AdMathEdu

ISSN
2088 - 687X

JURNAL PENDIDIKAN MATEMATIKA, ILMU MATEMATIKA
DAN MATEMATIKA TERAPAN

IMPLEMENTASI ALGORITMA EXPECTATION-MAXIMIZATION UNTUK ESTIMASI PARAMETER MODEL DISTRIBUSI CAMPURAN

Suparman

Program Studi Pendidikan Matematika FKIP UAD

Jl. Prof. Dr. Soepomo, SH. Janturan Yogyakarta

suparmancict@yahoo.co.id

ABSTRAK

Distribusi campuran merupakan distribusi yang sangat fleksibel untuk memodelkan data. Jika distribusi campuran dicocokkan terhadap data, maka umumnya parameternya tidak dketahui. Tulisan ini mengkaji masalah penaksiran parameter distribusi campuran.

Metode yang digunakan untuk mengestimasi parameter distribusi campuran adalah metode kemungkinan maksimum. Namun penaksir kemungkinan maksimum tidak dapat ditemukan secara analitik. Untuk mengatasi masalah tersebut diusulkan Algoritma EM. Algoritma EM terdiri atas dua tahap yaitu tahap ekspektasi dan tahap maksimisasi. Algoritma EM konvergen menuju penaksir kemungkinan maksimum.

Kinerja Algoritma EM diuji dengan menggunakan data simulasi. Hasil pengujian menunjukkan bahwa Algoritma EM dapat mengestimasi parameter distribusi campuran dengan baik. Selanjutnya Algoritma EM diimplementasikan pada data riil yang dijumpai dalam kehidupan sehari-hari.

Kata Kunci : Distribusi Campuran, Penaksir Kemungkinan Maksimum, Algoritma EM.

ABSTRACT

Mixture distribution is a very flexible distribution for modeling data. If the mixture distribution fitted to the data, the parameters are generally not known. This paper examines the estimation problem of the distribution mixture parameter.

The method used to estimate the parameters of a mixture distribution is the maximum likelihood method. However, the maximum likelihood estimator can not be found analytically. To overcome these problems, the EM algorithm proposed. EM algorithm consists of two phases: expectation and maximization stage. EM algorithm converges to the maximum likelihood estimator.

Performance of the EM algorithm is tested using simulated data. The results show that the EM algorithm can estimate the mixture distribution parameter. Further EM algorithm is implemented on real data encountered in everyday life.

Keywords : Mixture distributions, Maximum Likelihood Estimator, EM algorithm.

Pendahuluan

Model distribusi campuran (mixture distribution) merupakan model distribusi yang sering digunakan dalam bidang rekayasa dan kedokteran. Dalam bidang teknik industri, model distribusi campuran digunakan untuk penentuan spesifikasi di industri makanan. Dalam kedokteran, model distribusi campuran digunakan sebagai model distribusi dari sinyal resonansi magnetik. Dalam bidang teknik komputer, model distribusi campuran digunakan untuk mendeteksi struktur komunitas jaringan. Aplikasi dari distribusi campuran dapat ditemukan diberbagai literatur, misalnya Schlattman (2009) dan Mergersen *et al.* (2011).

Jika model distribusi campuran dicocokkan terhadap data, maka umumnya parameter model tidak diketahui. Mengingat begitu banyak model distribusi campuran, di sini akan dibatasi pada model distribusi campuran dengan bentuk fungsi kepadatan probabilitas sebagai berikut :

$$f(x|\ \theta) = \theta g(x) + (1-\theta)h(x)$$

di mana θ adalah parameter. Sedangkan g(x) dan h(x) keduanya merupakan distribusi yang diketahui. Misalkan $x_1, x_2, \cdots, x_n$ merupakan sampel random yang diambil dari suatu populasi yang berdistribusi campuran. Berdasarkan sampel random tersebut, permasalahan utama adalah cara mengestimasi parameter θ.

Estimasi parameter dengan menggunakan Metode Kemungkinan Maksimum tidak dapat ditentukan secara analitik karena fungsi kemungkinan untuk parameter θ mempunyai bentuk

$$L(\theta|\ x) = \prod_{i=1}^{n} [\theta g(x) + (1-\theta)h(x)]$$

Untuk mengatasi masalah tersebut, dalam penelitian ini digunakan Algoritma Expectation-Maximization (Dempster *et al.*, 1977). Algoritma Expectation-Maximization merupakan teknik komputasi yang sering digunakan dalam komputasi statistika (Gentle, 2002). Kinerja algoritma diuji dengan menggunakan data simulasi. Apabila dengan data simulasi, Algoritma Expectation-Maaximization dapat mengestimasi parameter θ dengan baik maka Algoritma Expextation-Maximization diimplementasikan pada data real.

Metode Penelitian

Penelitian dimulai dengan mengkaji berbagai pustaka terkait dengan metode estimasi kemungkinan maksimum, distribusi campuran, algoritma expectation-maximization dan estimasi kemungkinan maksimum dengan menggunakan algoritma expectation-maximization. Berdasarkan teori yang

dihasilkan dari berbagai kajian pustaka tersebut, selanjutnya dibuat program komputasinya dengan menggunakan MATLAB. Program komputer digunakan untuk menemukan penaksir kemungkinan maksimum distribusi campuran dengan menggunakan expectation-maximization.

Penaksir Kemungkinan Maksimum

Misalkan $x_1, x_2 . \cdots, x_n$ merupakan sampel random yang diambil dari suatu populasi dengan fungsi kepadatan $f(x|\ \theta)$. Metode kemungkinan maksimum adalah suatu metode untuk memperoleh penaksir untuk parameter θ sedemikian sehingga dapat membuat maksimum fungsi kemungkinan

$$L(\theta) = f(x_1, x_2, \cdots, x_n|\ \theta)$$

$$= \prod_{i=1} f(x_i|\ \theta)$$

Untuk beberapa distribusi, pencarian nilai maksimum untuk fungsi kemungkinan dapat dilakukan dengan menggunakan diferensial.

Namun untuk banyak distribusi, termasuk distribusi campuran, pencarian tidak dapat dilakukan dengan menggunakan diferensial. Salah satu metode yang dapat digunakan untuk mengatasi hal ini adalah algoritma Expectation-Maximization.

Distribusi Campuran

Misalkan x_1, x_2, …, x_n merupakan variabel random. Variabel random ini berdistribusi campuran dengan fungsi kepadatan probabilitas (Robert and Casella, 1999) :

$$f(x|\ \theta) = \theta g(x) + (1-\theta)h(x)$$

di mana θ adalah parameter. Sedangkan g(x) dan h(x) keduanya merupakan distribusi yang diketahui. Teori mengenai distribusi campuran dapat ditemukan dalam Box and Draper (2007), McLachlan and Khrishnan (2008) dan McLachlan and Peel (2000).

Dari fungsi kepadatan probabilitas $f(x|\ \theta)$, fungsi kemungkinan untuk parameter θ mengambil bentuk

$$L(\theta|\ x) = f(x_1, x_2, \cdots, x_n|\theta)$$

$$= \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \prod_{i=1}^{n} [\theta g(x_i) + (1-\theta)h(x_i)]$$

Penaksir kemungkinan maksimum untuk parameter θ adalah nilai θ yang membuat fungsi kemungkinan $L(\theta)$ maksimum. Namun hal ini tidak dapat ditentukan secara analitik. Untuk itu diusulkan penggunaan Algoritma Expectation-Maximization untuk menentukan penaksir kemungkinan maksimum untuk parameter θ. Teori mengenai Algoritma Expectation-

Maximization dapat ditemukan dalam berbagai literatur. Misalnya Gupta and Chen (2011), McLachlan and Khrishnan (2008) dan Watanabe and Yamaguchi (2004)..

Algoritma Expectation-Maximization

Untuk melakukan hal ini, mula-mula tambahkan $z_1, z_2, \ldots, z_n$ di mana z_i (i = 1, 2, ..., n) menunjukkan dari mana x_i berasal. Sehingga

$$x_i \mid z_i = 1 \sim g(x)$$

$$x_i \mid z_i = 0 \sim h(x)$$

Misalkan $L^c(\theta \mid x, z)$ menyatakan fungsi kemungkinan lengkap. Maka bentuk fungsi kemungkinan lengkap adalah

$$L^c(\theta \mid x, z) = \prod_{i=1}^{n} [z_i g(x_i) + (1 - z_i) h(x_i)] \times \theta^{z_i} (1-\theta)^{1-z_i}$$

Persamaan yang menjadi dasar dari Algoritma Expectation-Maximization adalah

$$k(z \mid \theta, x) = \frac{f(x, z \mid \theta)}{g(x \mid \theta)}$$

di mana $k(z \mid \theta, x)$ adalah distribusi bersyarat dari data z diberikan data x. Persamaan ini menghubungkan antara fungsi kemungkinan lengkap $L^c(\theta \mid x, z)$ dan fungsi kemungkinan $L(\theta \mid x)$, yaitu untuk suatu $\theta^{(0)}$:

$$\log L(\theta \mid x) = E_{z \mid x, \theta^{(0)}} [\log L^c(\theta \mid x, z)] - E_{z \mid x, \theta^{(0)}} [\log k(z \mid \theta, x)]$$

Di mana ekspektasi diambil terhadap distribusi $k(z \mid \theta, x)$. Pendekatan Algoritma Expectation-Maximization untuk memaksimumkan $L(\theta \mid x)$ cukup dimaksimumkan $E[\log L^c(\theta \mid x, z)]$.

Sehingga dimulai dari $\theta^{(0)}$, algoritma Expectation-Maximization terdiri atas dua langkah dan iterasi ini dilakukan hingga konvergen :

(1) Langkah Expectation : hitung

$$q^{(k)}(\theta) = E_{z \mid x, \theta^{(k-1)}} (\log L^c(\theta \mid x, z)$$

$$= \frac{\theta^{(k-1)} g(x)}{[\theta^{(k-1)} g(x) + (1 - \theta^{(k-1)}) h(x)]}$$

(2) Langkah Maximization : tentukan $\theta^{(k)}$ untuk memaksimumkan $q^{(k)}(\theta)$.

$$\theta^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\theta^{(k-1)} g(x_i)}{\theta^{(k-1)} g(x_i) + (1 - \theta^{(k-1)}) h(x_i)}$$

Barisan $\theta^{(1)}$, $\theta^{(2)}$, ... konvergen menuju maksimum lokal fungsi kemungkinan maksimum $L(\theta \mid x)$.

Berikut merupakan listing program yang ditulis dalam instruksi bahasa pemrograman MATLAB untuk penentuan penaksir kemungkinan maksimum untuk parameter distribusi campuran :

```
clear all
clc

n = 250;
```

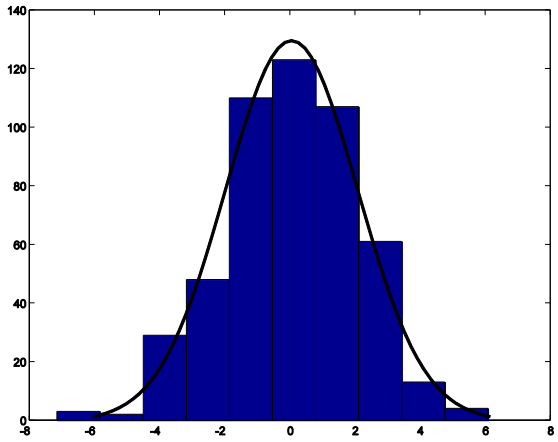
```
mu1 = 0; sigma1 = 2;
mu2 = 5; sigma2 = 3;
teta = 0.6;

g = normrnd(mu1, sigma1, n,1);
h = normrnd(mu2, sigma2, n,1);
x = teta*g+(1-teta)*h;

gx = normpdf(x,mu1,sigma1);
hx = normpdf(x,mu2,sigma2);
tetaj = 0.1;

m = 50;
mtetaj = zeros(m,1);
for j = 1:m,
    metaj(j) = mean((tetaj*gx)./(tetaj*gx
                +(1-tetaj)*hx));
end;
plot(mtetaj)
```

Hasil dan Pembahasan

Sebagai ilustrasi, di sini metode bootstrap akan diimplementasikan untuk menguji hipotesis mengenai dua mean populasi pada data sintesis (studi simulasi) dan data riil (studi kasus). Studi simulasi ditempuh untuk mengkonfirmasi kinerja dari pendekatan yang diusulkan apakah dapat bekerja dengan baik. Sedangkan studi kasus diberikan untuk memberikan contoh penerapan penelitian dalam memecahkan permasalahan dalam kehidupan sehari-hari. Komputasi ditulis dalam bahasa pemrograman MATLAB (Chapman, 2009).

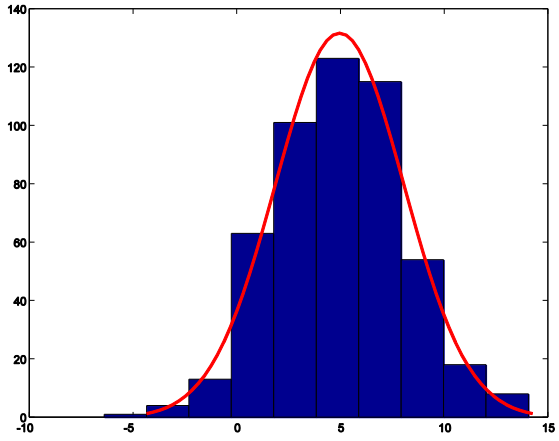
Data Sintesis Berdistribusi Campuran Normal

Gambar 1 menunjukkan 500 data simulasi dari populasi berdistribusi N(0,4).



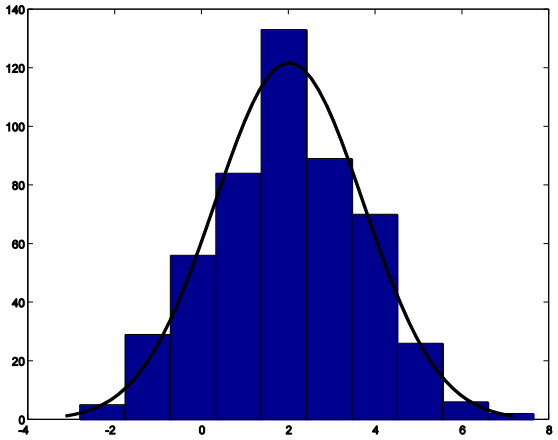
Gambar 1 : Superposisi antara histogram dan kurva untuk 500 data simulasi berdistribusi N(0,4)

Gambar 2 menunjukkan 500 data simulasi dari populasi berdistribusi N(5,9).



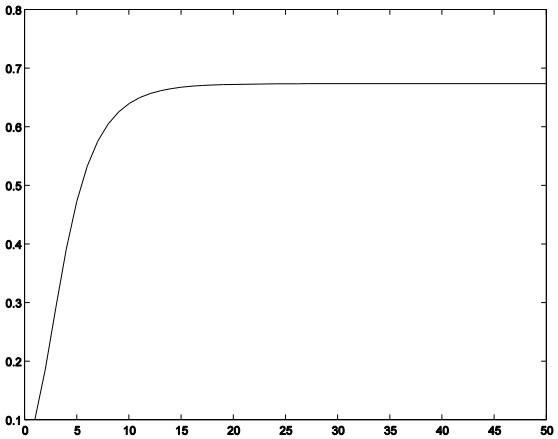
Gambar 2 : Superposisi antara histogram dan kurva untuk 500 data simulasi berdistribusi N(5,9)

Selanjutnya, dengan mengambil n = 500, θ = 0.6, g(x) = N(0,4), dan h(x) = N(5,9) dibuat data simulasi berdistribusi campuran. Datanya ditunjukkan pada Gambar 3



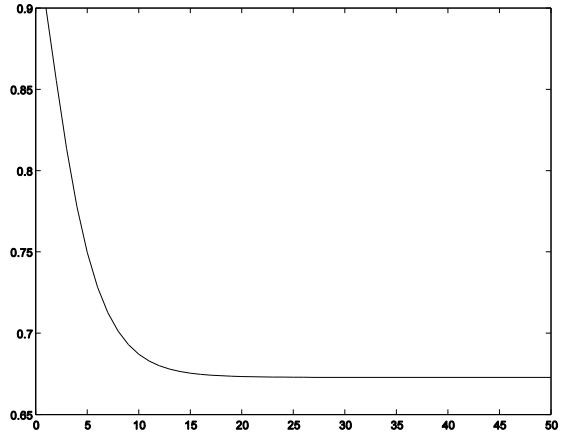
Gambar 3 : Superposisi antara histogram dan kurva untuk 500 data simulasi berdistribusi 0.6N(0,4)+0.4N(5,9)

Berdasarkan data ini, selanjutnya parameter θ diestimasi dengan menggunakan algoritma Expectation-Maximization. Dengan mengambil $\theta^{(0)}$ = 0.1, diperoleh $\hat{\theta} = 0.6726$. Hasil tiap iterasi dari algoritma Expectation-maximization ditunjukkan oleh Gambar 4



Gambar 4 : Lima puluh iterasi Algoritma EM dengan nilai awal 0.1 berdistribusi campuran normal.

Untuk menunjukkan bahwa pengambilan nilai awal dapat dilakukan secara sembarang, maka pada penghitungan penaksir parameter θ diulangi dengan menggunakan $\theta^{(0)}$ = 0.9. Ternyta diperoleh $\hat{\theta} = 0.6726$. Hasil tiap iterasi dari algoritma Expectation-maximization ditunjukkan oleh Gambar 5



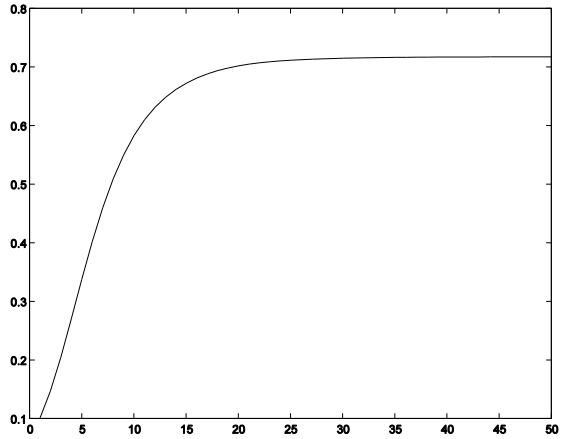
Gambar 5 : Lima puluh iterasi Algoritma EM dengan nilai awal 0.9 berdistribusi campuran normal.

Jadi tidak terdapat perbedaan nilai estimasi baik dimulai dari nilai awal 0.1 maupun 0.9.

Data Sintesis berdistribusi Campuran Exponensial

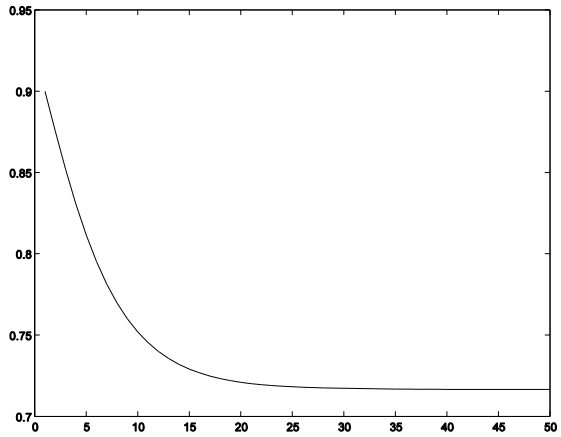
Selanjutnya, dengan mengambil n = 500, θ = 0.8, g(x) = EXP(1), dan h(x) = EXP(5)

dibuat data simulasi berdistribusi campuran. Berdasarkan data ini, selanjutnya parameter θ diestimasi dengan menggunakan algoritma Expectation-Maximization. Dengan mengambil $\theta^{(0)}$ = 0.1, diperoleh $\hat{\theta} = 0.7167$. Hasil tiap iterasi dari algoritma Expectation-maximization ditunjukkan oleh Gambar 6



Gambar 6 : Lima puluh iterasi Algoritma EM dengan nilai awal 0.1 dan data berdistribusi campuran eksponensial.

Untuk menunjukkan bahwa pengambilan nilai awal dapat dilakukan secara sembarang, maka pada penghitungan penaksir parameter θ diulangi dengan menggunakan $\theta^{(0)}$ = 0.9. Ternyata diperoleh $\hat{\theta} = 0.7167$. Hasil tiap iterasi dari algoritma Expectation-maximization ditunjukkan oleh Gambar 7



Gambar 7 : Lima puluh iterasi Algoritma EM dengan nilai awal 0.9 dan data berdistribusi campuran eksponensial.

Jadi tidak terdapat perbedaan nilai estimasi baik dimulai dari nilai awal 0.1 maupun 0.9.

Kesimpulan

Dalam artikel ini dikembangkan estimasi parameter distribusi campuran dengan menggunakan Algoritma EM.

Kinerja Algoritma EM diuji menggunakan data simulasi dan nampak bahwa Algoritma EM dapat mengestimasi parameter distribusi campuran dengan baik

Pustaka

Box, G.E.P. and Draper, N.R. (2007) Response, Mixtures, and Ridge Analyses**,** John Wiley & Sons.

Chapman, S.J. (2009) Essentials of Matlab Programming, Cengage Learning.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via EM Algorithm, J. Roy. Statist. Soc. Ser. B, 39, 1-38..

Gentle, J.E. (2002) Elements of Computational Statistics, Springer.

Gupta, M.R. and Chen, Y. (2011) Theory and Use of the EM Algorithm, Now Publishers

Mergersen, K.L., Robert, C.P. and Titterngton, D.M. (2011) Mixtures : Estimation and Applications, John Wiley and Sons.

McLachlan, G. and Khrishnan, T. (2008) The EM Algorithm and Extensions**,** John Wiley & Sons.

McLachlan, G. and Peel, D. (2000) Finite Mixture Models**,** John Wiley & Sons.

Robert, C.P. and Casella, G. (1999) Monte Carlo Statistical Methods**,** Springer Texts in Statistics.

Schlattman, P. (2009) Medical Applications of Finite Mixture Models**,** Springer.

Watanabe, M. and Yamaguchi, K. (2004) The EM Algorithm and Related Statistical Models**,** Marcel Dekker.