter bobot dari distribusi *finite mixture*, dengan parameter w_k adalah parameter *weight* atau bobot. Nilai-nilai w harus memenuhi $0 < w_k < 1$ dan $\sum w_k = 1$.

Jika diasumsikan semua komponen distribusi *finite mixture* berasal dari distribusi probabilitas yang memiliki vektor parameter θ, maka densitas *mixture* dapat ditulis dengan

$$g(x_i|\psi) = w_1 f_1(x_1|\theta_1) + \cdots + w_k f_k(x_k|\theta_k)$$

$$= \sum_{k}^{K} w_k f_k(x_k|\theta_k) \qquad (1)$$

dengan $\psi = [w, \theta]^T$, $\theta = [\theta_1, \ldots, \theta_k]^T$. (Fruhwirth-Schnatter, 2006). Model *finite mixture* adalah model statistika yang mengimplementasikan konsep distribusi *finite mixture* dalam pemodelannya.

2.2. Metode Estimasi Maksimum Likelihood.

Metode Estimasi Maksimum Likelihood dapat digunakan untuk memperkirakan parameter model dalam model *mixture*. Pada distribusi yang umum, estimasi parameter maksimum likelihood dihitung dari rumus sederhana yang melibatkan data.

Misal diberikan observasi x_i, *i=1, 2, ..., n* saling independent dan $\psi = [w, \theta]^T$, maka fungsi likelihood model *finite mixture* (1) didefinisikan sebagai berikut

$$L(\psi) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} w_k f_k(x_i|\theta_k) \right] \qquad (2)$$

Penduga maksimum likelihood $\hat{\psi}$ dalam pendekatan metode estimasi maksimum likelihood untuk estimasi parameter model *finite mixture* merupakan penyelesaian dari persamaan

$$\frac{\partial \ell(\psi)}{\partial(\psi)} = 0 \qquad (3)$$

dengan

$$\ell(\psi) = ln\, \mathrm{L}(\psi) = \sum_{i=1}^{n} ln \left(\sum_{k=1}^{K} w_k f_k(x_i|\theta_k) \right)$$

merupakan fungsi log-likelihood observasi (McLachlan dan Peel, 2000). Proses persamaan (3) tidak mudah dilakukan karena fungsi likelihood (2) tidak dalam bentuk *closed-form* (Celeux, 2018). Untuk menyelesaikan permasalahan tersebut digunakan algoritma *expectation-maximization* (EM).

2.3. Algoritma *Expectation Maximization* (EM)

1. Tahap *expectation* (E)

Pada tahap *expectation* dilakukan evaluasi hasil klaster dari inisiasi awal menggunakan parameternya dengan penduga $\hat{z}_{ik}{}^{(s)}$ yang merupakan probabilitas observasi dari x_i,

$$\hat{z}_{ik}{}^{(s)} = \frac{w_k{}^{(s)} f_k\left(x_i|\theta_k{}^{(s)}\right)}{\sum_{h=1}^{k} w_k{}^{(s)} f_k\left(x_i|\theta_k{}^{(s)}\right)}$$

2. Tahap *maximization* (M)

Pada tahap *maximization* diperkirakan kembali parameter $\hat{\theta}_k^{(s+1)}$ yang merupakan penyelesaian dari

$$\frac{\partial}{\partial\theta}\left\{ \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}{}^{(s)}\, ln[w_k f_k(x_i|\theta_k)] \right\} = 0$$

dan menduga parameter bobot w_k saat iterasi ke-(s+1) dengan

$$\hat{w}_k{}^{(s+1)} = \frac{\sum_{i=1}^{n} z_{ik}{}^{(s)}}{n}$$

(Dempster et all., 1977).

2.4. Uji *Goodness of Fit*.

Uji *Goodness of Fit* dilakukan untuk mengidentifikasi data berdistribusi univariat multimodal, dengan langkah-langkah sebagai berikut:

1. Menentukan uji hipotesis

H_0: Pola data mengikuti distribusi normal unimodal
H_1: Pola data tidak mengikuti distribusi normal unimodal

2. Menentukan tingkat signifikansi ($\propto$)
3. Menghitung statistik uji

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1]\left[\ln\left(F(Z_i)\right) + \ln\left(1-F(Z_{n+1-i})\right)\right]$$

dengan

$$Z_i = \frac{x_i - \bar{x}}{s}$$

A : statistik uji untuk metode Anderson-Darling
n : ukuran sampel
Z_i : data x_i yang telah di standarisasi
x_i : data ke-i yang telah diurutkan
$\bar{x}$: rata-rata data yang telah diurutkan
s : standar deviasi data
$F(Z_i)$: nilai fungsi distribusi kumulatif normal baku di Z_i

Modifikasi dari metode Anderson-Darling menggunakan rumus di bawah ini:

$$A^* = A\left(1 + \frac{0{,}75}{n} + \frac{2{,}25}{n^2}\right)$$

dan

$$c_\alpha = a_\alpha\left(1 - \frac{b_\alpha}{n} - \frac{d_\alpha}{n^2}\right)$$

dengan $a_\alpha, b_\alpha, d_\alpha$ ditunjukkan dalam tabel nilai kritis Anderson-Darling

4. Menentukan daerah kritis
   H_0 ditolak jika $A^* > c_\alpha$ dengan $P_{value} < \propto$
5. Menentukan kesimpulan

2.5. Uji Signifikansi Model *Finite Mixture*.

Uji signifikansi untuk model *finite mixture* dibuat berbasis *bootstrap likelihood ratio statistics* untuk mengetahui model *finite mixture* yang sesuai dalam pemodelan data (Yu, 2018). Uji hipotesisnya adalah sebagai berikut:

1. Menentukan uji hipotesis
   H_0: $K = K_0$ (Model *finite mixture* memiliki sebanyak K_0 komponen *mixture*.)
   H_1: $K = K_1 = K_0 + 1$ (Jumlah komponen *mixture* pada model adalah banyaknya komponen *mixture* dalam hipotesis null ditambahkan satu komponen)
2. Menentukan tingkat signifikansi ($\propto$)
3. Menghitung statistik uji

$$P_{value} = \frac{1}{B}\sum_{h=1}^{B} I\left(lrs_1^{(b)} > lrs_0\right),$$

dengan lrs_0 adalah *likelihood ratio statistic* dan B adalah banyaknya proses *bootstrap* yang dilakukan untuk membentuk vektor $lrs_1^{(1)}, \ldots, lrs_1^{(B)}$

4. Menentukan daerah kritis
   H_0 ditolak jika $P_{value} < \propto$
5. Menentukan kesimpulan

2.6. Pemilihan Model

Proses pemilihan model menggunakan metode berbasis kriteria informasi dan dua ukuran yang sering digunakan dalam menentukan banyaknya komponen *mixture* dalam model adalah *Akaike Information Criterion* (AIC) dan *Bayesian Information Criteria* (BIC).

$$AIC = -2lnL\left(\hat{\psi}\right) + 2p$$
$$BIC = -2lnL\left(\hat{\psi}\right) + p\ ln(n)$$

dengan $L\left(\hat{\psi}\right)$ adalah fungsi likelihood dari penduga maksimum likelihood $\hat{\psi}$, p adalah banyaknya parameter dalam model *finite mixture*, dan n adalah banyaknya data observasi.

2.7. Distribusi Normal

Distribusi Normal atau yang disebut juga Distribusi Gauss memiliki variabel random yang kontinu. Fungsi densitas peluang dari distribusi normal yaitu,

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

dengan x adalah variabel random kontinu dan $-\infty \leq x \leq \infty$. Mean μ dan variansi

σ^2 adalah parameter dari Distribusi Normal dengan $-\infty \leq \mu \leq \infty$ dan $\sigma^2 > 0$. $f(x;\mu,\sigma^2)$ dapat juga ditulis dengan $X \sim N(\mu,\sigma^2)$. (Bain and Engelhardt, 1992)

3. METODE PENELITIAN

Penelitian ini merupakan studi kasus data persentase nilai tambah dari PDB pada bidang industri di dunia tahun 2017. Data yang diperoleh merupakan data dari 178 negara, yang merupakan data sekunder dari *The World Bank* pada tahun 2017 dan dapat diakses pada alamat http://worldbank.org.

Estimasi parameter pemodelan *normal mixture* pada data nilai tambah PDB setiap negara di sektor industri tahun 2017 menggunakan estimasi maksimum likelihood dengan algoritma EM. Pengelompokan nilai tambah Produk Domestik Bruto (PDB) masing-masing negara dilakukan dengan menggunakan komponen yang terbentuk dari *normal mixture model* berdasarkan nilai AIC dan BIC terkecil.

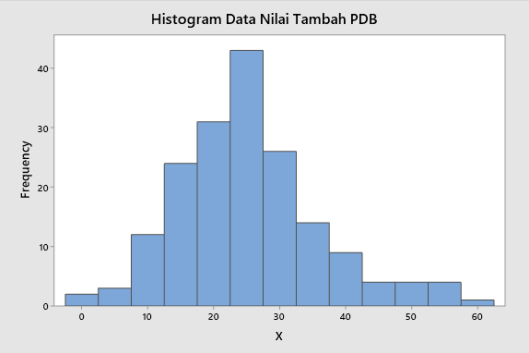
Berikut langkah-langkah yang dilakukan pada penelitian ini untuk pemodelan *normal mixture* pada data Nilai Tambah PDB di setiap negara tahun 2017:

1. Identifikasi pola distribusi univariat multimodal pada data persentase nilai tambah dari Produk Domestik Bruto pada bidang industri tahun 2017. Jika pola multimodal dipenuhi, maka data terindikasi model *finite mixture*.
2. Dilakukan pemilihan distribusi data berdasarkan pola data multimodal menggunakan uji *goodness-of-fit*.
3. Melakukan estimasi model *finite mixture* Normal dengan metode estimasi maksimum likelihood. Estimasi model dilakukan hingga diperoleh nilai estimator yang konvergen.
4. Menggunakan uji *bootstrap likelihood ratio test* untuk melakukan uji signifikansi model *finite mixture*.
5. Menentukan banyaknya komponen *mixture* berdasarkan nilai AIC dan BIC terkecil dan menginterpretasikan hasil pengelompokan tersebut.

4. HASIL DAN PEMBAHASAN

4.1. Identifikasi Pola Data Nilai Tambah Produk Domestik Bruto

Diberikan observasi x_i sebagai nilai tambah dari Produk Domestik Bruto (PDB) pada bidang industri di dunia tahun 2017, kemudian dilakukan identifikasi pola distribusi data untuk melihat adanya pola multimodal. Identifikasi pola distribusi nilai tambah dari PDB pada bidang industri di dunia tahun 2017 dilakukan melalui plot histogram pada Gambar 1. Dan dengan uji *goodness of fit.*



Gambar 1. Histogram Data Nilai Tambah Dari PDB Tahun 2017

Berdasarkan Gambar 1 terlihat bahwa grafik data menunjukkan adanya beberapa puncak distribusi atau bersifat multimodal. Selanjutnya dilakukan uji signifikansi *goodness of fit* untuk melihat terpenuhi atau tidaknya pola multimodal.

1. H_0: Pola data mengikuti distribusi normal unimodal
   H_1: Pola data tidak mengikuti distribusi normal unimodal
2. Menentukan tingkat signifikansi $(\propto)$
   $\propto = 0{,}05$
3. Menghitung statistik uji

Dari hasil output minitab didapatkan nilai A sebesar 1,557 dan P_{value} sebesar 0,005. Selanjutnya dihitung nilai A^* dan c_α:

$$A^* = 1{,}557\left(1 + \frac{0{,}75}{178} + \frac{2{,}25}{178^2}\right) = 1{,}564$$

dan

$$c_\alpha = 0{,}7514\left(1-\frac{0{,}795}{178}-\frac{0{,}89}{178^2}\right) = 0{,}748$$

4. Menentukan daerah kritis

H_0 ditolak jika $A^* > c_\alpha$ dan $P_{value} < 0{,}05$

5. Menentukan kesimpulan

Karena $A^* = 1{,}564 > c_\alpha = 0{,}748$ dan $P_{value} = 0{,}005 < 0{,}05$ maka H_0 ditolak yang berarti pola data tidak mengikuti distribusi normal unimodal.

4.2. Uji Signifikansi Model

Uji signifikasi dengan *bootstrap likelihood ratio statistics* dilakukan untuk melihat sesuai atau tidaknya data jika dimodelkan dengan *finite mixture*. Dalam uji *bootstrap likelihood ratio statistics*, hipotesis null menyatakan bahwa data mengikuti distribusi probabilitas tunggal, sedangkan hipotesis alternatif menyatakan bahwa data mengikuti model *finite mixture*. Berikut pernyataan uji hipotesis dan hasil komputasi dalam pengujian ini:

H_0: $K = 1$ (Model distribusi probabilitas tunggal)

H_1: $K = 2$ (Model *finite mixture* dengan dua komponen *mixture*.)

Tabel 1. Uji Hipotesis Signifikansi Model Finite Mixture

Distribusi Komponen Mixture	P_{value}	$\propto$	Kesimpulan
Normal	0,00	0,05	H_0 ditolak

Pada Tabel 1 diketahui bahwa nilai $p_{value} < \propto$ yang berarti data nilai tambah PDB tahun 2017 lebih sesuai dimodelkan dengan model *finite mixture* dengan komponen *mixture* berdistribusi normal.

4.3. Seleksi Model

Selanjutnya dilakukan proses seleksi model untuk mengetahui banyaknya komponen *mixture* yang sesuai menggunakan ukuran AIC dan BIC. Hasil perhitungan AIC dan BIC untuk setiap banyaknya komponen dituliskan dalam Tabel 2. Banyaknya komponen *mixture* yang dihitung hanya sampai 11 komponen saja karena pada komponen yang berjumlah lebih dari 11 tidak memberikan hasil perhitungan.

Tabel 2. Nilai AIC dan BIC dari Estimasi Model Finite Mixture

Banyak Komponen	AIC	BIC
2	1344,8892	1360,7699
3	1350,7674	1376,1766
4	1354,4481	1389,3857
5	1357,4983	1401,9644
6	1361,2463	1415,2409
7	1363,7034	1427,2264
8	1367,4624	1440,5138
9	1370,0969	1452,6769
10	1368,2701	1460,3785
11	1363,5489	1465,1857

Berdasarkan Tabel 2, nilai AIC dan BIC terendah untuk model *finite mixture* Normal ada pada dua komponen dengan AIC = 1344,8892 dan BIC = 1360,7699. Hasil komputasi estimasi model *finite mixture* diberikan oleh

$$g(x_i|\psi) = \hat{w}_1\,\mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1) + \hat{w}_2\,\mathcal{N}(\hat{\mu}_2, \hat{\sigma}_2) \quad (4)$$

dengan parameter bobot $\hat{w}_1 = 0{,}9208$ dan $\hat{w}_2 = 0{,}0792$ dan parameter-parameter dari distribusi Normal pada komponen pertama $\hat{\imath}_1 = 23{,}3425$ dan $\hat{o}_1 = 8{,}6684$, sedangkan pada komponen kedua $\hat{\imath}_2 = 49{,}5947$ dan $\hat{o}_2 = 6{,}1534$.

Model (4) menyatakan bahwa klaster pertama terdiri atas 92,08% dari populasi atau terdapat 163 negara yang rata-rata persentase nilai tambah PDB bidang industrinya sebesar, sedangkan klaster kedua terdiri atas 7,92% dari populasi atau terdapat 14 negara yang rata-rata persentase nilai tambah PDB bidang industrinya sebesar 49,5947%. Klaster pertama menginterpretasikan negara-negara dengan rata-rata nilai tambah PDB di bidang industrinya hanya sebesar 23,3425%, yang berarti kontribusi sektor industri terhadap PDB negara-negara dalam klaster pertama tidak lebih besar dibandingkan sektor lainnya. Sedangkan klaster kedua yang didominasi negara-negara di timur tengah menginterpretasikan negara-negara dengan rata-rata nilai tambah PDB di bidang industrinya sebesar 49,5947%, yang berarti kontribusi sektor industri

terhadap PDB negara-negara dalam klaster 2 paling besar diantara sektor lainnya.

5. KESIMPULAN

Data nilai tambah dari Produk Domestik Bruto (PDB) pada bidang industri di dunia tahun 2017 memiliki pola dsitribusi multimodal sehingga dapat dimodelkan dengan *finite mixture* normal dua komponen. Kedua komponen menggambarkan dua klaster berdasarkan rata-rata nilai tambah dari Produk Domestik Bruto (PDB) pada bidang industri. Klaster pertama terdapat 163 negara dengan rata-rata persentase nilai tambah PDB bidang industrinya sebesar 23,3425%. Klaster kedua terdapat 14 negara dengan rata-rata persentase nilai tambah PDB bidang industrinya sebesar 49,5947% yang anggotanya didominasi negara-negara di timur tengah.

6. REFERENSI

Bain, L.J. and Engelhardt, M. 1992. *Introduction to Probability and Mathematical Statistics 2 ed*. Duxbury Press, California.

Badan Pusat Statistik. 2019. *Pendapatan Nasional Indonesia tahun 2014-2018*. Jakarta: Badan Pusat Statistik.

Celeux, G. 2018. *EM Methods for Finite Mixtures In*: S. Fruhwirth-Schnatter, G. Celeux dan C.P. Robert, ed. Handbook of Mixture Analysis. CRC Press.

Dempster, A.P., Laird, N.M. dan Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*. Vol.39, No.1, Hal.1-22.

Fruhwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer.

McLachlan, G. J. dan Peel, D. 2000. *Finite mixture models*. New York: Wiley.

Shireman, E., Steinley, D., dan Brusco, M. J. 2015. *Examining the effect of initialization strategies on the performance of Normal mixture modeling*. USA: Psychonomic Society, Inc.

Susanto, I. dan Handajani, S.S. 2019. Artikel. Pemodelan Distribusi Pendapatan Rumah Tangga Per Kapita di Indonesia dengan Model *Finite Mixture*.

Thrun, M. C. 2019. Cluster Analysis of Per Capita Gross Domestic Products. *Entrepreneurial Business and Economics Review*. Vol. 7, No. 1.

Yu, Y. 2018. *mixR: Finite Mixture Modelling for Raw and Binned Data* [Daring]. R package version 0.1.1. Available from: https://cran.r-project.org/package=mixR.q

World Bank, (2012), Gross Domestic Product Deflator, diakses dari http://worldbank.org pada tanggal 11 November 2019.