Jurnal MIPA 37 (1): 71-78 (2014)

Jurnal MIPA

http://journal.unnes.ac.id/nju/index.php/JM

KONVERGENSI ESTIMATOR DALAM MODEL *MIXTURE* BERBASIS *MISSING DATA*

N Dwidayati✉, SH Kartiko, Subanar

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia

Info Artikel

Sejarah Artikel:
Diterima Februari 2014
Disetujui Maret 2014
Dipublikasikan April 2014

Keywords:
mixture;
missing data;
algoritma EM;
konvergensi.

Abstrak

Model *mixture* dapat mengestimasi proporsi pasien yang sembuh (*cured*) dan fungsi survival pasien tak sembuh (*uncured*). Pada kajian ini, model mixture dikembangkan untuk analisis cure rate berbasis missing data. Ada beberapa metode yang dapat digunakan untuk analisis *missing data.* Salah satu metode yang dapat digunakan adalah Algoritma EM, Metode ini didasarkan pada dua langkah, yaitu: (1) *Expectation Step* dan (2) *Maximization Step*. Algoritma EM merupakan pendekatan iterasi untuk mempelajari model dari data dengan nilai hilang melalui empat langkah, yaitu(1) pilih himpunan inisial dari parameter untuk sebuah model, (2) tentukan nilai ekspektasi untuk data hilang, (3) buat induksi parameter model baru dari gabungan nilai ekspekstasi dan data asli, dan (4) jika parameter tidak *converged*, ulangi langkah 2 menggunakan model baru. Berdasar kajian yang dilakukan dapat ditunjukkan bahwa pada algoritma EM, *log-likelihood* untuk *missing data* mengalami kenaikan setelah dilakukan setiap iterasi dari algoritmanya. Dengan demikian berdasar algoritma EM, barisan likelihood konvergen jika likelihood terbatas ke bawah.

Abstract

Model mixture can estimate the proportion of recovering (cured) patients and function of survival but do not recover (uncured) patients. In this study, a model mixture has been developed to analyze the curing rate based on missing data. There are some methods applicable to analyze missing data. One of the methods is EM Algorithm, This method is based on two (2) steps, i.e.: (1) Expectation Step and (2) Maximization Step. EM Algorithm is an iteration approach to study the model from data with missing values in four (4) steps, i.e. (1) to choose initial set from parameters for a model, (2) to determine the expectation value for missing data, (3) to make induction for the new model parameter from the combined expectation values and the original data, and (4) if parameter is not converged, repeat step 2 using new model. The current study indicated that for the EM algorithm, the log-likelihood function of the missing data has increased after each iteration of the algorithm. Thereby based on the EM algorithm, the on-line likelihood is convergent if the likelihood is limited downwards.

✉ Alamat korespondensi:
Gedung D7 Lantai 1, Kampus Unnes Sekaran,
Gunungpati, Semarang, 50229
E-mail: noengkd_unnes@yahoo.co.id

ISSN 0215-9945

PENDAHULUAN

Cure models adalah model survival yang dikembangkan untuk estimasi proporsi pasien yang sembuh dalam studiklinik. Model ini selain digunakan untuk mengestimasi proporsi pasien yang sembuh juga digunakan untuk mengestimasi probabilitas survival pasien yang tak sembuh sampai pada batas waktu yang diberikan. Model tersebut dikembangkan oleh Boag (1949) dengan mengestimasi proporsi pasien yang sembuh pada *treatment* penderita kanker mulut dan kerongkongan, leher, kandungan dan payudara. Model ini dinamakan model *mixture* karena dapat mengestimasi proporsi pasien yang sembuh dan fungsi survival pasien tak sembuh.

Model *mixture* didasarkan pada fungsi distribusi probalitas, baik diskret maupun kontinu dalam bentuk mixture:

$$f(.) = \sum_{i=1}^{n} p_i f_i(.) \quad (1)$$

dengan $f(.)$ menyatakan fungsi densitas yang menyatakan elemen-elemen *mixture*, dan p_i menyatakan *mixture weight* dengan $p > 0, \sum_{i=1}^{n} p_i = 1$ serta bilangan bulat n menyatakan banyak elemen *mixture.*

Model *mixture* dikatakan model *cure mixture* parametrik jika menggunakan distribusi probabilitas standar seperti distribusi eksponensial, Weibull, Gompertz dan *Generalized F*. Diskusi tentang model *mixture* parametrik dapat dilihat pada Boag (1949), Jones *et al.* (1981), Farewell (1982, 1986*),* Cantor & Shuster (1992), Ghitany *et al.* (1994).

Model *mixture* tanpa menggunakan distribusi probabilitas standar dinamakan model *cure mixture* nonparametrik. Model *mixture* nonparametrik untuk estimasi *cure rate* diperkenalkan oleh Kuk dan Chen (1992) dengan mengaplikasikan asumsi *proportional hazard* (PH) ke distribusi *failure time* pasien yang tak sembuh. Metode yang digunakan analog dengan yang digunakan dalam model PH Cox dan modelnya adalah nonparametrik. Walaupun demikian hasil dari metode mereka tak dapat lepas dari pendekatan Monte-Carlo dalam hal fungsi likelihood yang sering digunakan. Taylor (1995) menggunakan estimator survivor Kaplan-Meier untuk mengestimasi distribusi *failure time* pasien yang tak sembuh dan algoritma *EM* untuk mengestimasi γ, tetapi model ini tak dapat memberikan covariat dalam distribusi *failure time* pasien yang tak sembuh.

Salah satu aplikasi model *mixture* adalah analisis pada studi klinik dengan pasien berpotensi sembuh. Farewell (1986) mengaplikasikan model *mixture* parametrik yaitu model *mixture* Weibull untuk analisis pada studi klinik dari penderita kanker payudara sebagai kasus khusus. Model tersebut digunakan untuk menemukan efek kovariat pada *failure time* pasien yang tak sembuh dan pada *cure rate*. Fokus penelitian tersebut pada estimasi proporsi pasien yang sembuh dan distribusi *failure time* pasien yang tak sembuh

Misal T variabel random tak negatif, menyatakan *failure time* yang terobservasi dan $f(t \mid x, z)$ menyatakan fungsi kepadatan peluang (f.k.p.) serta $S(t|x,z)$ menyatakan fungsi survival T, dengan x dan z adalah nilai-nilai observasi dari dua vektor covariat yang pada distribusi T mungkin dependen. Dalam model *mixture* ini distribusi T mempunyai *mixture* hingga. Fungsi densitas dan fungsi survival dari T berturut-turut diberikan oleh:

$$f(t|x,z) = \pi(z)\, f_u(t|x) \quad (2)$$

$$S(t|x,z) = \pi(z)\, s_u(t|x) + 1 - \pi(z) \quad (3)$$

dengan $\pi(z)$ menyatakan proporsi pasien yang tidak sembuh yang dependen pada z, yang dalam bentuk logistik dinyatakan sebagai:

$$\log\left[\pi(z)/\{1-\pi(z)\}\right] = \gamma' z \quad (4)$$

$f_u(t|x)$ dan $s_u(t|x)$ berturut-turut menyatakan f.k.p. dan fungsi survival distribusi *failure time* pasien yang tak sembuh yang mungkin dependen pada nilai *x*.

Fungsi survival dan f.k.p pasien yang sembuh merupakan himpunan yang berturut-turut sama dengan satu dan nol untuk semua nilai hingga *t*, karena diasumsikan bahwa pasien tidak pernah kambuh atau tidak meninggal. Oleh karena itu, *failure time* dapat didefinisikan (dengan kesepakatan) tak hingga.

Secara spesifik $f_u(t|x)$ atau $s_u(t|x)$ dapat digunakan dalam model *mixture* baik parametrik maupun nonparametrik. Dalam model *mixture* parametrik, distribusi yang dimaksudkan diasumsikan sebagai distribusi *failure time* dari pasien yang tak sembuh. Fungsi kepadatan peluang

(f.k.p) $f_u(t|x)$ dan fungsi survival $s_u(t|x)$ diturunkan dari distribusi yang tergantung pada satu atau lebih parameter.

Peluang untuk sembuh, yang biasanya dikenal dengan *cure rate* atau *surviving fraction,* didefinisikan sebagai nilai asimtotik dari fungsi survival untuk t (waktu) menuju tak hingga, ditulis $\lim_{t\to\infty} S(t)$.

Misal T menyatakan waktu survival terobservasi. Inferensi statistika pada *cure rate* didasarkan pada sebarang fungsi survival $S(t) = P(T \geq t)$ dapat dinyatakan dalam bentuk:

$$S(t) = a + (1-a)S_0(t) \tag{5}$$

dengan $a = P(T = \infty)$ adalah peluang untuk sembuh, dan $S_0(t) = P(T \geq t \mid X < \infty)$.

Problem yang terkait dengan model parametrik adalah kesulitan dalam memverifikasi distribusi yang digunakan dalam model. Secara umum, distribusi yang digunakan adalah *generalized gamma*. (Yamaguchi,1992)

Distribusi *generalized gamma* adalah distribusi 3(tiga) parameter. Salah satu versi dari distribusi *generalized gamma* menggunakan parameter $k, \beta,$ dan θ. Fungsi densitas peluang dari distribusi *generalized gamma* adalah:

$$f(t) = \frac{\beta}{\theta\,\Gamma(k)}\left(\frac{t}{\theta}\right)^{k\beta-1} e^{-\left(\frac{t}{\theta}\right)^{\beta}} \tag{6}$$

dengan $\theta > 0$ parameter skala, $\beta > 0$ *dan* $k > 0$ *shape parameters* dan $\Gamma(k)$ fungsi gamma dari k yang didefinisikan dengan:

$$\Gamma(k) = \int_0^{\infty} t^{k-1} e^{-t}\, dt \tag{7}$$

Algoritma EM untuk model *mixture*

Misal dipunyai data dalam bentuk (t_i, δ, x_i, z_i), i = 1, 2, … n dengan t_i menyatakan *survival time* yang diobservasi untuk pasien ke-*i*, $\delta_i$menyatakan *censoring indicator* yang berharga 1 jika t_i tersensor dan 0 untuk t_i yang lain sedangkan x_i dan z_i menyatakan nilai terobservasi dari dua covariat.

Fungsi *likelihood* untuk model *mixture* adalah:

$$L = \prod_{i=1}^{n} \left|\pi(z_i) f_u(t_i|x_i)\right|^{\delta_i} \left[\pi(z_i) S_u(t_i|x_i) + 1 - \pi(z_i)\right]^{1-\delta_i} \tag{8}$$

Untuk mengestimasi parameter tak diketahui dalam model *mixture* digunakan algoritma *EM*. Ambil c_i indikator *cure status* pasien ke-*i* yang berharga 1 jika pasien sembuh dan 0 untuk yang lain, i =1, 2, …, n. Jika $\delta_i = 1$ maka $c_i = 0$, tetapi jika $\delta_i = 0$ maka c_i tidak terobservasi dan dapat bernilai 1 atau 0.

Sebagai catatan $1-= P(ci = 1|zi)$. Misal c=(c_1, …, c_n). Diberikan c_i yaitu data lengkap yang diperoleh maka fungsi log likelihood sebagai berikut.

L_c=log

$$\prod_{i=1}^{n} \left|\{\pi(z_i) f_u(t_i|x_i)\}^{1-c_i}\right|^{\delta_i} \left[\{1-\pi(z_i)\}^{c_i} \{\pi(z_i) S_u(t_i|x_i)\}^{1-c_{iu}}\right]^{1-\delta_i} \tag{9}$$

E-step dalam perhitungan algoritma *EM*, ekspektasi dari (8) digunakan untuk mengestimasi $f_u(t_i \mid x_i), S_u(t_i \mid x_i)$ dan $\pi(z_i)$. Diperoleh fungsi sebagai berikut.

$$L_p = \sum_{i=1}^{n} \left[g_i \log \pi(z_i) + (1-g_i) \log\{1-\pi(z_i)\}\right] \tag{10}$$

dan

$$L_S = \sum \left\{g_i \log S_u(t_i|x_i) + \delta_i \log h_u(t_i|x_i)\right\} \tag{11}$$

dengan $h_u(.) = f_u(.)/S_u(.)$ fungsi hazard distribusi *failure time* pasien yang tak sembuh, g_i diberikan oleh:

$$g_i = \frac{\delta_i + (1-\delta_i)\pi(z_i) S_u(t_i|x_i)}{\left\{1-\pi(z_i) + \pi(z_i) S_u(t_i|x_i)\right\}} \tag{12}$$

yang merupakan peluang pasien yang tidak sembuh. Oleh karena itu, *E-step* algoritma *EM* memuat peluang g_i setiap pasien.

M-step algoritma EM mencakup *maximize* (10) dan (11) terkait dengan $f_u(.), S_u(.)$ dan γ untuk g_i yang *fixed*. Dalam hal ini, algoritma EM digunakan untuk estimasi maksimum *likelihood* distribusi *failure time* pasien yang tak sembuh dan γ dapat diperoleh secara terpisah karena (9) hanya tergantung pada γ sedangkan (11) hanya tergantung pada distribusi *failure time* pasien yang tak sembuh. Algoritma *EM* pada model *mixture*

umum secara mendetail dapat dilihat pada Larson dan Dinse (1985).

Persamaan (10) dapat dimaksimumkan dengan metode optimasi Newton-Raphson, sedangkan maksimasi (11) tergantung pada bagaimana distribusi *failure time* pasien yang tak sembuh.

Bohning (2000) mengembangkan algoritma *EM* untuk k-komponen (*fixed)*. Analisis didasarkan pada likelihood lengkap dalam model *mixture.* Fungsi log-likelihood lengkap didefinisikan sebagai berikut.

$$l_{com}(P) = \sum_i \sum_j Z_{ij} \log(p_j) + \sum_i \sum_j Z_{ij} \log f(x_i, \lambda_j) \quad (13)$$

dengan Z_{ij} komponen indikator (j =1,..,k)

Pada *E-step* ditentukan ekspektasi bersyarat Z_{ij} diketahui P dan X yaitu $E\lfloor Z\_ij \mid P, X \rfloor$

Sedangkan pada M-step dilakukan maksimasi:

$$E\left[l_{com}(P)\right] = \sum_i \sum_j e_{ij} \log(p_j) + \sum_i \sum_j e_{ij} \log f(x_i, \lambda_j) \quad (14)$$

Berdasar maksimasi tersebut diperoleh iterasi baru:

$$p_j^{(new)} = \sum_i e_{ij} / n \quad (15)$$

$$\lambda_j^{(new)} = \text{tergantung pada bentuk } f(x, \lambda) =$$

[sering = $\sum_i e_{ij} x_i / \sum_i e_{ij}$]

Seidel et.al (2000) menemukan solusi dari problem maksima ganda. Distribusi nol dari *LRS* tergantung pada pemilihan nilai inisial untuk algoritma *EM*. Langkah–langkah yang ditempuh sebagai berikut. Sebelum menggunakan algoritma *EM*, pilih k-komponen (*fixed)*, kemudian pilih sebarang nilai awal:

$$P = (\lambda_1, \lambda_2, \ldots, \lambda_k; p_1, p_2, \ldots, p_k) \text{ untuk}$$

algoritma *EM*. Langkah 1: Gunakan algoritma *EM* untuk iteras P_{EM}. Langkah 2: Tentukan γ_{max} untuk maksimasi $d(P_{EM}, \gamma)$dalam γ dan tentukan γ_{min} untuk minimize $d(P_{EM}, \gamma)$. Langkah 3: Gantikan γ_{max} dengan γ_{min} sehingga diperoleh:

$$P = P_{EM} + P_{EM}(\lambda_{\min})\left\{ Q_{\lambda_{\max}} - Q_{\lambda_{\min}} \right\} \quad (16)$$

Kemudian ulangi langkah 1

Missing Data

Missing data (data hilang) merupakan bagian pada hampir semua penelitian dari waktu ke waktu (Jamshidian 2005). Mengapa ada data hilang dalam suatu penelitian? Data hilang bisa terjadi karena suatu peralatan/perlengkapan tidak berfungsi, suatu keadaan yang menakutkan, orang yang sedang diteliti tiba-tiba sakit atau meninggal (hilang dari pengamatan) atau pemasukan data yang tidak benar.

Dalam kasus studi klinik ada data yang hilang karena pasien-pasien dengan waktu penyembuhan yang lama meninggalkan kota, berpindah rumah sakit, atau berhenti berobat dalam jangka waktu tertentu kemudian melanjutkan kembali proses perawatan dan berpotensi untuk sembuh.

Howell (2007) membagi data hilang dalam 3(tiga) kategori., yaitu *MCAR*, *MAR* dan *MNAR*. Data dikatakan hilang lengkap secara acak (*missing completely at random = MCAR*) jika *mean* dari peluang bahwa obeservasi (X_i) hilang tidak berkorelasi dengan (X_i) atau nilai sebarang variabel yang lain. Data dikatakan hilang secara random (*Missing at Random*), jika *missingness* tidak tergantung pada nilai X_i setelah dikontrol dengan variabel yang lain. Jika data hiang bukan *MCAR* dan juga bukan *MAR*, maka data hilang dikategorikan dalam *MNAR* (*Missing Not at Random*).

Algoritma EM untuk Analisis *Missing Data*

Ada beberapa metode yang dapat digunakan untuk analisis data hilang. Cohen dan Cohen (2003) mempopulerkan metode yang didasarkan pada variabel boneka untuk memberi kode observasi yang hilang (*missing observation*). Metode ini banyak digunakan pada ilmu perilaku (*behavioral sciences*). Cohen dan Cohen (2003) mengaplikasikan regresi ganda serta analisis korelasi pada ilmu perilaku juga.

Strategi yang digunakan sebagai berikut. Misalkan beberapa data hilang pada variabel *X*,

yang merupakan salah satu dari beberapa variabel independen dalam analisis regresi. Buat variabel boneka *D* yang bernilai 1 jika data hilang dan bernilai nol jika data tak hilang. Buat variabel X^* sedemikian hingga :

$$X^* = \begin{cases} X; & jika\ data\ tak\ hilang \\ c; & jika\ data\ hilang \end{cases} \quad (17)$$

dengan *c* adalah sebarang konstanta.

Regresi variabel tergantung *Y* pada X^* dan sebarang variabel lain dalam model dapat diperoleh. Subtitusi nlai *c* untuk data hilang tidak layak dipandang sebagai imputasi karena koefisien X^* invarian pada pemilihan konstanta *c.* Dengan demikian model hanya tergantung pada pemilihan *c* sebagai koefisien dari *D* yang merupakan indikator nilai hilang.

Jika koefisien *D* dapat diinterpretasikan sebagai nilai prediksi *Y* untuk individu–individu dengan data hilang pada *X* dikurangi nilai prediksi *Y* untuk individu pada mean dari *X*, sebagai kontrol untuk variabel dalam model. Koefisien untuk X^* dapat dipandang sebagai estimasi efek dari *X* diantara subgrup data yang dipunyai.

Teknik ini dikenal sebagai metode indikator untuk data hilang. Pada pendekatan ini, estimator tak bias untuk parameternya tidak dapat ditemukan .Meskipun demikian teknik ini dapat dikembangkan dengan mudah pada kasus lebih dari satu variabel independen dengan data hilang.

Pada kasus dengan beberapa variabel mempunyai data hilang, jika digunakan banyak vaiabel boneka untuk untuk data hilang maka variabel-variabel tersebut mempunyai korelasi yang tinggi (semacam kasus data hilang pada variabel ganda). Untuk prosedur secara umum dalam menentukan estimator tak bias dari koefisien genap jika data MCAR dikembangkan Jones (1996)

Pada metode tradisional dikenal model regresi dan analisis varians (Sethi dan Seligman, 1993). Ada beberapa alternatif lain yang dapat digunakan untuk menentukan solusi dari masalah terkait. Beberapa metode lain yang dikembangkan sebagai berikut.

1) Listwise or Casewise Deletion

Metode ini digunakan untuk analisis data hilang pada sebarang variabel. Metode ini cukup konservatif, sederhana dan hanya membutuhkan sedikit waktu untuk komputasinya. Jika datanya *MNAR* maka estimator dari parameternya bias. Dalam kondisi ini dilakukan reduksi varians dan nilai R^2 dinaikkan seperi mereduksi power kesalahan baku dan uji-t dari fungsi ukuran sampel. Metode ini jarang direkomendasikan.

2) Pairwise Deletion

Metode ini didasarkan pada matriks korelasi, dimana korelasi diantara setiap pasangan variabel dihitung dari semua kasus dengan data valid atau antara dua variabel. Metode ini digunakan pada *MAR* dan *MCAR*, dan memiliki power yang lebih baik dibandingkan dengan metode listwise

3) Mean Subtitution

Metode ini didasarkan pada penukaran semua data hilang dari variabel yang diberikan dengan nilai mean dari variabel tersebut. Metode ini baik digunakan untuk analisis *MAR* dan data berdistribusi normal. Metode ini lebih baik jika dibandingkan dengan metode *pairwise deletion*. Sebagaimana pada metode *listwise*, jika proporsi data hilang naik, maka varians dapat direduksi dan R^2 dinaikkan

4) Imputation by Regression

Prediksi data hilang didasarkan pada persamaan regresi yang menggunakan semua variabel lain yang relevan dengan prediktor. Pendekatan yang lebih baik adalah menggunakan informasi pada variabel lain melalui regresi ganda, yang sering disebut imputasi mean bersyarat.

5) Propensity Score Method

Metode ini dapat digunakan jika data hilang mempunyai pola monoton. Metode ini didasarkan pada peluang bersyarat yang terkait dengan bagian *treatment* yang memberikan vektor covariat observasi. Pada metode ini, setiap variabel dengan nilai hilang, *propensity score* dihasilkan untuk setiap observasi untuk estimasi peluang bahwa observasi hilang. Imputasi *bootstrap* dengan aproksimasi Bayesian diaplikasikan pada setiap grup. Metode ini hanya menggunakan informasi covariat yang berkaitan dengan imputasi nilai variabel yang hilang, dan tidak menggunakan korelasi diantara variabel. Ini efektif untuk inferensi distribusi dari variabel yang diimputasi secara individual, sebagaimana analisis univariat, tetapi tiadak untuk analisis hubungan antar variabel sebagaimana dalam regresi linier. Allison

(2000) menyajikan metode ini untuk imputasi covariat yang hilang dalam seting ini.

6) Hot Deck Imputation

Metode ini didasarkan pada penggantian setiap nilai data hilang dengan sebuah nilai yang dipilih secara acak pada kasus dengan data lengkap. Lebih spesifik, misal *Y* variabel dengan data hilang dan *X* tanpa data hilang. Imputasi nilai hilang untuk *Y* dilakukan dengan menentukan himpunan kategorikal variabel *X* yang berkaitan dengan *Y* dari tabel kontingensi berdasar variabel *X*. Jika terdapat kasus dengan nilai *Y* hilang dalam bagian sel dalam tabel, ambil satu atau lebih kasus nonmissing dalam sel yang sama dan gunakan nilai *Y* untuk imputasi nilai *Y* yang hilang

7) Raw Maximum likelihood or Full Information Maximum Likelihood (FIML) Method

Metode ini menggunakan semua informasi dari data terobservasi, termasuk mean dan varians yang didasarkan pada data untuk setiap variabel.

8) Multiple Imputations

Metode ini merupakan generalisasi metode maximum likelihood berdasar matriks kovarians dan vektor mean (Scheuren 2005: 315-319)

9) Expectation Maximization (EM) Algorithm

Metode ini didasarkan pada 2 (dua) langkah, yaitu: (1) *Epectation Step* untuk menentukan distribusi untuk missing data berdasar nilai observasi yang diketahui dan mengestimasi parameternya, dan (2) *Maximization Step*, yaitu subtitusi data hilang dengan nilai ekpektasi yang sudah diperoleh dari langkah (1)

Algoritma EM merupakan pendekatan iterasi untuk mempelajari model dari data dengan nilai hilang melalui 4 (empat) langkah sebagai berikut.

(1). Pilih himpunan inisial dari parameter untuk sebuah model
(2). Tentukan nilai ekspektasi untuk data hilang
(3). Buat induksi parameter model baru dari gabungan nilai ekspekstasi dan data asli
(4). Jika parameter tidak converged, ulangi langkah 2 menggunakan model baru

Jika statistik cukup dari data lengkap merupakan fungsi linier dari data, maka algoritmanya dapat dipecahkan dengan sederhana. Dalam *E-step*, observasi dengan *missing data*, ekspektasi bersyarat dari data observasi yang diberikan ditukarkan dengan vektor parameter. Pada *M-step*, perbedaan dibuat antara data observasi dan data yang berkaitan.

Secara umum, implementasi algoritma *EM* dalam analisis dengan missing data sebagai berikut. Laju kekonvergenan algoritma *EM* sering sangat lamban (Dempster *et al.* 1977). Tidak seperti metode *least square* (*weighted*), pada algoritma *EM*, estimasi varians parameternya tidak dihasilkan. Jika observasi mempunyai *covariat mising*, maka asumsi–asumsi diperlukan untuk meringkas informasi dari observasi bagian.

Model Mixture untuk Analisis *Missing data*

Missing data merupakan bagian dari data tak lengkap (*incomplete data*), pada formulasi data tak lengkap, model mixture dapat dijelaskan sebagai berikut. Misalkan X ruang sample data lengkap dari x, Y ruang sample data terobservasi dan Z *hidden sample space*. Dipunyai $X = Y \times Z$ dan $x = (y, z)$. Densitas dari data terobservasi X dapat ditulis sebagai berikut (Picard 2007).

$$g(x,\theta) = f(y;\theta)h(z \mid y;\theta) \qquad (18)$$

Dengan $f(y^{\cdot},\theta)$ densitas data terobservasi dan $h(z|y^{\cdot},\theta)$ densitas bersyarat dari *missing observation*.

Misal $L(y^{\cdot},\theta)$fungsi likelihood dari *incomplete-data*/terobservasi dan $L^{c}(x^{\cdot},\theta)$fungsi likelihood dari *complete-data*/ tak terobservasi, maka:

$$\log L^{C}(x;\theta) = \log L(y;\theta) + \log h(z \mid y;\theta) \qquad (19)$$

dengan

$$\log L^{C}(x;\theta) = \sum_{i=1}^{n} \log g(x_i;\theta) \qquad (20)$$

dan

$$\log h(z \mid y;\theta) = \sum_{t=1}^{n}\sum_{p=1}^{P} z_{tp} \log E\{Z_{tp} \mid Y_t = y_t\} \qquad (21)$$

Variabel *hidden* tidak terobservasi, maka *EM* memuat *indirect optimization* dari *incomplete-data likelihood* melalui optimasi iterasi ekspektasi bersyarat *complete-data likelihood* menggunakan current fit untuk θ. Misal $\theta(h)$ nilai parameter pada iterasi h, maka:

$$\log L(y;\theta) = Q\left(\theta;\theta^{(h)}\right) - H\left(\theta;\theta^{(h)}\right) \qquad (22)$$

dengan perjanjian:

$$Q\left(\theta;\theta^{(h)}\right)= E_{\theta^{(h)}}\left\{\log L^{C}\left(X;\theta\right)\mid Y\right\} \quad (23)$$

$$H\left(\theta;\theta^{(h)}\right)= E_{\theta^{(h)}}\left\{\log h\left(Z\mid Y;\theta\right)\mid Y\right\} \quad (24)$$

Dengan $E_{\theta(h)}\{.\}$ menyatakan operator ekspektasi, current fit $\theta(h)$ untuk θ .

Pada hakekatnya algoritma *EM* memuat 2 (dua) langkah , (1) *E-step* menghitung $Q(\theta;\theta^{(h)})$ dan (2) M-step memilih $\theta^{(h+1)} = \underset{\theta}{Arg\max}\left\{Q\left(\theta;\theta^{(h)}\right)\right\}$. *E-steps* and *M-steps* diulang sampai mencapai nilai $\left|\theta^{(h+1)}-\theta^{(h)}\right|$ yang kecil.

Teorema:

Pada algoritma *EM*, *log-likelihood* untuk *incomplete data* mengalami kenaikan setelah dilakukan setiap iterasi dari algoritmanya

Bukti:

Bukti teorema ini didasarkan pada definisi *M-step*

$$Q\left(\theta,\theta^{(h+1)}\right)\geq Q\left(\theta,\theta^{(h)}\right)$$

Sementara aplikasi ketaksamaan Jensen memberikan:

$$H\left(\theta,\theta^{(h+1)}\right)\geq H\left(\theta,\theta^{(h)}\right)$$

Berdasarkan (21) diperoleh:

$$\log\ L\left(y,\theta^{(h+1)}\right)\leq \log\ L\left(y,\theta^{(h)}\right)$$

Ketaksamaan di atas memberikan bukti bahwa pada algoritma *EM*, barisan likelihood konvergen jika likelihood terbatas ke bawah.

PENUTUP

Pada hakekatnya Algoritma *EM* yang telah dipaparkan merupakan komputasi iterasi dari maximum likelihood estimators (*MLE*) jika observasi dikategorikan sebagai *incomplete data*. Ide dasar algoritma *EM* adalah mengkaitkan model data data lengkap dengan model *incomplete data* dengan komputasi seperti pada *MLE*

Berdasar kajian yang dilakukan dapat ditunjukkan bahwa pada algoritma *EM*, *log-likelihood* untuk *missing data* mengalami kenaikan setelah dilakukan setiap iterasi dari algoritmanya. Dengan demikian berdasar algoritma *EM*, barisan likelihood konvergen jika likelihood terbatas ke bawah.

DAFTAR PUSTAKA

Boag JW. 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J Royal Stat Soc.* 11: 15-53

Bohning D. 2000. *The Potential of Recent Developments in Nonparametric Mixture Distribution.* Euro-workshop on Statistical Modelling

Cantor AB & Shuster JJ. 1992. Perametric Versus Noparametric Methods for Cure Rates Based on Censored Survival Data*. Statistics in Medicine*. 11: 931-937

Cohen J & Cohen P. 1983. *Applied multiple regression/correlation analysis for the Behavioral Sciences, 2nd edition.* Hillsdale, NJ: Erlbaum

Cohen J & Cohen P, West SG & Aiken LS. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition.* Mahwah, N.J Lawrence Erlbaum

Dempster AP, Laird NM & Rubin DB. 1977. Maximum Likelihood from incomplete data via the EM algorithm (with discussion). *J Royal Stat Soc*. Series B. 39: 1-38

Kazemi I. 2005. *Methods for Missing Data.*Center for Applied Statistcs: Lancastar University

Farewell VT. 1982. The use ofmixture models for analysis of survival data with long-term survivors. *Biometrics* 38, 1041 - 1046.

Farewell VT. 1986. Mixture models in survival analysis. Are they worth the risk? *Can J Stat* 14: 257-262.

Ghitany ME, Maller RA and Zhou, S. 1994. Exponential mixture models with long-term survivor and covariates. *J Multivariate Analysis* 49, 218-241.

Howell D. 2007. *Treatment of Missing Data.* David.Howell's Statistical Home Page

Jamshidian M. 2005. *Statistician Works to Develop Method to Deal With Missing Data.* mori@fulleron.edu

Jones DR, Powles RL, Machin D & Sylvester RJ. 1981. On Estimating the proportion of cured patients in clinical studies. *Biometrie-Praximetrie* 21: 1-11.

Kuk AY & Chen C. 1992. A mixture model combining logistic regression and life model. *Biometrika* 79: 531-541.

Larson MG & Dinse GE. 1985. A Mixture Model for The Regression Analysis of Computing Risk Data*. Applied Statistics*. 34: 201-211

Picard F. 2007. *An Introduction to Mixture Models.* Statistics for Systems Biology Group. Research Report No.7

Scheuren F. 2005. Multiple Imputation: How it began and continue. *Am Stat*, 59: 315-319

Sethi S & Seligman MEP. 1993. Optimism and fundamentalism. *Psychol Sci,* 4: 256-259.

Taylor JMG. 1995. Semi-parametric Estimation in Failure-Time Mixture Models*. Biometrics*. 51: 899-907

Yamaguchi K. 1992. Accelerated Failure-Time Regression Models with a Regression models of Surviving Fraction. An application to analysis of Permanent Employment in Japan*. J Am Stat Assoc*. 37: 284-292