J. Ris. & Ap. Mat. Vol. 07 No. 02 (2023) pp. 116-127
Jurnal Riset dan Aplikasi Matematika
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

PEMODELAN JUMLAH KASUS BARU HARIAN COVID-19 DI INDONESIA MENGGUNAKAN *GAUSSIAN MIXTURE MODEL*

Fevi Novkaniza [1*], Nico, dan Rahmat Al Kafi [3]

[1,3]Departemen Matematika, FMIPA Universitas Indonesia, Depok 16424, Indonesia

*Corresponding author: fevi.novkaniza@sci.ui.ac.id

ABSTRAK

Penyakit COVID-19 adalah penyakit menular yang disebabkan oleh virus *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2). Selama masa pandemi COVID-19 terjadi beberapa kali lonjakan jumlah kasus baru COVID-19 yang menunjukkan adanya kesulitan dalam mengantisipasi peningkatan penyebaran COVID-19. Artikel ini membahas pemodelan jumlah kasus baru harian COVID-19 di Indonesia dari 1 Januari 2021 sampai 31 Maret 2022 menggunakan *Gaussian Mixture Model* (GMM). GMM merupakan salah satu *mixture model* dimana setiap komponen campuran diasumsikan berdistribusikan Gaussian. GMM dikonstruksi mengggunakan beberapa komponen campuran dan parameter dari setiap GMM diestimasi menggunakan metode *maximum likelihood estimation* (MLE) melalui algoritma *Expectation-Maximization* (EM). Berdasarkan nilai *Akaike Information Criteria* (AIC), diperoleh GMM dengan 4 komponen merupakan model terbaik untuk pemodelan data jumlah kasus baru harian COVID-19 di Indonesia. Berdasarkan menggunakan model GMM terbaik, diperoleh probabilitas jumlah kasus baru harian COVID-19 di Indonesia kurang dari jumlah kasus harian terendah adalah 0,01, lebih dari jumlah kasus harian rata-rata adalah 0,3 dan lebih dari jumlah kasus harian tertinggi adalah 0,017.
Kata kunci: COVID-19, *Expectation-Maximization*, Gaussian, komponen campuran

ABSTRACT

COVID-1 is an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). During the COVID-19 pandemic, there were several spikes in the number of new COVID-19 cases, which shows that there are difficulties in anticipating the increase in the spread of COVID-19. This article discusses modelling the number of daily new cases of COVID-19 in Indonesia from 1st January 2021 to March 31, 2022, using the Gaussian Mixture Model (GMM). GMM is a mixture model where each mixture component is assumed to have a Gaussian distribution. GMM is constructed using several mixed components, and the parameters of each GMM are estimated using the maximum likelihood estimation (MLE) method via the Expectation-Maximization (EM) algorithm. Based on the Akaike Information Criteria (AIC) values, it was found that GMM with 4 components is the best model for modelling data on the number of daily new cases of COVID-19 in Indonesia. Based on the best GMM model, the probability that the number of new daily COVID-19 cases in Indonesia is less than the lowest number of daily cases is 0.01, more than the average number of daily cases is 0.3, and more than the highest number of daily cases is 0.017.
***Keywords**: COVID-19, Expectation-Maximization, Gaussian, mixture component*

1 Pendahuluan

Coronavirus Disease 2019 (COVID-19) adalah penyakit menular yang disebabkan oleh virus *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) dengan gejala umum

2020 *Mathematics Subject Classification*: 62H30
Tanggal masuk: 21-09-23; direvisi: 09-10-23; diterima: 11-10-23

berupa demam, batuk, sakit kepala, kelelahan, gangguan pernapasan, hingga anosmia. Virus SARS-CoV-2 dapat menyebar dari mulut maupun hidung dari orang yang terinfeksi melalui partikel cairan kecil atau aerosol yang dikeluarkan ketika orang tersebut batuk, bersin, berbicara hingga bernapas. Menurut *Center for Disease Control and Prevention* (CDC), kasus pertama COVID-19 ditemukan di Wuhan, China pada 31 Desember 2019 dan secara resmi dinyatakan sebagai pandemi oleh WHO pada 11 Maret 2020. Di Indonesia, kasus COVID-19 pertama kali terdeteksi pada 2 Maret 2020, yaitu 2 orang warga yang berdomisili di Depok diketahui mengidap virus SARS Cov-2. Kedua pengidap COVID-19 tersebut memiliki riwayat berinteraksi dengan warga negara Jepang yang diketahui lebih dulu menderita penyakit tersebut. Penyebaran COVID-19 di Indonesia juga cukup cepat hingga memicu pemerintah DKI Jakarta untuk mulai memberlakukan Pembatasan Sosisal Berskala Besar (PSBB) pada 10 April 2020. Menyusul pemberlakuan PSBB di DKI Jakarta, Presiden Jokowi menetapkan pandemi koronavirus di Indonesia sebagai bencana nasional pada 13 April 2020 dan mulai memberlakukan PSBB diluar DKI Jakarta.

Berdasarkan data dari WHO, tingkat penyebaran COVID-19 di dunia mencapai titik yang sangat rendah yaitu dibawah 600.000 kasus harian pada Oktober dan November 2021. Namun, pada bulan November 2021, Profesor Tulio de Oliveira menemukan varian baru COVID-19 yaitu varian Omicron yang kemudian menyebabkan peningkatan yang signifikan pada kasus harian COVID-19 hingga mencapai titik tertingginya yaitu 4.042.762 kasus harian pada Januari 2022. Hal ini juga terjadi di Indonesia dimana kasus baru harian COVID-19 sangat rendah selama Oktober hingga Desember 2021 bahkan pemerintah melakukan pelonggaran terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Namun setelah mulai masuknya COVID-19 varian omicron, tingkat penyebaran COVID-19 di Indonesia meningkat hingga mencapai tingkat tertingginya yaitu 64.718 kasus baru pada 16 Februari 2022. Dengan ini dapat disimpulkan bahwa jumlah kasus baru COVID-19 yang rendah tidak menjamin bahwa pelonjakan kasus baru COVID-19 tidak akan kembali terjadi.

Pertanggal 30 Desember 2022, di Indonesia sudah terdapat total 6.719.327 kasus terkonfirmasi dengan 552 kasus baru harian dan total 160.593 kasus kematian dengan 10 kasus kematian baru harian. Jumlah kasus baru harian di Indonesia sudah sangat rendah jika dibandingkan dengan jumlah kasus baru harian pada saat puncak penyebaran COVID-19 varian omicron. Karenanya, Presiden Joko Widodo secara resmi mencabut Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di seluruh wilayah Indonesia. Namun, berkaca pada sejarah penyebaran COVID-19 varian omicron dimana jumlah kasus baru melonjak setelah pelonggaran PPKM yang menunjukkan bahwa adanya kesulitan dalam mengantisipasi lonjakan kasus baru COVID-19. Untuk itu diperlukan pemodelan jumlah kasus baru harian COVID-19 di Indonesia sehingga dapat digunakan dalam prediksi terjadinya kemunculan kasus baru harian COVID-19 yang tinggi di Indonesia.

Jumlah kasus baru harian COVID-19 adalah data numerik yang dapat dimodelkan menggunakan distribusi probabilitas. Salah satu distribusi probabilitas yang sering digunakan adalah distribusi Normal atau Gaussian. Distribusi Gaussian pertama kali ditemukan oleh Carl Friedrich Gauss pada 1809. Distribusi gaussian adalah distribusi probabilitas berbentuk kurva lonceng atau *bell-shape* yang simetris terhadap rata-ratanya [1]. Distribusi normal adalah salah satu distribusi kontinu yang paling penting dan paling banyak digunakan diantara semua distribusi probabilitas. Jumlah kasus baru harian COVID-19 yang dilaporkan merupakan variabel acak diskrit, sedangkan pada kenyataannya bisa saja terjadi kasus baru harian yang jumlahnya jauh lebih banyak namun tidak dilaporkan. Akibatnya, distribusi diskrit kurang cocok digunakan untuk memodelkan data jumlah kasus baru harian COVID-19 karena data tersebut bisa memiliki ukuran yang sangat besar. Oleh karena itu, dapat diasumsikan bahwa data jumlah kasus baru harian COVID-19 mengikuti distribusi Gaussian.

Namun, pada kenyataannya bisa saja terdapat keadaan dimana suatu data memiliki lebih dari satu karakteristik sehingga tidak dapat dimodelkan dengan hanya satu distribusi saja. Sehingga diperlukan suatu model yang terdiri dari campuran beberapa distribusi yaitu model

campuran (*mixture model)*. Model ini pertama kali digunakan oleh Karl Pearson pada 1894 yang menggunakan campuran dari dua distribusi Gaussian untuk memodelkan data kepiting yang dicurigai memiliki 2 subpopulasi. *Mixture Model* adalah suatu model yang umumnya digunakan untuk pemodelan data dengan populasi heterogen [2]. Secara matematis, *Mixture Model* merupakan penjumlahan linear berbobot dari beberapa fungsi distribusi yang berada pada dimensi yang sama. Beberapa contoh penggunaan *Mixture Model* adalah pemodelan ombak laut menggunakan campuran dari distribusi lognormal bivariat dan pemodelan proses simulasi *Distribution Activation Energy Model* (DAEM) dan proses pirolisis *Distillers Dried Grains with Solubles* (DDGS) yang menggunakan model campuran dari beberapa distribusi turunan Weibull [3], [4].

Salah satu model campuran yang paling banyak digunakan adalah *Gaussian Mixture Model* (GMM). GMM adalah model probabilitas yang menggunakan asumsi bahwa data mengikuti karakteristik dari beberapa distribusi Gaussian dengan parameter yang tidak diketahui [5]. GMM telah banyak digunakan dalam berbagai penelitian, seperti penelitian [6] yang menggunakan GMM untuk mengkonstruksi model *Value-at-Risk* (VaR) dan *Expected Shortfall* (ES). Namun permasalahan dalam penggunaan GMM adalah parameternya yang cukup banyak sehingga menyebabkan parameter GMM tidak dapat diestimasi secara langsung melalui metode *maximum likelihood estimation* (MLE). Namun hal ini dapat diatasi dengan menggunakan algoritma *Expectation-Maximization* (EM).

Algoritma EM pertama kali dipopulerkan oleh [7]. Menurut [8], algoritma EM merupakan salah satu metode komputasi yang telah banyak diterapkan pada penelitian-penelitian di bidang *machine learning*, kedokteran, ekonomi, hingga sosiologi dan sebagainya. Secara garis besar, algoritma EM merupakan iterasi pada dua tahap, yatiu ekspektasi (*Expectation)* dan maksimalisasi (*Maximization)*. Langkah ekspektasi merupakan estimasi nilai ekspektasi variabel laten lalu membentuk persamaan *log-likelihood* dari data lengkap. Selanjutnya, tahap maksimalisasi, merupakan langkah mencari estimasi parameter-parameter model yang memaksimalkan persamaan *log-likelihood* data lengkap yang diperoleh dari langkah ekspektasi. Pada penelitian ini, GMM digunakan untuk memodelkan data jumlah kasus baru harian COVID-19 harian di Indonesia dari 1 Januari 2021 sampai 31 Maret 2022 dengan interval waktu 455 hari yang diperoleh dari website *https://covid19.who.int/data*.

2 Tinjauan Pustaka

Pada GMM, parameter tidak dapat diestimasi secara langsung menggunakan metode MLE. Oleh karena itu, perlu digunakan suatu metode iteratif yaitu algoritma EM. Untuk menggunakan algoritma EM, GMM harus dipandang sebagai model dengan variabel laten yaitu dengan mengasumsikan terdapat suatu variabel laten $Z_i \in \{1,..,K\}$ yang merepresentasikan komponen campuran untuk X_i dimana $Z_i = k$ menunjukkan bahwa $X_i = x$ berasal dari komponen ke-k. Lalu definisikan variabel indikator $I(Z_i = k) = \begin{cases} 0, & Z_i \neq k \\ 1, & Z_i = k \end{cases}$, $k = 1,..K$, dimana $p(Z_i) = \prod_{k=1}^{K} \pi_k^{I(Z_i=k)}$. Lalu misalkan observasi X_i berasal dari GMM dengan K komponen campuran dan probabilitas bersyarat $X_i = x$ diketahui $Z_i = k$ adalah

$$p(X_i = x|Z_i = k) = N(x|\mu_k, \sigma_k^2),$$

maka diperoleh fungsi likelihood *complete data* (X, Z) dari GMM sebagai berikut

$$L(X, Z|\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} N(x_i|\mu_k, \sigma_k^2)^{I(Z_i=k)} {\pi_k}^{I(Z_i=k)} \qquad (1)$$

dan *log-likelihood* sebagai berikut

$$l(X, Z|\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \log \pi_k N(x_i|\mu_k, \sigma_k^2) \qquad (2)$$

Selanjutnya, dilakukan iterasi antara langkah ekspektasi (*expectation)* yaitu mencari fungsi ekspektasi *log-likelihood* dari *complete data* dan langkah maksimalisasi (*maximization*)

yaitu memaksimumkan fungsi yang diperoleh dari langkah ekspektasi. Secara lengkap, iterasi ke-(t) dari algoritma EM berupa kedua tahap berikut

1. Langkah *Expectation*

   Mengkonstruksi ekspektasi *log-likelihood* dari *complete data* (X, Z) yang akan dinotasikan sebagai Q yaitu

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right) = E\left[l(X, Z|\boldsymbol{\theta})\middle|X, \boldsymbol{\theta}^{(t-1)}\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} p\left(Z_i = k\middle|x_i, \boldsymbol{\theta}^{(t-1)}\right)\log \pi_k N(x_i|\mu_k, \sigma_k^2)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t-1)}\log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t-1)}\log N(x_i|\mu_k, \sigma_k^2),$$

   dimana $r_{ik}^{(t-1)}$ adalah distribusi posterior dari Z dengan nilai estimasi parameter yang diperoleh dari iterasi ke-$(t-1)$ yaitu

$$r_{ik}^{(t-1)} = p\left(Z_i = k\middle|x_i, \boldsymbol{\theta}^{(t-1)}\right) = \frac{p\left(x_i\middle|Z_i = k, \boldsymbol{\theta}^{(t-1)}\right)p\left(Z_i = k|\boldsymbol{\theta}^{(t-1)}\right)}{p(x_i|\boldsymbol{\theta}^{(t-1)})}$$

$$= \frac{\pi_k^{(t-1)} N\left(x_i\middle|\mu_k^{(t-1)}, {\sigma_k^2}^{(t-1)}\right)}{\sum_{k=1}^{K}\pi_k^{(t-1)} N\left(x_i\middle|\mu_k^{(t-1)}, {\sigma_k^2}^{(t-1)}\right)},$$

   dan $N_k^{(t)} = \sum_{i=1}^{n} r_{ik}^{(t)}$

2. Langkah *Maximization*

   Pada langkah ini dicari estimasi nilai parameter $\mu_k, \sigma_k, \pi_k,\ k = 1, \dots, K$ yang memaksimumkan fungsi Q yang diperoleh pada langkah *expectation*

   - Estimasi dari parameter mean dapat diperoleh melalui persamaan berikut:

$$\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right)}{\partial \mu_k} = 0$$

$$\Leftrightarrow \widehat{\mu_k} = \frac{\sum_{i=1}^{n} r_{ik}^{(t-1)} x_i}{N_k^{(t-1)}}, \qquad k = 1, \dots, K$$

   - Estimasi dari parameter variansi dapat diperoleh melalui persamaan berikut

$$\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right)}{\partial \sigma_k} = 0$$

$$\Leftrightarrow \sigma_k^2 = \frac{\sum_{i=1}^{n} r_{ik}^{(t-1)}(x_i - \mu_k)^2}{N_k^{(t-1)}}$$

   - Estimasi dari parameter bobot campuran dapat diperoleh menggunakan metode pengali lagrange. Perhatikan bahwa bobot campuran memiliki syarat sebagai berikut

$$\sum_k \pi_k = 1 \Leftrightarrow \sum_k \pi_k - 1 = 0$$

Dengan menggunakan ekspektasi *log-likelihood* Q, dapat diperoleh fungsi lagrange sebagai berikut

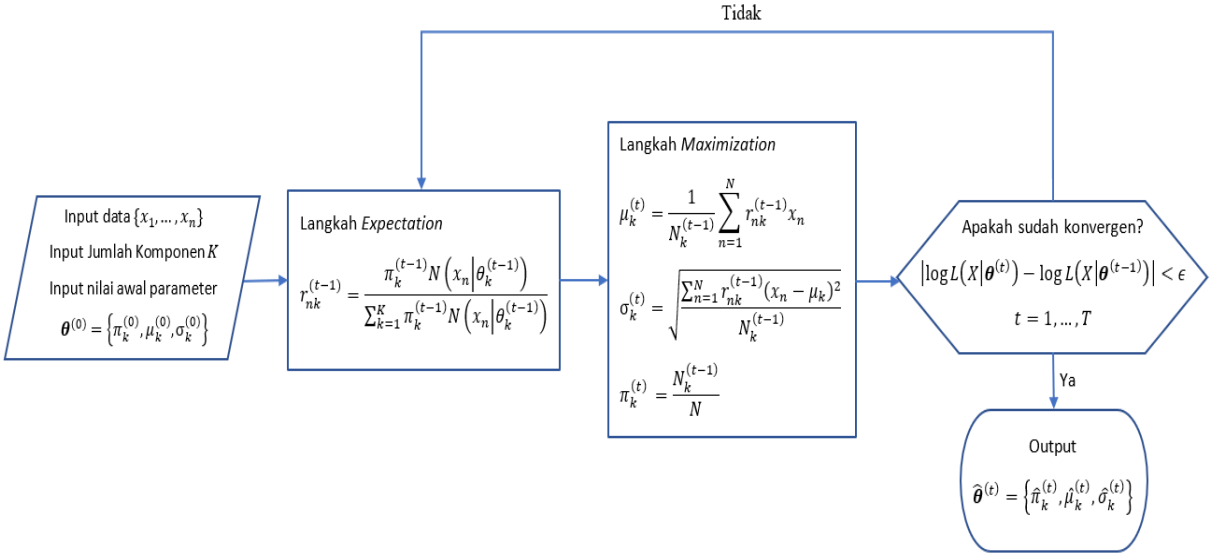
$$\mathcal{L} = Q + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t-1)}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t-1)}\log N(x_i|\mu_k,\sigma_k^2) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

Kemudian, dengan menggunakan metode pengali lagrange diperoleh estimasi parameter bobot campuran sebagai berikut

$$\widehat{\pi_k} = \frac{\sum_{i=1}^{n} r_{ik}^{(t-1)}}{n} = \frac{N_k^{(t-1)}}{n} \qquad (3)$$

Dalam penggunaan algoritma EM, perlu dilakukan inisialisasi nilai awal parameter dari GMM yang dinotasikan dengan $\boldsymbol{\theta}^{(0)} = \{\mu_k^{(0)}, \sigma_k^{(0)}, \pi_k^{(0)},\ k = 1, \dots, K\}$ yang kemudian digunakan pada langkah *expectation* pada iterasi pertama. Prosedur estimasi parameter GMM menggunakan algoritma EM dapat digambarkan dalam diagram alir sebagai berikut:



Gambar 1. Diagram Alur Algoritma EM untuk GMM

Berdasarkan nilai *Akaike Information Criterion* (AIC) [9] terkecil dengan formulasi:

$$AIC = -2\log L(X|\boldsymbol{\theta}) + 2k \qquad (4)$$

dimana $L(X|\boldsymbol{\theta})$ merupakan fungsi likelihood dari GMM, dan k merupakan jumlah parameter yang diestimasi, dapat ditentukan model terbaik berdasarkan nilai AIC yang terkecil. Model GMM terbaik selanjutnya digunakan untuk estimasi probabilitas menggunakan persamaan berikut:

$$P(X \le m) = \int_0^m \sum_{k=1}^{K}\pi_k N(x|\mu_k,\sigma_k^2) \qquad (5)$$

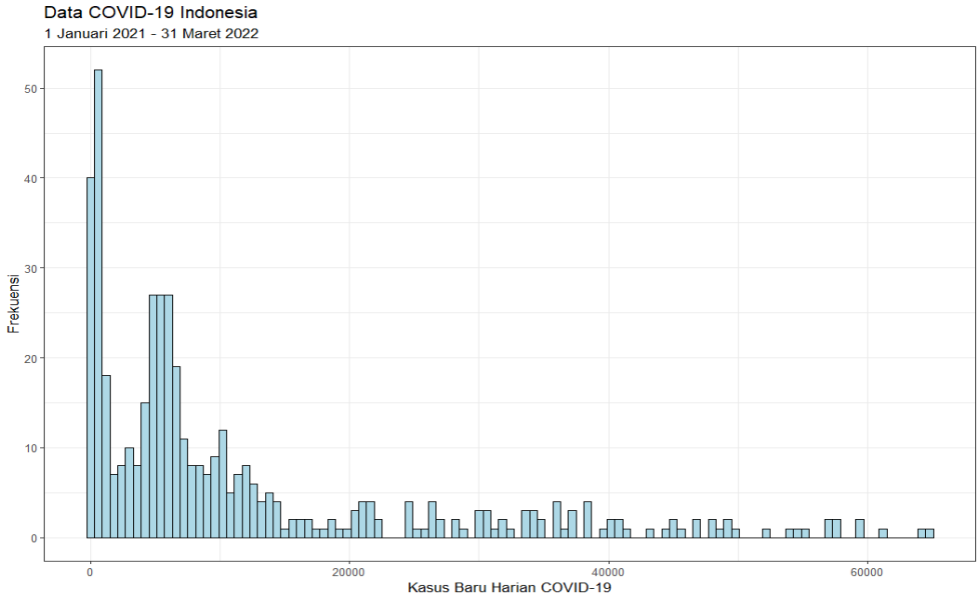
3 Hasil dan Pembahasan

Berdasarkan data jumlah kasus baru harian COVID-19 harian di Indonesia dari 1 Januari 2021 sampai 31 Maret 2022 dengan interval waktu 455 hari yang diperoleh dari website

https://covid19.who.int/data, statistik deskriptif dan histogram disajikan pada Tabel 1 dan Gambar 2 berikut ini:

Tabel 1. Statistik Deskriptif Data

Jumlah Observasi	455
Minimum	92
Median	5944
Rata-rata	11582
Maksimum	64718



Gambar 2. Histogram Data Jumlah Kasus Baru Harian COVID-19

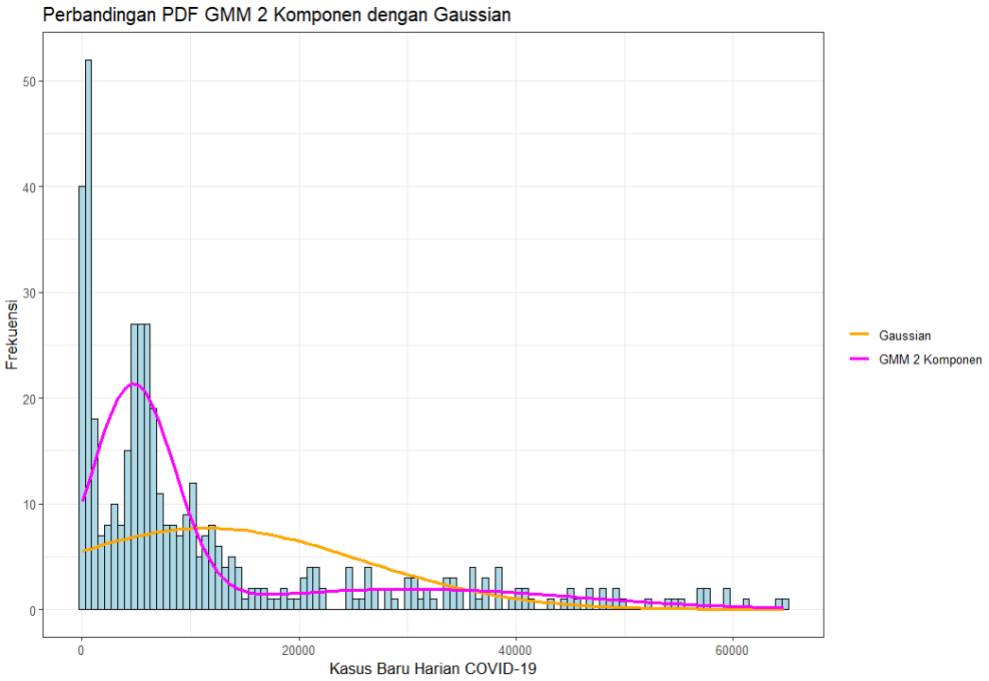
Dari Gambar 2, terlihat bahwa data jumlah kasus baru harian COVID-19 di Indonesia memiliki 3 gelombang berbeda yang masing-masing cukup simetris sehingga tersebut dapat dimodelkan dengan GMM dengan 3 komponen campuran. Sebagai pembanding juga dikonstruksi 3 model GMM dengan 2 komponen, 3 komponen dan 4 komponen. Estimasi parameter dan fungsi densitas masing-masing model dapat diperoleh menggunakan algoritma EM yang dijalankan menggunakan perangkat lunak R.

3.1 GMM dengan 2 Komponen Campuran

Dengan mengasumsikan bahwa data jumlah kasus baru harian COVID-19 terdiri atas 2 karakteristik sub-data yang berbeda, maka dapat dikontruksi GMM dengan 2 komponen campuran. Hasil estimasi pdf dari GMM dengan 2 komponen campuran adalah sebagai berikut:

$$\hat{p}(x) = 0{,}2684453\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-30277{,}48}{15158{,}21}\right)^2\right)}{15158{,}21\sqrt{2\pi}}\right) + 0{,}7315547\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-4719{,}283}{3809{,}355}\right)^2\right)}{3809{,}355\sqrt{2\pi}}\right).$$

Grafik estimasi pdf dari GMM 2 komponen campuran terdapat pada Gambar 3 berikut ini:



Gambar 3. Plot Perbandingan Estimasi pdf GMM 2 Komponen dengan pdf Gaussian

Dari Gambar 3, terlihat bahwa estimasi pdf GMM dengan 2 komponen lebih sesuai dengan data dibanding pdf Gaussian. Namun, puncak tertinggi dari data belum mampu diakomodir dengan hanya 1 distribusi Gaussian maupun GMM dengan 2 komponen campuran.

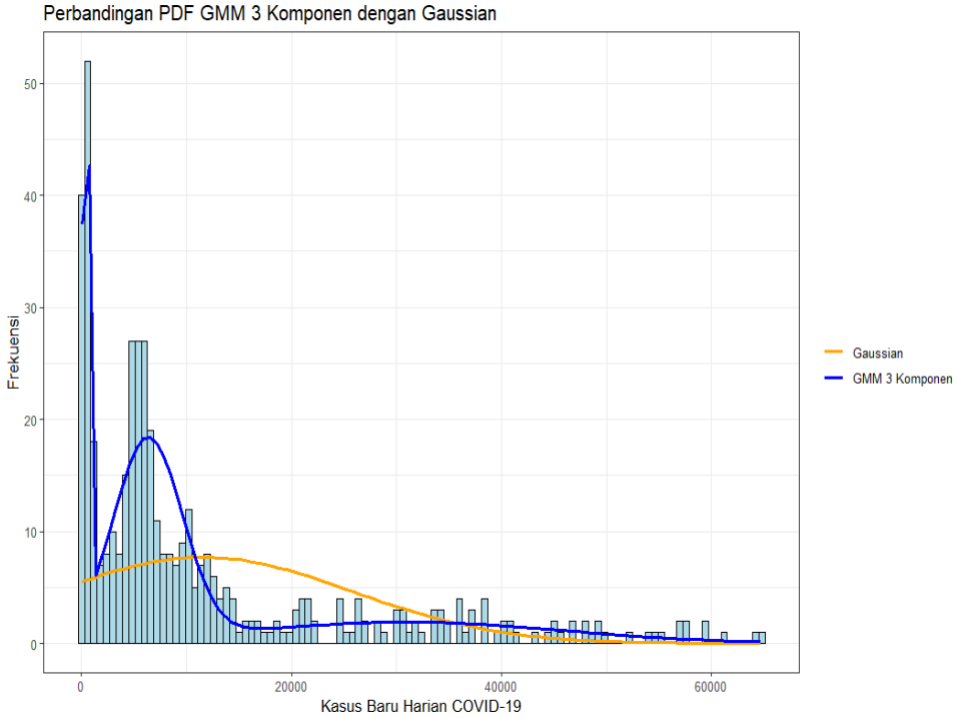
3.2 GMM 3 Komponen Campuran

Dengan mengasumsikan bahwa data jumlah kasus baru harian COVID-19 terdiri atas 3 karakteristik sub-data yang berbeda, maka dapat dikontruksi GMM dengan 3 komponen campuran. Hasil estimasi pdf dari GMM dengan 2 komponen campuran adalah sebagai berikut:

$$\hat{p}(x) = 0{,}2605101\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-30986{,}07}{14795{,}18}\right)^2\right)}{14795{,}18\sqrt{2\pi}}\right) + 0{,}5405994\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-6335{,}185}{3280{,}602}\right)^2\right)}{3280{,}602\sqrt{2\pi}}\right)$$

$$+ 0{,}1988905\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-424{,}8165}{237{,}9972}\right)^2\right)}{237{,}9972\sqrt{2\pi}}\right).$$

Selanjutnya grafik estimasi pdf dari GMM 3 komponen campuran terdapat pada Gambar 4 berikut ini:



Gambar 4. Plot Perbandingan Estimasi pdf GMM 3 Komponen dengan pdf Gaussian

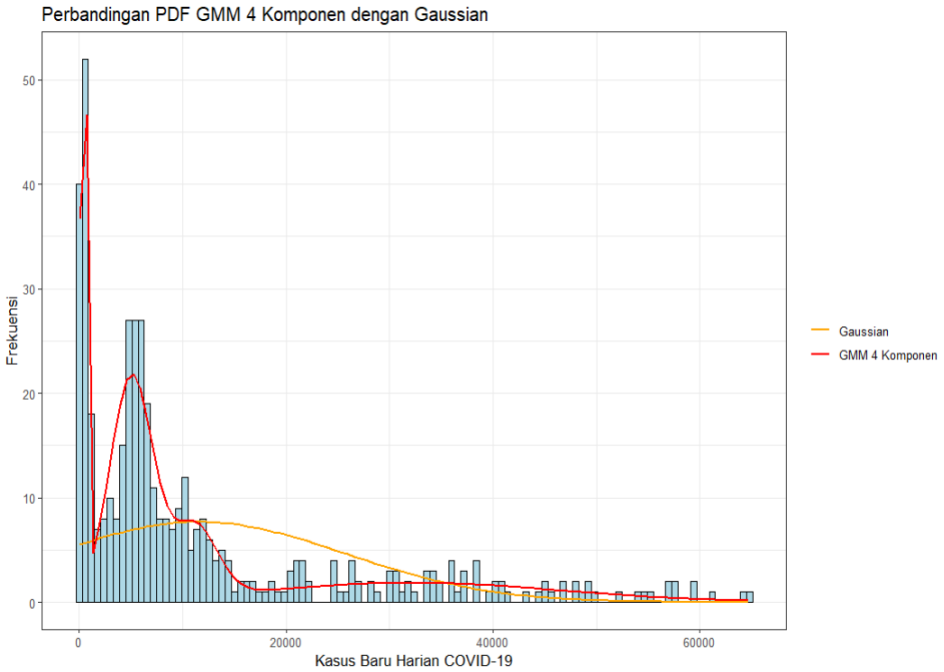
Pada Gambar 4, terlihat bahwa seperti GMM 3 komponen, model ini lebih sesuai dengan karakteristik data dan puncak tertinggi dari data juga sudah bisa diakomodir oleh model.

3.3 GMM 4 Komponen

Dengan mengasumsikan bahwa data jumlah kasus baru harian COVID-19 terdiri atas 4 karakteristik sub-data yang berbeda, maka dapat dikontruksi GMM dengan 4 komponen campuran. Hasil estimasi pdf dari GMM dengan 4 komponen campuran adalah sebagai berikut:

$$\hat{p}(x) = 0{,}1356412\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-11175{,}604}{2167{,}584}\right)^2\right)}{2167{,}584\sqrt{2\pi}}\right) + 0{,}2086614\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-436{,}7362}{250{,}9271}\right)^2\right)}{250{,}9271\sqrt{2\pi}}\right)$$
$$+ 0{,}2453927\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-32157{,}68}{14407{,}62}\right)^2\right)}{14407{,}62\sqrt{2\pi}}\right) + 0{,}4103048\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-5077{,}432}{2086{,}84}\right)^2\right)}{2086{,}84\sqrt{2\pi}}\right).$$

Selanjutnya, dapat diperoleh grafik estimasi pdf dari GMM 4 komponen yang ditunjukkan pada Gambar 5 berikut ini:



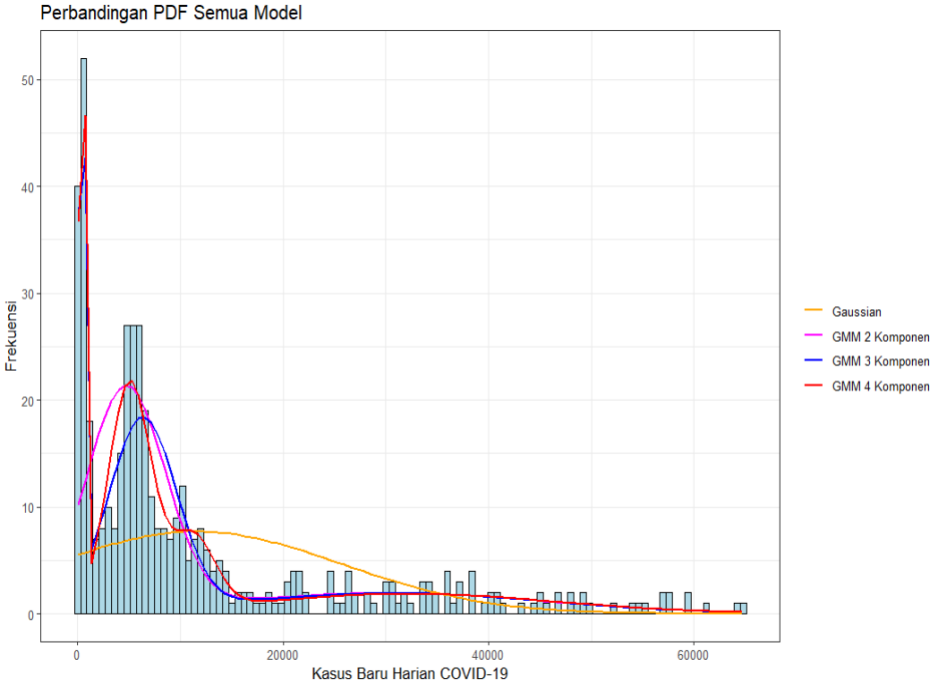
Gambar 5. Plot Perbandingan Estimasi pdf GMM 3 Komponen dengan pdf Gaussian

Pada Gambar 5, terlihat bahwa seperti GMM 3 komponen, model ini juga sesuai dengan karakteristik data dan puncak tertinggi dari data juga sudah bisa diakomodir oleh model. Namun perlu ditentukan model GMM terbaik berdasarkan kriteria nilai AIC yang disajikan pada Tabel 2 sebagai berikut:

Tabel 2. Perbandingan Nilai AIC

Model	AIC
Distribusi Gaussian	9992,35
GMM 2 Komponen	9535,45
GMM 3 Komponen	9303,03
GMM 4 Komponen	9281,98

Berdasarkan perbandingan nilai AIC, dapat disimpulkan bahwa GMM dengan 4 komponen merupakan model terbaik. Hal ini juga terlihat pada grafik perbandingan estimasi pdf semua model yang ditunjukkan pada Gambar 6 sebagai berikut:



Gambar 6. Plot Perbandingan Estimasi pdf Semua Model

Pada Gambar 6, terlihat bahwa garis merah yaitu estimasi pdf GMM dengan 4 komponen memiliki karakteristik yang paling menyerupai karakteristik data dan ini sesuai dengan hasil perbandingan nilai AIC. Berdasarkan model GMM dengan 4 komponen campuran, selanjutnya dilakukan estimasi probabilitas jumlah kasus baru harian COVID-19 di Indonesia berdasarkan persamaan (5) yaitu:

$$\hat{P}(X \le m) = \int_0^m 0{,}356412\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-11175{,}604}{2167{,}584}\right)^2\right)}{2167{,}584\sqrt{2\pi}}\right) + 0{,}2086614\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-436{,}7362}{250{,}9271}\right)^2\right)}{250{,}9271\sqrt{2\pi}}\right)$$
$$+ 0{,}2453927\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-32157{,}68}{14407{,}62}\right)^2\right)}{14407{,}62\sqrt{2\pi}}\right) + 0{,}4103048\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{x-5077{,}432}{2086{,}84}\right)^2\right)}{2086{,}84\sqrt{2\pi}}\right)dx$$

Tabel 3 memuat estimasi probabilitas untuk beberapa nilai m berdasarkan statistik deskriptif data yaitu:

Tabel 3. Estimasi Probabilitas Terjadinya Sejumlah Kasus Baru Harian COVID-19 di Indonesia

m	$\hat{P}(X \le m)$
92	0,009598
11.582	0,299443
64.718	0,017669

Berdasarkan Tabel 3, terlihat bahwa peluang jumlah kasus baru harian COVID-19 di Indonesia mencapai tingkat terendah hanya sebesar 0,009598 yaitu sangat kecil. Namun, dapat dilihat juga bahwa peluang untuk jumlah kasus melebihi tingkat tertinggi juga hanya sebesar 0,017669 yang dapat dikatakan juga sangat kecil. Sedangkan, peluang jumlah kasus melebihi rata-rata adalah sebesar 0,299443 yang artinya lebih besar kemungkinan bahwa kasus baru COVID-19 di Indonesia akan berjumlah kurang dari rata-rata.

4 Kesimpulan dan Saran

Dalam pemodelan jumlah kasus baru harian COVID-19 di Indonesia dengan menggunakan data dari 1 Januari 2021 hingga 31 Maret 2022, diperoleh GMM dengan 4 komponen sebagai model terbaik. Berdasarkan model tersebut, diperoleh hasil sebagai berikut

a. Estimasi probabilitas jumlah kasus baru harian COVID-19 di Indonesia kurang dari jumlah kasus harian terendah adalah 0,009598,
b. Estimasi probabilitas jumlah kasus baru harian COVID-19 di Indonesia lebih banyak dari jumlah kasus harian rata-rata adalah 0,299443
c. Estimasi probabilitas jumlah kasus baru harian COVID-19 di Indonesia lebih banyak dari jumlah kasus harian tertinggi adalah 0,017669.

Walaupun probabilitas jumlah kasus baru harian COVID-19 di Indonesia mencapai tingkat tertinggi cukup rendah, tindakan pencegahan penyebaran COVID-19 tetap harus dipertahankan karena probabilitas jumlah kasus baru harian COVID-19 di Indonesia mencapai tingkat terendah juga sangat rendah sehingga mengindikasikan bahwa pandemi COVID-19 masih belum berakhir dan tetap diperlukan pemantauan terhadap kondisi kesehatan masyarakat di Indonesia. Selain pemodelan penyakit COVID-19, GMM dapat digunakan untuk pemodelan penyakit lainnya seperti penyakit Alzheimer [10], Parkinson [11] dan penyakit lainnya. Hal ini karena model campuran tidak memerlukan informasi di subpopulasi mana suatu titik data

References

[1] P. S. Mann, *Introductory Statistics* (7th ed.), Wiley, 2010.

[2] J. Shen, *Finite Mixture Regression Models And Applications: Detection Limit And Goodness-Of-Fit Test*, University of Medicine and Dentistry of New Jersey, 2011.

[3] W. Huang and S. Dong, "Join Distribution of Individual Wave Heights and Periods in Mixed Sea States Using Finite Mixture Models", *Coastal Engineering*, 161, 2020. https://doi.org/10.1016/j.coastaleng.2020.103773

[4] L. Tian, R. Li, and P. Ma, Insight Into Derivative Weibull Mixture Model in Describing Simulated Distributed Activation Energy Model and Distillers Dried Grains With Solubles Pyrolysis Processes. *Waste Management*, 153, 219-228. https://doi.org/10.1016/j.wasman.2022.09.010

[5] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[6] S. M. Seyfi, A. Sharifi, and H. Arian, "Portfolio Value-at-Risk and Expected-Shortfall Using An Efficient Simulation Approach Based On Gaussian Mixture Model", *Mathematics and Computers in Simulation*, 190, pp. 1056-1079, 2021. https://doi.org/10.1016/j.matcom.2021.05.029

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B*(Methodological), Vol. 39, 1-38, 1977. http://dx.doi.org/10.2307/2984875

[8] T. K. Moon, "The Expectation-Maximization Algorithm", *Signal Processing*, 6, 47-60, 1996. https://doi.org/10.1109/79.543975

[9] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, AC- 19, 716-723, 1974. http://dx.doi.org/10.1109/TAC.1974.1100705.

[10] R. Li, R. Perneczky, I. Yakushev, S. Förster, A. Kurz, et al, "Gaussian Mixture Models and Model Selection for [18F] Fluorodeoxyglucose Positron Emission Tomography Classification in Alzheimer's Disease", *PLOS ONE* 10(4): e0122731, 2015.

https://doi.org/10.1371/journal.pone.0122731

[11] J. Laeticia, M. Graziella, C. Jean-Christophe, V. Marie, L. Stephane, et.al, "Comparison of telephone recordings and professional microphone recordings for early detection of Parkinson's disease, using mel-frequency cepstral coefficients with Gaussian mixture models". *INTERSPEECH 2019: 20th annual conference of the International Speech Communication Association*, Sept 2019, Graz, Austria, pp.3033-3037, 2020.