Jurnal Sains Matematika dan Statistika, Vol. 3, No. 1, Januari 2017
ISSN 1693-2390 print/ISSN 2407-0939 online

Teknik Mengatasi Data Hilang dengan Metode Algoritma EM

Juliana Sari[1], Rado Yendra[2]
[1, 2] Jurusan Matematika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau
Jl. HR. Soebrantas No. 155 Simpang Baru, Panam, Pekanbaru, 28293
Email: sari.julianna95@gmail.com, yendra75@yahoo.com.sg.

ABSTRAK

Data hilang merupakan informasi yang tidak tersedia untuk sebuah kasus tertentu. Salah satu metode yang digunakan untuk mengatasi data hilang adalah Algoritma EM. Penelitian ini bertujuan untuk mendapatkan teknik mengatasi data hilang dengan metode Algoritma EM. Metode Algoritma EM merupakan sebuah metode optimisasi iteratif yang terbagi atas dua tahapan yaitu tahap ekspektasi dan tahap maksimisasi. Data yang digunakan pada penelitian ini adalah data matriks $4x3$ dengan 3 data hilang. Hasil pada penelitian ini diperoleh nilai akhir sampai 7 iterasi, selanjutnya analisis uji χ^2 diperoleh $F_{hitung} < F_{tabel}$ dengan nilai $-11.7318 \leq 0.71$ maka dapat disimpulkan untuk terima H_0 dengan interpretasi tidak terdapat perbedaan rata-rata nilai awal antara sebelum penambahan data hilang dan sesudah penambahan data hilang.

Kata Kunci: *Algoritma EM, Data Hilang,Matriks, Uji* χ^2.

ABSTRACT

Missing data is information that is not available for a particular case. One of the methods used to solve the missing data is the EM Algorithm. This study aims to obtain techniques to solve lost data by EM Algorithm method. The EM Algorithm Method is an iterative optimization method which is divided into two stages: the expectation stage and the maximization stage. The data used in this research is matrix $4x3$ *data with 3 missing data. The results of this study obtained the final value of up to 7 iterations, then the test analysis* χ^2 *obtained* $F_{hitung} < F_{tabel}$ *with the value* $-11.7318 \leq 0.71$ *it can be concluded to receive* H_0 *by interpretation there is no difference in the average initial value between before the addition of missing data and after the addition of missing data.*

Keywords: EM Algorithm, Missing Data, Matrix, Test. χ^2.

Pendahuluan

Permasalahan data hilang pertama kali diperkenalkan oleh Orchard dan Woodbury pada tahun 1972. Data hilang dapat disebabkan oleh beberapa hal yang tidak bisa diramalkan dan tidak bisa dihindari. Misalnya pengukuran yang mungkin tidak lengkap, kesalahan yang terjadi pada prosedur pengumpulan data atau karena responden menolak untuk menjawab beberapa pertanyaan tertentu dalam survei atau karena munculnya hipotesis baru yang menarik setelah pengumpulan data dilakukan. Little dan Rubin (1987) memperkenalkan berbagai macam metode untuk mengatasi data hilang (*missing data*), diantaranya adalah: *complete case analysis.*

Penelitian data hilangsudahbanyakdilakukanolehpeneliti yang terdahuludiantaranya, S. Zacksdari N.Y, USA danJosemar Rodrigues dariBzil The University of Sao Paulo (1985) dalampenelitian*A Note On The Missing Value Principle And The EM-Algoritm For estimation And*

Prediction In Sampling From Finite Populations With AMultinormalSuperpopulation Model, Donald B. Rubin (1976) dalampenelitian*Inference and Missing Data*, A. P. Dempster; N. M. Laird; D. B. Rubin (1977) dalamPenelitian*Maximum Likelihood from incomplete data via the EM Algorithm*.

Metode dan Bahan Penelitian

1. Algoritma EM

Algoritma EM adalah sebuah metode optimisasi iteratif untuk estimasi Maksimum Likelihood (ML) yang berguna dalam permasalahan data yang tidak lengkap (*incomplete data*). Kasus khusus dimana algoritma EM digunakan untuk memprediksi rata-rata populasi dan varians tidak diketahui dan harus diperkirakan mempunyai tahap Ekspektasi *(Expectation Step)* dan tahap Maksimisasi *(Maximization Step).*

1.1 Tahap Ekspektasi atau *Expectation Step* (E Step)

Tahapan-tahapan ekspektasi data hilang denganAlgoritma EM adalah :

a. Hitung nilai parameter dari data yang ada.

$$\tilde{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1,2,...p \tag{1}$$

$$\tilde{\sigma}_k = \tilde{\sigma}_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2 \quad k = 1,2,...p \tag{2}$$

$$\tilde{\sigma}_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) i = 1,2,...p \; k = 1,2,...p \tag{3}$$

dengan

$\tilde{\mu}$ = rata-rata mean

$\tilde{\sigma}_k$ = varians

$\tilde{\sigma}_{ik}$ = kovarians

b. Masukkan ke persamaan

Untuk setiap $x_j^{(1)}$ adalah komponen yang hilang, dan $x_j^{(2)}$ adalah komponen yang ada. Untuk memprediksi $\tilde{\mu}$ dan $\tilde{\Sigma}$ digunakan mean distribusi bersyarat $x^{(1)}$ dan diberikan $x^{(2)}$ untuk menduga nilai yang hilang. Sehingga:

$$\begin{aligned}\tilde{x}_j^{(1)} &= E\left(X_j^{(1)} \middle| x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}\right) \\ &= \tilde{\mu}^{(1)} + \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\left(x_j^{(2)} - \tilde{\mu}^{(2)}\right)\end{aligned} \tag{4}$$

Memprediksi kontribusi $x_j^{(1)}$ untuk T_1 :

$$\begin{aligned}\overline{x_j^{(1)} x_j^{(1)}} &= E\left(X_j^{(1)} X_j^{(1)'} \middle| x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}\right) \\ &= \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21} + \tilde{X}_j^{(1)}\tilde{X}_j^{(1)'}\end{aligned} \tag{5}$$

$$\begin{aligned}\overline{x_j^{(1)} x_j^{(2)}} &= E\left(X_j^{(1)} X_j^{(2)'} \middle| x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}\right) \\ &= \tilde{x}_j^{(1)}\tilde{x}_j^{(2)'}\end{aligned} \tag{6}$$

Memprediksi kontribusi $x_j^{(1)}$ untuk T_2 :

Kontribusi pertama dijumlahkan untuk setiap x_j dengan komponen yang hilang. Hasil ini digabungkan dengan data sampel menghasilkan T_1 dan T_2 Menentukan matriks T_1 dan T_2 menggunakan rumus :

$$\tilde{T}_1 = \begin{bmatrix} \tilde{x}_{11} + x_{21} + x_{31} + \tilde{x}_{41} \\ x_{12} + x_{22} + x_{32} + \tilde{x}_{42} \\ x_{13} + x_{23} + x_{33} + x_{43} \end{bmatrix} \tag{7}$$

$$\tilde{T}_2 = \begin{bmatrix} \overline{x_{11}^2} + x_{21}^2 + x_{31}^2 + \overline{x_{41}^2} & & \\ \overline{x_{11}x_{12}} + x_{21}x_{22} + x_{31}x_{32} + \overline{x_{41}x_{42}} & x_{12}^2 + x_{22}^2 + x_{32}^2 + \overline{x_{42}^2} & \\ \overline{x_{11}x_{13}} + x_{21}x_{23} + x_{31}x_{33} + \overline{x_{41}x_{43}} & x_{12}x_{13} + x_{22}x_{23} + x_{32}x_{33} + \overline{x_{42}x_{43}} & \\ & & x_{13}^2 + x_{23}^2 + x_{33}^2 + x_{43}^2 \end{bmatrix} \tag{8}$$

1.2 Tahap Maksimisasi atau *Maximization Step* (M Step)

$$\tilde{\mu} = \frac{\tilde{T}_1}{n} \tag{9}$$

$$\tilde{\Sigma} = \frac{1}{n}\tilde{T}_2 - \tilde{\mu}\tilde{\mu}' \tag{10}$$

2. Uji χ^2

Uji χ^2 adalah pengujian hipotesis mengenai perbandingan antara frekuensi observasi / yang benar-benar terjadi dengan frekuensi harapan / ekspektasi. Nilai χ^2 adalah nilai kuadrat. Oleh karena itu nilai χ^2 selalu positif. Uji χ^2 digunakan untuk menunjukkan apakah ada pengaruh data hilang terhadap nilai awal dalam sebuah data dan kemudian dibandingkan dengan nilai hitung dengan rumus :

$$n(\tilde{\mu} - \mu)' \tilde{\Sigma}^{-1}(\tilde{\mu} - \mu) \leq \chi_p^2(\alpha) \tag{11}$$

Hasil dan Pembahasan

Dalam menyelesaikan teknik data hilang dengan menggunakan metode algoritma EM terdiri dari beberapa langkah : mendapatkan rata-rata sampel awal dari data yang tidak lengkap, ganti rata-rata sampel awal untuk memperoleh perkiraan varians dan kovarians awal, gunakan nilai awal $\tilde{\mu}$ dan $\tilde{\Sigma}$ untuk memprediksi nilai yang hilang,prediksi komponen yang hilang pada x_1 dengan mempartisi nilai awal $\tilde{\mu}$ dan $\tilde{\Sigma}$, substitusikan nilai prediksi komponen yang hilang pada x_1 dan x_4 terhadap T_1 dan T_2 selanjutnya, langkah estimasi dengan menstubtitusikan hasil-hasil kedalam persamaan (2.17) dan (2.18), diperoleh nilai $\tilde{\mu}$ dan $\tilde{\Sigma}$ dari langkah estimasi, periksa apakah nilai $\tilde{\mu}$ dan $\tilde{\Sigma}$ sudah konvergen. Jika belum, lakukan iterasi sampai nilai $\tilde{\mu}$ dan $\tilde{\Sigma}$ konvergen. Lakukan hipotesis dan kesimpulan.
Sebagai contoh diberikan data dengan 3 data hilang

$$X = \begin{bmatrix} - & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ - & - & 5 \end{bmatrix}$$

Hasil dari penyelesaian contoh diatas dengan menggunakan metode Algoritma EM diperoleh bahwa $\tilde{\sigma}_{11} = 0.60$ dan $\tilde{\sigma}_{22} = 0.60$ lebih besar dari estimasi iterasi ketujuh observasi yang hilang dan diperoleh nilai elemen-elemen $\tilde{\mu}$ dan $\tilde{\Sigma}$ sudah konvergen. Oleh karena elemen- elemen $\tilde{\mu}$ dan $\tilde{\Sigma}$ sudah konvergen, maka iterasi berhenti pada iterasi ketujuh.
Selanjutnya, untuk melihat penyelesaian masalah data hilang dengan metode Algoritma EM menggunakan program MAPLE dapat dilihat pada Lampiran.

Untuk melihat apakah data yang hilang berpengaruh atau tidak terhadap nilai awal maka dilakukan uji χ^2.

1. Hipotesis
   $H_0 : \mu = \tilde{\mu}$
   (data hilang tidak berpengaruh terhadap nilai awal)
   $H1 : \mu \neq \tilde{\mu}$
   (data hilang berpengaruh terhadap nilai awal)
2. Taraf Signifikan
   $\alpha = 5\%$.
3. Statistik Uji
   menggunakan uji χ^2
4. Statistik Hitung
   Diperoleh dengan metode Algoritma EM dan disubstitusikan kedalam persamaan :

$$n(\tilde{\mu} - \mu)' \tilde{\Sigma}^{-1} (\tilde{\mu} - \mu) \leq \chi_p^2(\alpha)$$

$$4\left(\begin{bmatrix} 6.06 \\ 1.12 \\ 4.00 \end{bmatrix} - \begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix}\right)' \begin{bmatrix} 0.60 & 0.38 & 1.22 \\ 0.38 & 0.60 & 0.87 \\ 1.22 & 0.87 & 2.50 \end{bmatrix}' \left(\begin{bmatrix} 6.06 \\ 1.12 \\ 4 \end{bmatrix} - \begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix}\right) \leq \chi_p^2(0.05)$$

$$4\left(\begin{bmatrix} 0.06 \\ 0.12 \\ 0 \end{bmatrix}\right)' \begin{bmatrix} 0.60 & 0.38 & 1.22 \\ 0.38 & 0.60 & 0.87 \\ 1.22 & 0.87 & 2.50 \end{bmatrix} \left(\begin{bmatrix} 0.06 \\ 0.12 \\ 0 \end{bmatrix}\right) \leq \chi_3^2(0.05)$$

$$-11.7318 \leq 0.71$$

5. Kesimpulan
   Jadi didapat $F_{hitung} < F_{tabel}$ maka dapat disimpulkan untuk terima H_0 dengan interpretasi tidak terdapat perbedaan rata-rata antara sebelum penambahan data hilang dan sesudah penambahan data hilang.

Kesimpulan

Hasil penyelesaian masalah data hilang menggunakan metode Algoritma EM diperoleh nilai $\tilde{\mu}$ dan $\tilde{\sum}$ konvergen pada iterasi ketujuh. Selanjutnya analisis uji χ^2 yang dilakukan, diperoleh $F_{hitung} < F_{tabel}$ dengan nilai $-11.7318 \leq 0.71$ maka dapat disimpulkan untuk terima H_0 dengan interpretasi tidak terdapat perbedaan rata-rata nilai awal antara sebelum penambahan data hilang dan sesudah penambahan data hilang.

Daftar Pustaka

[1] Assauri, Sofjan. *"Aljabar Linear DasarEkonometri".*Edisikedua, halaman 40.Penerbit : CV. Rajawali, Jakarta. 1983.
[2] Dempster, A. P, N. M. Laird, D. B. Rubin, "Maximum Likelihood From Incomplete Data Via The EM Algorithm". *Journal Of The Royal Statistical Society. Series B 39:1-38. 1977.*
[3] Fatimah, Imas. *"Data Hilang DalamR ancangan Percobaan".* Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam. Bogor.2003.
[4] Jhonson, Richard A, &Wichern, Dean W. *"Applied Multivariate Statistical Analysis"* Edisikesembilan.Amerika. 2007.
[5] Little, Roderick, J. A & Rubin, Donal B. *"Statistical Analysis With Missing Data".* California. 1987.
[6] Pudjiastuti, BSW. *"Matriks :TeoridanAplikasi".*Penerbit :GrahaIlmu,Yogyakarta. 2006.
[7] Susila,INyoman. *"Matriks :TeoridanSoal-Soal".*Penerbit :Erlangga, Jakarta. 1984.
[8] Sutojo, Bowo, dkk. *"TeoridanAplikasiAljabar Linier danMatriks".*Penerbit :Andi, Yogyakarta. 2010.