EM Algorithm for Missing Data

EM Imputation Examples

STAT 32950-24620

Spring 2023 (5/4)

Missing Data Problem

A simple example

There are $n = 3$ observations on bivariate $X = [X_1 \ X_2]'$.

$$\begin{array}{c} \\ obs_1 \\ obs_2 \\ obs_3 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} ? & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \end{array}$$

There is a missing value of component variable X_1 in the first observation.

Objective: Find a good imputation value for x_{11}.

Example 1 of multivariate normal missing data

Assuming the observations are independent.
Assuming the observations are from the same distribution. (i.i.d.)

Assuming the missing x_{11} of variable X_1 is **missing at random**.

A reasonable candidate of estimate:
— the sample average of all observed values of variable X_1.

$\Rightarrow$ Initial estimation:

$$\tilde{x}_{11} = 3$$

Under the initial estimate,

$$\begin{bmatrix} ? & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} 3 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix}$$

Initialization: Impute with variable mean

$$X \sim \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim (\mu, \Sigma)$$

Consider the case that the component variables X_1 and X_2 are correlated.

Σ is not a diagonal matrix.

The estimated μ and Σ should give us information whether the imputed value is appropriate.

To find a better estimate taking the correlation between variables X_1 and X_2 into consideration, we need to estimate the parameters under the current estimate.

From initial impute to parameter estimates

Assume the data is from a bivariate normal $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N(\mu, \Sigma)$.

The parameters are in μ and Σ.

Recall the maximum likelihood estimation for multivariate normal:

$$\hat{\mu} = \bar{X}$$

$$\hat{\Sigma}_{ML} = \frac{n-1}{n} S$$

Parameter estimates after the initial impute

Use the data imputed with initial value to obtain the initial ML estimates of μ and Σ:

$$\tilde{\mu} = \begin{bmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 14/3 \end{bmatrix} = \begin{bmatrix} 3.00 \\ 4.67 \end{bmatrix}$$

$$\tilde{S} = \begin{bmatrix} 1 & 1 \\ 1 & 6.33 \end{bmatrix}, \qquad \tilde{\Sigma}_{ML} = \left(\frac{2}{3}\right) \tilde{S} = \begin{bmatrix} 0.67 & 0.67 \\ 0.67 & 4.22 \end{bmatrix}$$

So the initial estimate of the parameters and thus the joint distribution of X_1, X_2 are

$$\begin{bmatrix} \tilde{X}_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{bmatrix}, \begin{bmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{bmatrix}\right) = N\left(\begin{bmatrix} 3 \\ 4.67 \end{bmatrix}, \begin{bmatrix} 0.67 & 0.67 \\ 0.67 & 4.22 \end{bmatrix}\right)$$

E-step, new impute

The EM iteration process uses the conditional distribution

$$X_1|X_2 = x_2 \ \sim \ N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

The next step (the first estimate in the iteration process) of x_{11} uses the Expected (E) value of the conditional distribution.

$$\begin{aligned} \tilde{x}'_{11} &= E[X_{11}|X_2 = x_{12}] = E[X_1|X_2 = 2] \\ &= \tilde{\mu}_1 + \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}(2 - \tilde{\mu}_2) \\ &= 3 + 0.67 \times 4.22^{-1}(2 - 4.67) = 2.58 \end{aligned}$$

$$\begin{bmatrix} ? & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 3 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 2.58 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix}$$

M-step, new estimated parameters

To find the next estimate by the second estimation, we need to know the parameter estimations under the current estimate.

The current (first iteration) estimate gives the estimates of the following Maximum (M) Likelihood estimates of the parameters:

$$\tilde{\mu}' = \begin{bmatrix} \tilde{\mu}'_1 \\ \tilde{\mu}'_2 \end{bmatrix} = \begin{bmatrix} 2.86 \\ 4.67 \end{bmatrix}$$

$$\tilde{\Sigma}'_{ML} = \left(\frac{2}{3}\right) \tilde{S}' = \begin{bmatrix} 0.70 & 1.04 \\ 1.04 & 4.22 \end{bmatrix}$$

First EM iteration results

For the mean parameter

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

The estimation process up to the first iteration proceeds as

$$\begin{bmatrix} ? \\ \mu_2 \end{bmatrix} = \begin{bmatrix} ? \\ 4.67 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 3 \\ 4.67 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 2.86 \\ 4.67 \end{bmatrix}$$

For the covariance matrix

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

the estimation process up to the first iteration proceeds as

$$\begin{bmatrix} ? & ? \\ ? & \Sigma_{22} \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 0.67 & 0.67 \\ 0.67 & 4.22 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 0.70 & 1.04 \\ 1.04 & 4.22 \end{bmatrix}$$

Imputed data sequence

$$\begin{bmatrix} ? & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 3 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 2.58 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix}$$

$$\xrightarrow{\text{second iteration}} \begin{bmatrix} 2.20 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{third iteration}} \begin{bmatrix} 1.87 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix}$$

$$\vdots$$

$$\xrightarrow{\text{40th iteration}} \begin{bmatrix} -0.948 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix} \xrightarrow{\text{41st iteration}} \begin{bmatrix} -0.953 & 2 \\ 2 & 5 \\ 4 & 7 \end{bmatrix}$$

Imputed data sequence stopping criteria

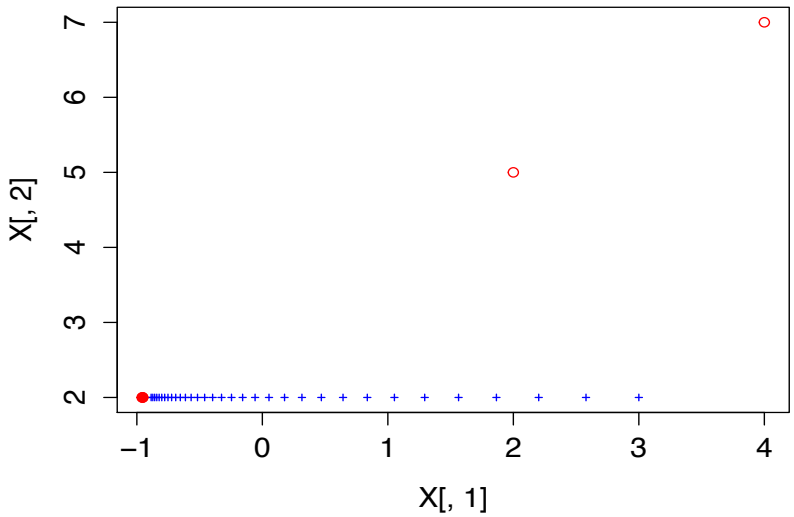
Set the precision threshold $\delta = 0.005$.

Since

$$\left| \tilde{x}_{11}^{(42)} - \tilde{x}_{11}^{(41)} \right| < \delta$$

The process stops after 41 iterations.

EM Imputation for the missing observation



Example 2: $n = 4$ $p = 3$

4 observations from $X = [X_1\ X_2\ X_3]' \sim N_3(\mu, \Sigma)$.

$$\begin{array}{c} \\ obs_1 \\ obs_2 \\ obs_3 \\ obs_4 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} ? & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ ? & ? & 5 \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} \end{array}$$

Objective: Find good imputations for the value of x_{11} in observation 1, and for the values of x_{41}, x_{42} in observation 4.

Initial imputes

Initial value of estimation:

$$\tilde{x}_{11} = 6$$

$$\tilde{x}_{41} = 6, \quad \tilde{x}_{42} = 1.$$

Under the initial values,

$$\begin{bmatrix} ? & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ ? & ? & 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 6 & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ 6 & 1 & 5 \end{bmatrix}$$

Parameter estimates from Initial imputes

Use the initial values to obtain the initial ML estimates of μ and Σ:

$$\tilde{\mu} = \begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix}$$

$$\tilde{\Sigma}_{ML} = \left(\frac{3}{4}\right) \tilde{S} = \begin{bmatrix} 0.50 & 0.25 & 1.00 \\ 0.25 & 0.50 & 0.75 \\ 1.00 & 0.75 & 2.50 \end{bmatrix}$$

So the initial estimate of the parameters are obtained, and the joint distribution of $(X_1, X_2, X_3)'$ is

$$\sim N\left(\begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.50 & 0.25 & 1.00 \\ 0.25 & 0.50 & 0.75 \\ 1.00 & 0.75 & 2.50 \end{bmatrix} \right)$$

Imputing missing value in observation 1

To impute the missing x_{11} of variable X_1 in observation 1, we partition $(X_1, X_2, X_3)'$ as $(X_{(1)}, X_{(2)})'$, where $X_{(1)}$ is the missing variable X_1.

$$\left[\begin{array}{c} X_1 \\ \hline X_2 \\ X_3 \end{array}\right] = \left[\begin{array}{c} X_{(1)} \\ \hline X_{(2)} \end{array}\right]$$

Correspondingly,

$$\left[\begin{array}{c|cc} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \hline \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{array}\right] = \left[\begin{array}{c|c} \Sigma_{(11)} & \Sigma_{(12)} \\ \hline \Sigma_{(21)} & \Sigma_{(22)} \end{array}\right]$$

Under the normality assumption,

$$\begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{bmatrix} \right)$$

The corresponding conditional distribution is

$$X_{(1)}|X_{(2)} = x_{(2)} \sim N\left(\mu_{(1)} + \Sigma_{(12)}\Sigma_{(22)}^{-1}\left(x_{(2)} - \mu_{(2)}\right), \Sigma_{(11)} - \Sigma_{(12)}\Sigma_{(22)}^{-1}\Sigma_{(21)}\right)$$

Using the estimates based on the initial values of the mean

$$\begin{bmatrix} \tilde{\mu}_{(1)} \\ \hline \tilde{\mu}_{(2)} \end{bmatrix} = \begin{bmatrix} 6 \\ \hline 1 \\ 4 \end{bmatrix}$$

and covariance matrix estimates

$$\begin{bmatrix} \tilde{\Sigma}_{(11)} & \tilde{\Sigma}_{(12)} \\ \hline \tilde{\Sigma}_{(21)} & \tilde{\Sigma}_{(22)} \end{bmatrix} = \left[\begin{array}{c|cc} 0.50 & 0.25 & 1.00 \\ \hline 0.25 & 0.50 & 0.75 \\ 1.00 & 0.75 & 2.50 \end{array}\right]$$

Using the parameter estimates based on the initial values, we obtain the estimates for x_{11} after the first iteration:

$$\begin{aligned} \tilde{x}'_{11} &= E\left[\tilde{X}_{(1)}|X_{(2)} = x_{(2)}\right] \\ &= E\left[\tilde{X}_{(1)} \mid X_{(2)} = (0,3)'\right] \\ &= \tilde{\mu}_{(1)} + \tilde{\Sigma}_{(12)}\tilde{\Sigma}_{(22)}^{-1}\left(x_{(2)} - \mu_{(2)}\right) \\ &= 6 + [0.25\ 1]\begin{bmatrix} 0.5 & 0.75 \\ 0.75 & 2.5 \end{bmatrix}^{-1}\left(\begin{bmatrix} 0 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 4 \end{bmatrix}\right) \\ &= 5.73 \end{aligned}$$

Imputing 2 missing values in observation 4

To impute the missing x_{41}, x_{42} of variable X_1, X_2 in observation 4, we partition $(X_1, X_2, X_3)'$ as $(X_{(1)}, X_{(2)})'$, where $X_{(1)}$ represents the missing vector $(X_1, X_2)'$.

$$\begin{bmatrix} X_1 \\ X_2 \\ \hline X_3 \end{bmatrix} = \begin{bmatrix} X_{(1)} \\ \hline X_{(2)} \end{bmatrix}$$

Correspondingly,

$$\left[\begin{array}{cc|c} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \hline \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{array}\right] = \begin{bmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \hline \Sigma_{(21)} & \Sigma_{(22)} \end{bmatrix}$$

Under the normality assumption,

$$\begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{bmatrix}\right)$$

The corresponding conditional distribution is

$$X_{(1)}|X_{(2)} = x_{(2)} \sim N\left(\mu_{(1)} + \Sigma_{(12)}\Sigma_{(22)}^{-1}\left(x_{(2)} - \mu_{(2)}\right), \Sigma_{(11)} - \Sigma_{(12)}\Sigma_{(22)}^{-1}\Sigma_{(21)}\right)$$

Using the estimates based on the initial values of the mean

$$\begin{bmatrix} \tilde{\mu}_{(1)} \\ \hline \tilde{\mu}_{(2)} \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ \hline 4 \end{bmatrix}$$

and covariance matrix estimates

$$\begin{bmatrix} \tilde{\Sigma}_{(11)} & \tilde{\Sigma}_{(12)} \\ \hline \tilde{\Sigma}_{(21)} & \tilde{\Sigma}_{(22)} \end{bmatrix} = \left[\begin{array}{cc|c} 0.50 & 0.25 & 1.00 \\ 0.25 & 0.50 & 0.75 \\ \hline 1.00 & 0.75 & 2.50 \end{array}\right]$$

we obtain the estimates for $(x_{41}, x_{42})'$ after the first iteration:

$$\begin{aligned}
\tilde{x}'_{11} &= E\left[\tilde{X}_{(1)} | X_{(2)} = x_{(2)}\right] \\
&= E\left[\tilde{X}_{(1)} \mid X_{(2)} = 5\right] \\
&= \tilde{\mu}_{(1)} + \tilde{\Sigma}_{(12)}\tilde{\Sigma}_{(22)}^{-1}\left(x_{(2)} - \mu_{(2)}\right) \\
&= \begin{bmatrix} 6 \\ 1 \end{bmatrix} + \begin{bmatrix} 1.00 \\ 0.75 \end{bmatrix} (2.5)^{-1}(5-4) \\
&= \begin{bmatrix} 6.4 \\ 1.3 \end{bmatrix}
\end{aligned}$$

The imputed data after the first iteration is

$$\begin{bmatrix} 5.73 & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ 6.4 & 1.3 & 5 \end{bmatrix}$$

from which we obtain the parameter estimates.

$$\begin{bmatrix} \tilde{\mu}'_1 \\ \tilde{\mu}'_2 \\ \tilde{\mu}_3 \end{bmatrix} = \begin{bmatrix} 6.03 \\ 1.08 \\ 4 \end{bmatrix}$$

and covariance matrix estimates

$$\tilde{\Sigma}' = \begin{bmatrix} 0.558 & 0.346 & 1.168 \\ 0.346 & 0.517 & 0.825 \\ 1.168 & 0.825 & 2.50 \end{bmatrix}$$

Up to iteration 1 — Data imputation

$$\begin{bmatrix} ? & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ ? & ? & 5 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 6 & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ 6 & 1 & 5 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 5.73 & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ 6.4 & 1.3 & 5 \end{bmatrix}$$

Notice that the first iteration estimations are obtained via two observation-specific partitions of the mean vector and covariance matrix, and by two observation-specific imputations based on the corresponding conditional distributions.

Up to iteration 1 — Parameter estimations

from which we obtain the parameter estimates.

$$\begin{bmatrix} ? \\ ? \\ \mu_3 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 6.03 \\ 1.08 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & 2.50 \end{bmatrix} \xrightarrow{\text{initial value}} \begin{bmatrix} 0.50 & 0.25 & 1.00 \\ 0.25 & 0.50 & 0.75 \\ 1.00 & 0.75 & 2.50 \end{bmatrix} \xrightarrow{\text{first iteration}} \begin{bmatrix} 0.558 & 0.346 & 1.168 \\ 0.346 & 0.517 & 0.825 \\ 1.168 & 0.825 & 2.50 \end{bmatrix}$$

Next: Iteration 2, ...