

Multiple Imputation Algorithms for General Missing Pattern in the Presence of High-dimensional Data

Yi Deng

July 24, 2015

1 General Missing Pattern Setup

Consider a data set \mathbf{Z} with p variables, $\mathbf{z}_1, \dots, \mathbf{z}_p$. Assume the first l ($l \leq p$) variables have some missing values. We denote the missing components and observed components for variable j by $\mathbf{z}_{j,mis}$ and $\mathbf{z}_{j,obs}$. Suppose we have r_j observed values in variable \mathbf{z}_j .

2 Multiple Imputation through Indirect Use of Regularized Regression (IURR)

2.1 We start the iterative procedure with some initial values. Impute all the elements in $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(0)} = \text{mean}(\mathbf{z}_{j,obs})$, ($j = 1, 2, \dots, l$). Define the initial completed dataset after this imputation as $\mathbf{Z}^{(0)}$, with $\mathbf{z}_j^{(0)} = (\mathbf{z}_{j,mis}^{(0)}, \mathbf{z}_{j,obs})$.

2.2 In the m -th iteration:

For variable j , ($j = 1, \dots, l$), define $\mathbf{W} = \{\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_{j-1}^{(m)}, \mathbf{z}_{j+1}^{(m-1)}, \dots, \mathbf{z}_l^{(m-1)}, \mathbf{z}_{l+1}, \dots, \mathbf{z}_p\}$. Denote by \mathbf{W}_{mis} the component of \mathbf{W} corresponding to $\mathbf{z}_{j,mis}$ and by \mathbf{W}_{obs} the component of \mathbf{W} corresponds to $\mathbf{z}_{j,obs}$.

(i) If \mathbf{Z}_j follows a Gaussian distribution.

We use a regularized regression method to fit a multiple linear regression model considering $\mathbf{z}_{j,obs}$ as the outcome variable and \mathbf{W}_{obs} as the predictor variable, and identify the active set, $\hat{\mathcal{S}}_j^{(m)}$.

Let $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ denote the subset of \mathbf{W} that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$\mathbf{z}_{j,obs} = \theta_{0,j} \mathbf{1}_{r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs} \boldsymbol{\theta}_j + \boldsymbol{\epsilon}_j, \quad (1)$$

where $\boldsymbol{\epsilon}_j \sim N(\mathbf{0}, \sigma_j^2 \mathbf{I}_{r_j})$ and $\mathbf{1}_{r_j}$ is a vector of length r_j with all entries one.

Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j, \sigma_j^2)$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_{0,j}, \boldsymbol{\theta}_j, \sigma_j^2) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (1) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters.

Conduct one imputation for $\mathbf{z}_{j,mis}$: randomly draw $(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j, \hat{\sigma}_j^2)$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $N(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis} \hat{\boldsymbol{\theta}}_j, \hat{\sigma}_j^2 \mathbf{I}_{n-r_j})$. Let $\mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs})$.

(ii) If \mathbf{Z}_j follows a Bernoulli distribution.

We use a regularized regression method to fit a multiple logistic regression model considering $\mathbf{z}_{j,obs}$ as the outcome variable and \mathbf{W}_{obs} as the predictor variable, and identify the active set, $\hat{\mathcal{S}}_j^{(m)}$.

Let $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ denote the subset of \mathbf{W} that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$\text{logit}(\Pr(\mathbf{z}_{j,obs} = 1 | \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs})) = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs} \boldsymbol{\theta}_j, \quad (2)$$

Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j)$ by using a standard inference procedure such as maximum likelihood.

$$((\theta_{0,j}, \boldsymbol{\theta}_j) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (2) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters.

Conduct one imputation for $\mathbf{z}_{j,mis}$: randomly draw $(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j)$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution

$$\text{Bernoulli}\left(\frac{\exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis} \hat{\boldsymbol{\theta}}_j)}{1 + \exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis} \hat{\boldsymbol{\theta}}_j)}\right). \text{ Let } \mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs}).$$

(iii) If \mathbf{Z}_j follows a Poisson distribution.

We use a regularized regression method to fit a multiple Poisson regression model considering $\mathbf{z}_{j,obs}$ as the outcome variable and \mathbf{W}_{obs} as the predictor variable, and identify the active set, $\hat{\mathcal{S}}_j^{(m)}$.

Let $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ denote the subset of \mathbf{W} that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$\log(\mathbf{E}[\mathbf{z}_{j,obs} | \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs}]) = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},obs} \boldsymbol{\theta}_j, \quad (3)$$

Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j)$ by using a standard inference procedure such as maximum likelihood.

$$((\theta_{0,j}, \boldsymbol{\theta}_j) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (3) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters.

Conduct one imputation for $\mathbf{z}_{j,mis}$: randomly draw $(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j)$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution

$$\text{Poisson}(\exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{\hat{\mathcal{S}}_j^{(m)},mis} \hat{\boldsymbol{\theta}}_j)). \text{ Let } \mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs}).$$

- 2.3 Denote the updated data set after the m -th iteration by $\mathbf{Z}^{(m)}$. Repeat the procedures iteratively. After the iterations converge, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.

3 Multiple Imputation through Direct Use of Regularized Regression (DURR)

- 3.1 We start the iterative procedure with some initial values. Impute all the elements in $\mathbf{z}_{mis,j}$ with $\mathbf{z}_{mis,j}^{(0)} = \text{mean}(\mathbf{z}_{obs,j})$, ($j = 1, 2, \dots, l$). Define the initial completed dataset after this imputation as $\mathbf{Z}^{(0)}$.

- 3.2 In the m -th iteration:

For variable j , ($j = 1, \dots, l$), define $\mathbf{W} = \{\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_{j-1}^{(m)}, \mathbf{z}_{j+1}^{(m-1)}, \dots, \mathbf{z}_l^{(m-1)}, \mathbf{z}_{l+1}, \dots, \mathbf{z}_p\}$. Denote by \mathbf{W}_{mis} the component of \mathbf{W} corresponding to $\mathbf{z}_{j,mis}$. We generate a bootstrap data set $\{\mathbf{W}^*, \mathbf{z}_j^*\}$ of size n by randomly drawing n observations from $\{\mathbf{W}, \mathbf{z}_j^{(m-1)}\}$ with replacement. Denote the observed values of \mathbf{z}_j^* by $\mathbf{z}_{j,obs}^*$ and the corresponding component of \mathbf{W}^* by \mathbf{W}_{obs}^* . Suppose we have r_j and r_j^* observed values in variable \mathbf{z}_j and \mathbf{z}_j^* , respectively.

- (i) If \mathbf{Z}_j follows a Gaussian distribution. The model is

$$\mathbf{z}_{j,obs}^* = \theta_{0,j} \mathbf{1}_{r_j^*} + \mathbf{W}_{obs}^* \boldsymbol{\theta}_j + \boldsymbol{\epsilon}_j, \quad (4)$$

where $\boldsymbol{\epsilon}_j \sim N(\mathbf{0}, \sigma_j^2 \mathbf{I}_{r_j^*})$.

Use a regularized regression method to fit model (4) and get parameter estimates as follows:

$$(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j) = \underset{(\theta_{0,j}, \boldsymbol{\theta}_j)}{\operatorname{argmin}} [-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; \mathbf{z}_{j,obs}^*, \mathbf{W}_{obs}^*) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function. Then consider the mean of squared residuals as an estimate of σ_j^2 , denoted by $\hat{\sigma}_j^2$.

Impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $N(\hat{\theta}_{0,j} \mathbf{1}_{n-r_j} + \mathbf{W}_{mis} \hat{\boldsymbol{\theta}}_j, \hat{\sigma}_j^2 \mathbf{I}_{n-r_j})$, noting that imputation is conducted on the original data set \mathbf{W}_{mis} , not the bootstrap data set \mathbf{W}^* . Let $\mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs}^*)$.

- (ii) If \mathbf{Z}_j follows a Bernoulli distribution. The model is

$$\text{logit}(\mathbf{z}_{j,obs}^* = 1 | \mathbf{W}_{obs}^*) = \theta_{0,j} \mathbf{1}_{r_j} + \mathbf{W}_{obs}^* \boldsymbol{\theta}_j, \quad (5)$$

Use a regularized regression method to fit model (5) and get parameter estimates as follows:

$$(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j) = \underset{(\theta_{0,j}, \boldsymbol{\theta}_j)}{\operatorname{argmin}} [-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; \mathbf{z}_{j,obs}^*, \mathbf{W}_{obs}^*) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function.

Impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution

$Bernoulli(\frac{\exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{mis}\hat{\boldsymbol{\theta}}_j)}{1 + \exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{mis}\hat{\boldsymbol{\theta}}_j)})$, noting that imputation is conducted on the original data set \mathbf{W}_{mis} , not the bootstrap data set \mathbf{W}^* . Let $\mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs})$.

(iii) If \mathbf{Z}_j follows a Poisson distribution. The model is

$$\log(\mathbf{E}[\mathbf{z}_{j,obs}^* | \mathbf{W}_{obs}^*]) = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{obs}^* \boldsymbol{\theta}_j, \quad (6)$$

Use a regularized regression method to fit model (6) and get parameter estimates as follows:

$$(\hat{\theta}_{0,j}, \hat{\boldsymbol{\theta}}_j) = \underset{(\theta_{0,j}, \boldsymbol{\theta}_j)}{\operatorname{argmin}} [-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; \mathbf{z}_{j,obs}^*, \mathbf{W}_{obs}^*) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function.

Impute $\mathbf{z}_{j,mis}$ with $\mathbf{z}_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $Poisson(\exp(\hat{\theta}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}_{mis}\hat{\boldsymbol{\theta}}_j))$, noting that imputation is conducted on the original data set \mathbf{W}_{mis} , not the bootstrap data set \mathbf{W}^* . Let $\mathbf{z}_j^{(m)} = (\mathbf{z}_{j,mis}^{(m)}, \mathbf{z}_{j,obs})$.

3.3 Denote the updated data set after the m -th iteration by $\mathbf{Z}^{(m)}$. Repeat the procedures iteratively. After the iterations converge, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.