# Estimasi Parameter Distribusi Campuran BiWeibull

A. S. Awalluddin[1, a)]

[1]*Jurusan Matematika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung*

[a)]*aasolih@uinsgd.ac.id*

**Abstrak**

Salah satu pendekatan distribusi yang banyak digunakan dalam analsis data statistik adalah Distribusi Weibull. Pendekatan distribusi Weibull tunggal untuk populasi yang didalamnya terdapat beberapa sub populasi kurang baik digunakan, karena adanya perbedaan karakteristik data antar sub populasi tersebut, sehingga memungkinkan memiliki parameter yang berbeda. Distribusi campuran (mixture distribution) diharapkan dapat mencadi solusi pendekatan bentuk distribusi dengan kondisi data berasal dari beberapa sub populasi. Sebagai contoh distribusi usia atau masa pengobatan (sampai mati) pasien kanker atau Aids, dengan sub populasi adalah ras atau kelompok usia. Tulisan ini membahas distribusi campuran Weibull, khususnya model distribusi campuran biWeibull yang akan digunakan dalam studi kasus data yang terdiri dari dua sub populasi. Model biWeibull tidak lain adalah kombinasi linier dari dua distribusi Weibull (biWeibull) dengan bentuk berikut :

$$g\left(x,\overline{\alpha},\overline{\beta}\right)=\pi\, Weib\left(x,\alpha_1,\beta_1\right)+\left(1-\pi\right)Weib\left(x,\alpha_2,\beta_2\right)$$

dimana g(.) adalah distribusi campurannya, selanjutnya $\pi$ , $\overline{\alpha}$ , dan $\overline{\beta}$ masing-masing dikenal sebagai parameter proporsi, parameter vektor skala dan vektor bentuk (*shape*) yang berukuran dua. Metode estimasi parameter model yang digunakan Tahapan estimasi parameter model diuraikan dalam tulisan ini. Di lapangan, masalah yang timbul selanjutnya adalah bagaimana menaksir ketiga parameter model tersebut jika data ada di tangan. Dalam tulisan ini algoritma yang digunakan untuk menaksir parameter di atas adalah algoritma EM (*Expectation Maximization*). Studi kasus sebagai penerapan algoritma dilakukan untuk kasus pendekatan distribusi campuran biWeibull.

*Keywords: Distribusi Weibull, fungsi Kemungkinan, Algoritm Expectation Maximization (EM)*



**Abstract**

One distribution approach that is widely used in statistical data analysis is Weibull Distribution. The single Weibull distribution approach for the population in which there are several sub-populations is not good to use, because there are differences in the characteristics of the data between the sub-populations, so it is possible to have different parameters. The mixture distribution is expected to be able to find a solution to the distribution form with the condition of the data coming from several sub-populations. For example, age distribution or treatment period (to death) of cancer patients or Aids, with sub-populations being race or age group. This paper discusses the Weibull mixture distribution, specifically the biWeibull mixture distribution model that will be used in case study data consisting of two sub-populations. The biWeibull model is nothing but a linear combination of two Weibull (biWeibull) distribution with the following forms:

$$g\left(x,\overline{\alpha},\overline{\beta}\right)=\pi\, Weib\left(x,\alpha_1,\beta_1\right)+\left(1-\pi\right)Weib\left(x,\alpha_2,\beta_2\right)$$

where g(.) is the mixture distribution, then, and each of them is known as the proportion parameter, the scale vector parameter and the two shape vector. The model parameter estimation method used in the stages of model parameter estimation is described in this paper. In the field, the problem that arises next is how to estimate the three parameters of the model if the data is in hand. In this paper

the algorithm used to estimate the parameters above is the EM (Expectation Maximization) algorithm. The case study as an application of the algorithm was carried out for the case of the biWeibull mixture distribution approach.

*Keywords: Weibull Distribution, Possible function, Algorithm Expectation Maximization (EM)*

**Pendahuluan**

Dalam menentukan distribusi data penelitian, telah dikenal beberapa bentuk baik diskrit maupun kontinu. Selama ini model fungsi distribusi peluang dari data diasumsikan sebagai fungsi distribusi peluang tunggal, misalkan distribusi normal. Tetapi kenyataannya, seringkali plot distribusi data tidak sesuai dengan asumsi fungsi distribusi peluang tunggal. Hal ini menyebabkan hasil analisis tidak mewakili keadaan sebenarnya. Keadaan ini terjadi karena keheterogenan data atau adanya beberapa sub populasi dalam penelitian. Untuk memodelkan kondisi seperti ini pendekatan model distribusi campuran (*mixture distribution models*) yang merupakan kombinasi linier dari dua atau lebih fungsi densitas peluang kontinu (fungsi peluang untuk kasus data diskrit) lebih tepat. Bentuk fungsi distribusi campuran dapat dituliskan sebagai berikut :

$$g(x,\theta) = \pi_1 f_1(x,\theta_1) + \pi_2 f_2(x,\theta_2) + .... + \pi_k f_k(x,\theta_k) \quad \text{dimana} \quad \sum_{i=1}^{k} \pi_i = 1$$

Beberapa contoh kasus yang menyebabkan munculnya distribusi campuran diantaranya dalam tulisan Everitt (2001). Ia menjelaskan bahwa pada tahun 1890 Prof. W.R Weldon berkonsultasi dengan seorang statistikawan Karl Pearson mengenai pengukuran perbandingan panjang kepala dengan badan dari 1000 kepiting. Dari plot data diperoleh bahwa grafik menunjukkan menceng kanan. Weldon menduga bahwa kemencengan ini akibat data berasal dari dua tipe kepiting, tetapi dalam penelitian tidak diketahui dari tipe yang mana data berasal. Pearson berpendapat bahwa data dapat dimodelkan sebagai kombinasi linier dua fungsi distribusi peluang normal dengan bobot berbeda yang merupakan proporsi dari dua tipe kepiting, dengan fungsi peluang sebagai berikut :

$$g(x,\bar{\mu},\bar{\sigma}) = \pi_1 N(x,\mu_1,\sigma_1) + (1-\pi_1) N(x,\mu_2,\sigma_2)$$

Keatinge [1] banyak menemukan dalam kajian aktuaria, terutama untuk mengetahui distribusi *loss models*. Seringkali Keatinge tidak memperoleh model yang cocok ketika data diasumsikan berdistribusi tunggal. Ketika data yang diperoleh dipandang berasal dari beberapa sub populasi, sehingga fungsi distribusi data dimodelkan sebagai kombinasi linier dari fungsi distribusi sub populasi dengan bobot yang berbeda menghasilkan model yang lebih cocok. Oleh karena itu pendekatan distribusi campuran dapat memberikan model yang lebih realistis.



Analisis lebih lanjut dalam fungsi densitas peluang distribusi campuran adalah menentukan taksiran parameter yang tidak sederhana. Algoritma *Expectation Maximization* (EM) untuk struktur data tak lengkap digunakan dalam menentukan taksiran parameter-parameternya. Beberapa jenis distribusi, termasuk distribusi Weibull, khususnya distribusi tunggal banyak digunakan dalam analisis data masa hidup atau dalam model ketahanan (*survival models*)[2].

**Distribusi Campuran Berhingga (*Finite Mixture Distribution*)**

Distribusi campuran (*mixture distribution*) adalah kombinasi linier dari dua atau lebih distribusi statistik. Hal ini bisa terjadi karena diperkirakan bahwa distribusi populasi *tidak* berasal dari satu distribusi tertentu saja.

**Definisi** *Misalkan variabel acak **X** memiliki nilai dalam sebuah ruang sample **S**, dan distribusinya dapat dituliskan dalam sebuah fungsi peluang , berikut :*

$$g(x,\psi) = \pi_1 f_1(x) + \pi_2 f_2(x) + ... + \pi_k f_k(x) \quad (x \in S) \qquad (1)$$

*dengan $\psi$ adalah vektor parameter yang melibatkan $\pi_i$ dan paremetr yang ada dalam $f_i(x)$*

*dengan $0 \le \pi_i \le 1$, $i = 1,2,...,k$ dan $\sum_{i=1}^{k} \pi_i = 1$*

*dikatakan bahwa X merupakan distribusi campuran berhingga (**finite mixture distribution**), dan* $g(\cdot)$ *adalah fungsi densitas campuran berhingga (**finite mixture density function**). Parameter* $\pi_1, \pi_2, ..., \pi_k$ *dinamakan bobot atau proporsi campuran (**mixing weight or mixing proportion**) dan* $f_1(\cdot), f_2(\cdot), ..., f_k(\cdot)$ *masing-masing adalah fungsi densitas campuran (**component density function** of the mixture) untuk komponen i=1,2,…,k.*

Secara umum tidak ada syarat bahwa fungsi densitas distribusi campuran berasal dari kombinasi linier fungsi densitas peluang yang sama, walaupun demikian dalam pembahasan tulisan ini akan digunakan $f_1(x), f_2(x), ..., f_k(x)$ yang memiliki bentuk yang sama yaitu fungsi densitas Weibull dua parameter dengan parameter setiap fungsi berbeda. Kombinasi linier dari fungsi peluang setiap komponen dengan proporsi $\pi_i$, akan memenuhi syarat integral fungsi peluang adalah satu, akibatnya :

$$\int_{-\infty}^{\infty} g(x,\psi)dx = \int_{-\infty}^{\infty}\left[\sum_{i=1}^{k}\pi_i f_i(x)\right]dx = \sum_{i=1}^{k}\pi_i \int_{-\infty}^{\infty} f_i(x)dx = \sum_{i=1}^{k}\pi_i = 1$$

Komponen densitas dapat dituliskan dengan $f_i(x) = f_i(x,\theta_i)$ dengan $\theta_i$ menotasikan parameter-parameter dalam $f_i(x)$.

Fungsi densitas campuran *berhingga* dapat dituliskan dalam bentuk :

$$g(x,\psi) = \sum_{i=1}^{k}\pi_i f_i(x,\theta_i) \text{ dengan } (x \in S) \quad (2)$$

dengan $\psi = \left((\pi_1, \theta_1^T), ..., (\pi_k, \theta_k^T)\right)^T$ adalah himpunan seluruh parameter-parameter yang berbeda dalam model campuran.

**Fungsi Kemungkinan (*Likelihood*) Distribusi Campuran Weibull Berhingga Data Tak Lengkap**

Sebelum menentukan fungsi kemungkinan untuk data tak lengkap, akan dijelaskan terlebih dahulu definisi dari data tak lengkap.

**Definisi** Du [3] *Misalkan suatu sample acak* $x_1, x_2, ..., x_n$ *berasal dari populasi campuran, maka data tak lengkap (Incomplete data) dapat didefinisikan sebagai :*

$$\{y_j; j = 1,2,...,n\} = \{(x_j, z_j); j = 1,2,...,n\},$$

*untuk setiap* $z_j = (z_{1j}, ...., z_{kj})$ *adalah vektor indikator dari kategori ke-i (i=1,2,..,k), tidak diketahui.*

Fungsi likelihood yang dibentuk oleh data tak lengkap $(y_1, y_2, ...., y_n)$ dapat dituliskan sebagai berikut :

$$L(\psi; y_1, ...., y_n) = \prod_{j=1}^{n} g(x_j, \psi) = \prod_{j=1}^{n}\left[\sum_{i=1}^{k}\pi_i\ f_i(x_j, \theta_i)\right] \quad (3)$$

atau

$$\ell(\psi) = \sum_{j=1}^{n}\log\left[\sum_{i=1}^{k}\pi_i f_i(x_j, \theta_i)\right] \quad (4)$$

Untuk struktur tak lengkap dapat digambarkan pada tabel 1 berikut:

**Tabel 1 Struktur data tak lengkap (incomplete data)**

| *j* | X | ($z_{ij}$) | | | |
|---|---|---|---|---|---|
| | | Kategori (*i*) | | | |
| | | *1* | *2* | ... | *k* |

| | | | | | |
|---|---|---|---|---|---|
| 1 | $x_1$ | | | | |
| 2 | $x_2$ | | | | |
| . | . | | | | |
| . | . | | | | |
| . | . | | | | |
| $n$ | $x_n$ | | | | |

Ketika $f$ merupakan fungsi distribusi Weibull dengan dua parameter yang mempunyai bentuk fungsi peluang (Klein, 1997) :

$$f(x) = \alpha\beta\, x^{\beta-1} \exp\left(-\alpha\, x^{\beta}\right) \quad (5)$$

dimana :

$x = \text{Data variabel acak dengan } x \geq 0$

$\alpha = \text{Parameter "Scale" (Skala) sebagai karakteristik data hidup dengan } \alpha > 0$

$\beta = \text{Parameter "Slope" atau "Shape" (bentuk) dengan } \beta > 0$

maka fungsi densitas distribusi campuran Weibull didefinisikan sebagai berikut :

$$g(x,\psi) = \sum_{i=1}^{k} \pi_i \alpha_i \beta_i\, x^{\beta_i - 1} \exp\left(-\alpha_i\, x^{\beta_i}\right) \;;\; (x \in S) \quad (6)$$

dimana $\psi = \left((\pi_1, \alpha_1, \beta_1), \ldots, (\pi_k, \alpha_k, \beta_k)\right)^T$ . Dengan demikian logaritma fungsi kemungkinan dari persamaan (2) dengan fungsi densitas distribusi campuran Weibull data tak lengkap adalah sebagai berikut :

$$\ell(\psi) = \sum_{j=1}^{n} \log\left[\sum_{i=1}^{k} \pi_i \left\{\alpha_i \beta_i\, x_j^{\beta_i - 1} \exp\left(-\alpha_i\, x_j^{\beta_i}\right)\right\}\right] \quad (7)$$

**Algoritma Expectation Maximization (EM)**

Algoritma EM adalah alternatif algoritma MLE dengan data tak lengkap atau fungsi likelihood yang melibatkan variabel *latent/hidden*. Algoritma EM pertama kali dikenalkan oleh Dempster, Laird dan Rubin [4]. Untuk setiap iterasi algoritma EM, terdapat dua tahap, dinamakan *expectation step* atau *E-step* dan *maximization step* atau *M*-step, oleh karena itu dinamakan algoritma EM.



Secara umum prosedur algoritma EM dapat di tuliskan sebagai berikut :

*E-step*: dalam E-*step* dihitung ekspektasi likelihood untuk data lengkap (juga dinamakan fungsi *Q*), dengan kata lain membangun fungsi *Q* yang juga disebut "*Expected complete log-likelihood*" didasarkan pada parameter penduga awal (*previous guess*).

*M-step*: menaksir ulang (*re-estimate*) seluruh parameter dengan memaksimumkan fungsi *Q*,

Iterasi EM : proses *E-step* dan *M-step* ini terus diulang secara iterasi sampai hasil konvergen.

Misalkan fungsi log likelihood data lengkap untuk $\psi$ adalah $\log L_c(\psi) = \ell_c(\psi)$. Algoritma EM sebagai pendekatan penyelesaian masalah likelihood untuk data tak lengkap secara tidak langsung dilakukan secara iterasi pada $\ell_c(\psi)$ dengan menghitung ekspektasi bersyarat *x* yang merupakan data pengamatan untuk data tak lengkap, sampai ditemukan kecocokan parameter $\psi$ pada iterasi yang ke (*s+1*) dari algoritma EM.

Langkah perhitungan yang dilakukan adalah sebagai berikut :

E-Step : Hitung $Q(\psi;\psi_{(s)})$ , dimana

$$Q(\psi;\psi_{(s)}) = E_{\psi_{(s)}}\left\{\ell_c(\psi)|x\right\}$$

M-Step : Pilih $\psi_{(s+1)}$ untuk setiap nilai $\psi \in \Omega$ yang memaksimumkan $Q(\psi;\psi_{(s)})$, sehingga

$$Q(\psi_{(s+1)};\psi_{(s)}) \geq Q(\psi;\psi_{(s)}) \qquad \forall \psi \in \Omega$$

Operator $E_{\psi_{(s)}}$ adalah ekspektasi terhadap vektor parameter $\psi_{(s)}$. *E* dan *M-step* dilakukan secara berulang sampai konvergen. Kekonvergenan dapat ditentukan dengan menggunkan kriteria kekonvergenan $\|\psi_{(s+1)} - \psi_{(s)}\| < \varepsilon$ untuk $\varepsilon > 0$.

**Algoritma Generalized EM (GEM)**

Dempster,et.al [4] mendefinisikan algoritma GEM, untuk M-step memerlukan $\psi_{(s+1)}$ untuk dipilh sehingga

$$Q(\psi_{(s+1)};\psi_{(s)}) \geq Q(\psi_{(s)};\psi_{(s)})$$

diperoleh. Pemilihan $\psi_{(s+1)}$ untuk menaikan fungsi *Q*, $Q(\psi;\psi_{(s)})$, pada nilai $\psi = \psi_{(s)}$, daripada untuk memaksimumkan pada seluruh $\psi \in \Omega$.

Kondisi pada $\psi_{(s+1)}$ cukup untuk menjamin bahwa :

$$L(\psi_{(s+1)}) \geq L(\psi_{(s)})$$

Karena itu likelihood $L(\psi)$ tidak menurun setelah iterasi GEM, dan juga nilai likelihood barisan GEM akan konvergen jika terbatas atas.

**Kemonotonan Algoritma EM**

Fungsi likelihood untuk data tak lengkap tidak menurun setelah iterasi EM, yaitu :

$$\ell(\psi_{(s+1)};x) \geq \ell(\psi_{(s)};x) \text{ ; } s = 0,1,2,... \qquad (8)$$

Untuk membuktikan hal ini, misalkan

$$h(y|x;\psi) = \frac{k(y;\psi)}{g(x;\psi)}$$

sebagai densitas bersyarat *Y* dengan *X=x* diberikan. Diperoleh log likelihood sebagai berikut :

$$\begin{aligned}\ell(\psi;x) &= \log g(x;\psi) \text{ dengan } g(x;\psi) = L(\psi;x)\\ &= \log k(y;\psi) - \log h(y|x;\psi)\\ &= \ell(\psi;y) - \log h(y|x;\psi)\end{aligned}$$



Dengan mengambil ekspektasi dari persamaan (8) dan memperhatikan distribusi bersyarat dari *Y* dengan *X=x* diberikan, menggunakan taksiran $\psi_{(s)}$ untuk $\psi$ ,diperoleh :

$$\begin{aligned}\ell(\psi;x) &= E_{\psi_{(s)}}\left[\ell(\psi;y)|x\right] - E_{\psi_{(s)}}\left[log\ h(y|x;\psi)|x\right]\\ &= Q(\psi;\psi_{(s)}) - H(\psi;\psi_{(s)})\end{aligned}$$

dimana

$$H(\psi;\psi_{(s)}) = E_{\psi_{(s)}}\left[log\ h(y|x;\psi)|x\right]$$

Dari persamaan (IV.6), diperoleh

$$\begin{aligned}\ell(\psi_{(s+1)};x) - \ell(\psi_{(s)};x) &= \{Q(\psi_{(s+1)};\psi_{(s)}) - Q(\psi_{(s)};\psi_{(s)})\}\\ &\quad - \{H(\psi_{(s+1)};\psi_{(s)}) - H(\psi_{(s)};\psi_{(s)})\}\end{aligned}$$

Selisih fungsi *Q* pada ruas kanan adalah non negatif ketika $\psi_{(s+1)}$ dipilih, maka

$$Q(\psi_{(s+1)};\psi_{(s)}) \geq Q(\psi_{(s)};\psi_{(s)}) \qquad (9)$$

akibatnya, persamaan akan diperoleh jika selisih fungsi *H* pada ruas kanan persamaan (9) adalah non positif, yaitu :

$$H(\psi_{(s+1)};\psi_{(s)}) - H(\psi_{(s)};\psi_{(s)}) \leq 0$$

Sekarang, untuk setiap $\psi$ , diperoleh

$$H(\psi_{(s+1)};\psi_{(s)}) - H(\psi_{(s)};\psi_{(s)}) = E\left[\log\left\{h(y|x;\psi)/\, h(y|x;\psi_{(s)})\right| x;\psi_{(s)}\right\}\Big]$$

$$\leq \log\left[E\left\{h(y|x;\psi)/\, h(y|x;\psi_{(s)})\right| x;\psi_{(s)}\right\}\Big]$$

$$= \log \int_{y=\Phi(x)} h(y|x;\psi)dx$$

= 0

Ketaksamaan (10) sebagai akibat dari ketaksamaan Jansen dan kecekungan dari fungsi logaritma. Dari ketaksamaan (8) , menunjukkan bahwa likelihood $L(\psi)$ tidak menurun setelah iterasi EM. Likelihood akan naik jika ketaksamaan (9) sempurna (*strict*). Jadi $L(\psi_{(s)})$ konvergen secara monoton jika terbatas diatas.

**Algoritma EM untuk Distribusi campuran**

Dalam menentukan taksiran parameter distribusi campuran Weibull data tak lengkap digunakan penaksiran dengan memaksimukan fungsi menggunakan algoritma EM, yang terdiri dari *E-step* dan *M-step* .

Dalam struktur data tak lengkap *z* sebagai *hidden variable* tidak terobservasi, sehingga dapat dihitung dengan melakukan panaksiran densitas marjinal dari *x*

$$\sum_z f(x,z;\psi)$$

selanjutnya digunakan metoda maksimum likelihood. Untuk melakukan hal ini, diasumsikan sebuah nilai parameter penduga (*guess*) $\psi_{(0)}$ dengan nilai observasi hasil pengukuran yaitu $x_1,...,x_n$ . Kemudian dibangun fungsi $Q(\psi,\psi_{(0)},(x_1,...,x_n))$ dan diperoleh taksiran parameter dengan memaksimalkan *Q*.

Pada *E-step*, dapat ditentukan *Q* sebagai nilai ekspektasi log likelihood data lengkap dengan syarat *x* yang diperoleh dari pengukuran, sehingga :

$$Q(\psi,\psi_{(0)},(x_1,...,x_n)) = E\left[\ell_c(\psi)\ \middle|\ x=(x_1,...,x_n)\right]$$

dengan $\ell_c(\psi) = \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij}\log(\pi_i f_i(x_j;\theta_i))$

sebagai fungsi log-likelihood untuk data lengkap

$$= E\left[\sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij}\log(\pi_i f_i(x_j;\theta_i))\middle| x=(x_1,...,x_n)\right]$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{z} z_{ij}\log(\pi_i f_i(x_j;\theta_i)).f_i(z_{ij}|x_j;\psi_{(0)})$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{z} z_{ij}.f_i(z_{ij}|x_j;\psi_{(0)}).\log(\pi_i f_i(x_j;\theta_i)).$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{z} z_{ij}.\frac{f_i(x_j|z_{ij};\psi_{(0)})f_i(z_{ij})}{f_i(x_j;\psi_{(0)})}.\log(\pi_i f_i(x_j;\theta_i)).$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}.\frac{\pi_{i(0)} f_i(x_j;\theta_{i(0)})}{\sum_{i=1}^{k} f_i(x_j|z_{ij};\psi_{(0)})f_i(z_{ij})}\log\pi_i\ f_i(x_j;\theta_i)$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}.\frac{\pi_{i(0)} f_i(x_j;\theta_{i(0)})}{\sum_{i=1}^{k} \pi_{i(0)} f_i(x_j;\theta_{i(0)})}\log\pi_i\ f_i(x_j;\theta_i)$$

tulis $Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)=\sum_{j=1}^{n}\sum_{i=1}^{k}.\left\langle z_{ij}\right\rangle \log \pi_i \; f_i\left(x_j;\theta_i\right)$

dengan $\left\langle z_{ij}\right\rangle=\dfrac{\pi_{i(0)}\, f_i\left(x_j;\theta_{i(0)}\right)}{\sum_{i=1}^{k}\pi_{i(0)}\, f_i\left(x_j;\theta_{i(0)}\right)}$

Selanjutnya untuk menentukan taksiran parameter, dilakukan dengan memaksimumkan fungsi *Q* terhadap masing-masing parameternya. Langkah ini dinamakan sebagai *M-step*.

**Penaksiran Parameter Distribusi Campuran Weibull**

Dua langkah dalam penyelesaian algoritma EM, yaitu E-*step* dan M-*step* sebagaimana telah dijelaskan. Untuk fungsi distribusi Weibull campuran langkah E-*step* dimulai dengan membangun fungsi *Q*, sehingga diperoleh persamaan

$$Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)=\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log \pi_i+\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log\, f_i\left(x_j;\theta_i\right) \qquad (10)$$

dengan $\theta_i=(\alpha_i,\beta_i)$ dan $\pi_{i(0)}$, $\alpha_{i(0)}$ dan $\beta_{i(0)}$ adalah *initial value* parameter untuk *i = 1, 2,….,k*.

dengan $\left\langle z_{ij}\right\rangle=\dfrac{\pi_{i(0)}\, f_i\left(x_j;\alpha_{i(0)},\beta_{i(0)}\right)}{\sum_{i=1}^{k}\pi_{i(0)}\, f_i\left(x_j;\alpha_{i(0)},\beta_{i(0)}\right)}$, $f$ adalah fungsi Weibull dua parameter.

Taksiran $\pi_i$, $\alpha_i$ dan $\beta_i$ diperoleh dengan memaksimumkan fungsi $Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)$ dari persamaan (V.1) langkah ini dinamakan M-step, sehingga diperoleh

$$\frac{\partial}{\partial\pi_i}Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)=\frac{\partial}{\partial\pi_i}\left(\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log \pi_i\right)$$

dengan batasan bahwa $\sum_{i=1}^{k}\pi_i=1$, penyelesaian di atas dapat menggunakan metoda *lagrange multiplier*, sehingga

$$\frac{\partial}{\partial\pi_i}Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)=\frac{\partial}{\partial\pi_i}\left[\left(\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log \pi_i\right)+\lambda\left(1-\sum_{i=1}^{k}\pi_i\right)\right]=0$$



dengan $\lambda$ *lagrange multiplier*, diperoleh

$$\hat{\pi}_i=\frac{\sum_{j=1}^{n}\left\langle z_{ij}\right\rangle}{\sum_{i=1}^{k}\sum_{j=1}^{n}\left\langle z_{ij}\right\rangle}=\frac{\sum_{j=1}^{n}\left\langle z_{ij}\right\rangle}{n} \qquad (11)$$

Selanjutnya dari persamaan (10) dapat diperoleh taksiran $\alpha_i$ dengan memaksimumkan fungsi terhadap $\alpha_i$, sehingga

$$\frac{\partial}{\partial\alpha_i}Q\left(\psi,\psi_{(0)},(x_1,...,x_n)\right)=\frac{\partial}{\partial\alpha_i}\left[\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log \pi_i+\sum_{j=1}^{n}\sum_{i=1}^{k}\left\langle z_{ij}\right\rangle \log\left(\alpha_i\beta_i x_j^{\beta_i-1}\exp\left(-\alpha_i x_j^{\beta_i}\right)\right)\right]=0$$

diperoleh

$$\hat{\alpha}_i=\frac{\sum_{j=1}^{n}\left\langle z_{ij}\right\rangle}{\sum_{j=1}^{n}\left\langle z_{ij}\right\rangle x_j^{\beta_i}} \qquad (12)$$

demikian juga dalam menentukan taksiran $\beta_i$, diperoleh

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle}{\hat{\alpha}_i \sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i} \log x_j - \sum_{j=1}^{n}\langle z_{ij}\rangle \log x_j} \quad (13)$$

substitusi persamaan (V.3) terhadap (V.4), sehingga

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle \sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i}}{\sum_{j=1}^{n}\langle z_{ij}\rangle \sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i} \log x_j - \left(\sum_{j=1}^{n}\langle z_{ij}\rangle \log x_j\right)\left(\sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i}\right)} \quad (14)$$

untuk memperoleh taksiran masing-masing parameter, maka dilakukan iterasi dengan prosedur untuk setiap *i = 1, 2, . . ., k* sebagai berikut :

1. Tentukan data $x_j$ untuk *j=1,2,....,n*
2. Tentukan nilai inisial (*starting point*) $\pi_{i(0)}, \alpha_{i(0)}$ dan $\beta_{i(0)}$
3. Hitung secara iterasi persamaan $\langle z_{ij}\rangle$ sebagai *E-step*
4. Hitung secara iterasi persamaan (V.2), (V.3) dan (V.4) atau (V.5) untuk memperoleh nilai taksiran parameter sebagai *M-step*
5. Lakukan langkah 3 dan 4 sampai taksiran parameter konvergen
6. Diperoleh $\hat{\pi}_i, \hat{\alpha}_i$ dan $\hat{\beta}_i$

dari prosedur di atas dapat juga dibuat algoritma program sebagai berikut :

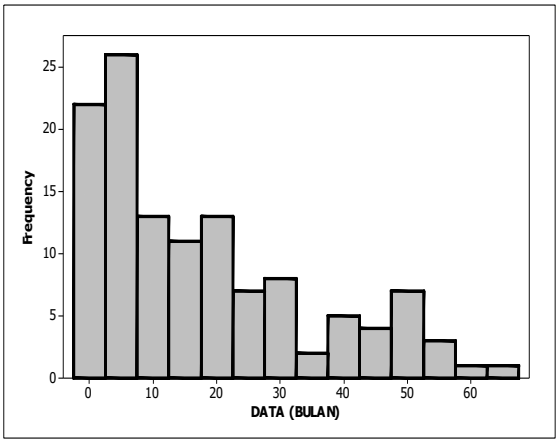
**Inisialisasi** : $\pi_{i(0)}, \alpha_{i(0)}, \beta_{i(0)}$

1. **repeat**
2. **for** i = 1 to k **do** // E-step
3. **for** j = 1 to n **do**
4. $$\langle z_{ij}\rangle = \frac{\pi_{i(0)} f_i\left(x_j;\alpha_{i(0)},\beta_{i(0)}\right)}{\sum_{j=1}^{k}\pi_{i(0)} f_i\left(x_j;\alpha_{i(0)},\beta_{i(0)}\right)}$$
$$f_i\left(x_j,\alpha_{i(0)},\beta_{i(0)}\right) = \alpha_{i(0)}\beta_{i(0)} x_j^{\beta_{i(0)}-1} \exp\left(-\alpha_{i(0)} x_j^{\beta_{i(0)}}\right)$$
5. **for** i = 1 to k **do** // M-step
6. $$\hat{\pi}_i = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle}{\sum_{i=1}^{k}\sum_{j=1}^{n}\langle z_{ij}\rangle} = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle}{n}$$
$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle}{\sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i}}$$
$$\hat{\beta}_i = \frac{\sum_{j=1}^{n}\langle z_{ij}\rangle \sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i}}{\sum_{j=1}^{n}\langle z_{ij}\rangle \sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i} \log x_j - \left(\sum_{j=1}^{n}\langle z_{ij}\rangle \log x_j\right)\left(\sum_{j=1}^{n}\langle z_{ij}\rangle x_j^{\beta_i}\right)}$$

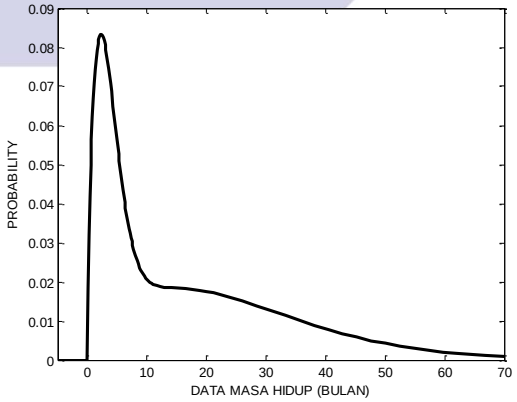
7. **until** model parameter converge

**Studi Kasus Distribusi Campuran BiWeibull**

Sebagai contoh kasus penerapan algoritma, diambil 123 data masa hidup pasien transpalansi ginjall dalam bulan periode 1982 – 1992 [2]. Histogram data dapat dilihat pada Gambar 1 , terlihat bahwa data cenderung menceng kanan.




**Gambar 1**
**Histogram Data Masa Hidup Pasien**




**Gambar 2**
**Plot Fungsi Distribusi Campuran biWeibull**

Pendekatan Distribusi tunggal untuk model seperti ini tampaknya terlalu dipaksakan dan belum cukup memuaskan karena adanya pengelompokkan lain pada data di atas 20 bulan, yang kita bisa anggap adanya dua sub populasi. Penggunaan distribusi campuran dengan kombinasi liner dua fungsi yang berbeda parameter lebih realistis.

Hasil estimasi parameter distribusi campuran dengan algoritma EM dan perbandingan distribusi tunggal dan campuran biWeibull ditunjukkan pada Tabel 2. Uji kecocokan (*goodness of fit*) model dengan *chi square test* [5] dilakukan untuk menentukan model terbaik. Tabel 3 menunjukkan hasil uji seluruh model distribusi (tunggal dan campuran). Model distribusi campuran bi Weibull memberikan hasil uji yang baik dengan *p-value* sebesar 0.096. Plot fungsi dengan estimasi parameter yang dihasilkan untuk distribusi campuran biWeibull pada Gambar 2 terlihat terjadi pengelompokan data menjadi dua kelompok, hal ini lebih realistis untuk menggambarkan pendekatan model distribusi kasus data di atas sesuai dengan sebaran data yang ditunjukan histogram data pada Gambar 2.



**Tabel 2**
**Hasil Estimasi Parameter Distribusi**

| Weibull (Tunggal) | BiWeibull (Campuran) |
|---|---|
| $\hat{\alpha}$ = 18,015 | $\hat{\pi}_1$ =0,329 $\hat{\pi}_2$ =0,671 |
| $\hat{\beta}$ = 1,021 | $\hat{\alpha}_1$ =3,964 $\hat{\alpha}_2$ =27,741 |
| | $\hat{\beta}_1$ =1,596 $\hat{\beta}_2$ =1,602 |

**Tabel 3**
**Hasil Uji Kecocokan (goodness of fit)** $\chi^2$

| (Weibull) | $\chi^2$ | *p-value* | *H₀* |
|---|---|---|---|
| tunggal | 26,6389 | 0,0052 | tolak |
| campuran | 14,8233 | 0,0959 | terima |

**Kesimpulan**

Tulisan ini membahas langkah-langkah estimasi parameter distribusi campuran Weibull secara umum dengan menggunakan algoritma EM. Sifat-sifat algoritma EM secara teoretis diuraikan. Studi kasus penerapan algoritma EM untuk estimasi model campuran Weibull menunjukkan bahwa untuk kasus dengan data yang yang terdapat sub populasi pendekatan distribusi biWeibull lebih baik dibandingkan dengan pendekatan Weibull tunggal.

**Referensi**

[1] Keatinge, C.L. (1999), *Modeling Losses With The Mixed Exponential Distribution*, www.casact.org/pubs/proceed/proceed99/99578.pdf, 654-698

[2] Klein, J.P., Moeschberger, M.L. (1997), *Survival Analysis : Techniques for Censored and Truncated Data*, Springer-Verlag Newyork, Inc, 8-9

[3] Du, J. (2002), Combined *Algorithms for Constrained Estimation of finite Mixture Distributions With Grouped Data and Conditional Data*, Thesis Master of Science, McMaster University Hamilton, Ontario, 1-12, 24-29

[4] Dempster, A.P., Laird, N.M., Rubin, D.B.(1977), *Maximum Likelihood from Incomplete Data via EM Algorithm*., J.R Statists.Soc., B39, 1-38

[5] Ebeling, C. E. (1997), *An Introduction to Reliability and Maintainability Engeneering*, McGraw-Hill International Inc, New York, 58-65, 392-401