

1 A Simplified Data Setup

We consider the same simplified data setup from the paper, where only one variable has missing values with the others fully observed. Without loss of generality, we assume that \mathbf{z}_1 is continuous and contains missing values with the first r components observed, $\mathbf{z}_{obs,1} = (z_{1,1}, \dots, z_{r,1})^T$, and the remaining $n - r$ components missing, $\mathbf{z}_{mis,1} = (z_{r+1,1}, \dots, z_{n,1})^T$. Define the complement data set for \mathbf{z}_1 as $\mathbf{Z}_{-1} = (\mathbf{z}_{1,-1}, \mathbf{z}_{2,-1}, \dots, \mathbf{z}_{n,-1})^T = (\mathbf{Z}_{obs,-1}, \mathbf{Z}_{mis,-1})^T$ with $\mathbf{z}_{i,-1} = (z_{i,2}, z_{i,3}, \dots, z_{i,p})^T$, and define the complement data sets for $\mathbf{z}_{obs,1}$ and $\mathbf{z}_{mis,1}$ as $\mathbf{Z}_{obs,-1} = (\mathbf{z}_{1,-1}, \dots, \mathbf{z}_{r,-1})^T$ and $\mathbf{Z}_{mis,-1} = (\mathbf{z}_{r+1,-1}, \dots, \mathbf{z}_{n,-1})^T$, respectively. Then the observed data are $\mathbf{Z}_{obs} = (\mathbf{z}_{obs,1}, \mathbf{Z}_{obs,-1}, \mathbf{Z}_{mis,-1})$ and the missing data are $\mathbf{Z}_{mis} = (\mathbf{z}_{mis,1})$; there are r complete cases and $n - r$ incomplete cases with \mathbf{z}_1 missing.

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

2.1 Gaussian

2.1.1 Generate a bootstrap data set $\mathbf{Z}^{(m)}$ of size n by randomly drawing n observations from \mathbf{Z} with replacement.

2.1.2

$$\mathbf{z}_{obs,1}^{(m)} = \alpha_0^{(m)} + \mathbf{Z}_{obs,-1}^{(m)} \boldsymbol{\alpha}^{(m)} + \boldsymbol{\epsilon}^{(m)}, \quad (1)$$

where $\boldsymbol{\epsilon}^{(m)} \sim N(\mathbf{0}, \sigma^{2(m)} \mathbf{I}_r)$ and $\boldsymbol{\alpha}^{(m)} = (\alpha_1^{(m)}, \dots, \alpha_{p-1}^{(m)})^T$.

$$(\hat{\alpha}_0^{(m)}, \hat{\boldsymbol{\alpha}}^{(m)}) = \underset{(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)})}{\operatorname{argmin}} [-\ell(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)}; \mathbf{z}_{obs,1}^{(m)}, \mathbf{Z}_{obs,-1}^{(m)}) + P_\lambda(\boldsymbol{\alpha}^{(m)})] \quad (2)$$

Where $P_\lambda(\boldsymbol{\alpha}^{(m)})$ is a regularization function.

$$\hat{\mathbf{z}}_{obs,1}^{(m)} = \hat{\alpha}_0^{(m)} + \mathbf{Z}_{obs,-1}^{(m)} \hat{\boldsymbol{\alpha}}^{(m)}$$

$$\hat{\sigma}^{2(m)} = (\mathbf{z}_{obs,1}^{(m)} - \hat{\mathbf{z}}_{obs,1}^{(m)})^T (\mathbf{z}_{obs,1}^{(m)} - \hat{\mathbf{z}}_{obs,1}^{(m)}) / r$$

2.1.3 Impute $\mathbf{z}_{mis,1}$ with $\mathbf{z}_{mis,1}^{(m)}$ by drawing randomly from the predictive distribution $N(\hat{\alpha}_0^{(m)} + \mathbf{Z}_{mis,-1} \hat{\boldsymbol{\alpha}}^{(m)}, \hat{\sigma}^{2(m)} \mathbf{I}_r)$, noting that imputation is conducted on the original data set, not the bootstrap data set.

2.2 Binary

2.2.1 Generate a bootstrap data set $\mathbf{Z}^{(m)}$ of size n by randomly drawing n observations from \mathbf{Z} with replacement.

2.2.2 Suppose

$$z_{i,1}^{(m)} \sim \text{Bernoulli}(\pi_i^{(m)}) \quad (3)$$

$$\log\left(\frac{\pi_i^{(m)}}{1 - \pi_i^{(m)}}\right) = \alpha_0^{(m)} + \mathbf{z}_{i,-1}^{(m)} \boldsymbol{\alpha}^{(m)} \quad i = (1, \dots, r) \quad (4)$$

Then,

$$(\hat{\alpha}_0^{(m)}, \hat{\boldsymbol{\alpha}}^{(m)}) = \underset{(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)})}{\operatorname{argmin}} [-\ell(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)}; \mathbf{z}_{obs,1}^{(m)}, \mathbf{Z}_{obs,-1}^{(m)}) + P_\lambda(\boldsymbol{\alpha}^{(m)})] \quad (5)$$

Where $P_\lambda(\boldsymbol{\alpha}^{(m)})$ is a regularization function.

We can get $\hat{\pi}_j = \frac{\exp(\hat{\alpha}_0^{(m)} + \mathbf{z}_{j,-1} \hat{\boldsymbol{\alpha}}^{(m)})}{1 + \exp(\hat{\alpha}_0^{(m)} + \mathbf{z}_{j,-1} \hat{\boldsymbol{\alpha}}^{(m)})} \quad j = (r + 1, \dots, n)$

2.2.3 For j in $r + 1, \dots, n$, impute $\mathbf{z}_{j,1}$ with $\mathbf{z}_{j,1}^{(m)}$ by drawing randomly from the predictive distribution $\text{Bernoulli}(\hat{\pi}_j)$, noting that imputation is conducted on the original data set, not the bootstrap data set.

2.3 Poisson

2.3.1 Generate a bootstrap data set $\mathbf{Z}^{(m)}$ of size n by randomly drawing n observations from \mathbf{Z} with replacement.

2.3.2 Suppose

$$z_{i,1}^{(m)} \sim \text{Poisson}(\mu_i^{(m)}) \quad (6)$$

$$\log(\mu_i^{(m)}) = \alpha_0^{(m)} + \mathbf{z}_{i,-1}^{(m)} \boldsymbol{\alpha}^{(m)} \quad i = (1, \dots, r) \quad (7)$$

Then,

$$(\hat{\alpha}_0^{(m)}, \hat{\boldsymbol{\alpha}}^{(m)}) = \underset{(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)})}{\operatorname{argmin}} [-\ell(\alpha_0^{(m)}, \boldsymbol{\alpha}^{(m)}; \mathbf{z}_{obs,1}^{(m)}, \mathbf{Z}_{obs,-1}^{(m)}) + P_\lambda(\boldsymbol{\alpha}^{(m)})] \quad (8)$$

Where $P_\lambda(\boldsymbol{\alpha}^{(m)})$ is a regularization function.

We can get $\hat{\mu}_j = \exp(\hat{\alpha}_0^{(m)} + \mathbf{z}_{j,-1} \hat{\boldsymbol{\alpha}}^{(m)}) \quad j = (r + 1, \dots, n)$

2.3.3 For j in $r + 1, \dots, n$, impute $\mathbf{z}_{j,1}$ with $\mathbf{z}_{j,1}^{(m)}$ by drawing randomly from the predictive distribution $\text{Poisson}(\hat{\mu}_j)$, noting that imputation is conducted on the original data set, not the bootstrap data set.

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

3.1 Gaussian

3.1.1 Suppose

$$\mathbf{z}_{obs,1} = \alpha_0 + \mathbf{Z}_{obs,-1}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad (9)$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_r)$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{p-1})^T$.

Use a regularized regression method to fit model (9) and identify the active set, $\hat{\mathcal{S}}$.

3.1.2 Then the model is

$$\mathbf{z}_{obs,1} = \theta_0 + \mathbf{Z}_{obs,\hat{\mathcal{S}}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (10)$$

Approximate the distribution of $(\theta_0, \boldsymbol{\theta}, \sigma^2)$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_0, \boldsymbol{\theta}, \sigma^2) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE}) \quad (11)$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (10) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters.

3.1.3 Conduct multiple imputation for $\mathbf{z}_{mis,1}$: in the m -th imputation, randomly draw $(\hat{\theta}_0^{(m)}, \hat{\boldsymbol{\theta}}^{(m)}, \hat{\sigma}^{2(m)})$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and subsequently impute $\mathbf{z}_{mis,1}$ with $\mathbf{z}_{mis,1}^{(m)}$ by drawing randomly from the predictive distribution $N(\hat{\theta}_0^{(m)} + \mathbf{Z}_{mis,\hat{\mathcal{S}}}\hat{\boldsymbol{\theta}}^{(m)}, \hat{\sigma}^{2(m)} \mathbf{I}_r)$.

3.2 Binary

3.2.1 Suppose

$$z_{i,1}^{(m)} \sim \text{Bernoulli}(\pi_i^{(m)}) \quad (12)$$

$$\log\left(\frac{\pi_i^{(m)}}{1 - \pi_i^{(m)}}\right) = \alpha_0^{(m)} + \mathbf{z}_{i,-1}^{(m)}\boldsymbol{\alpha}^{(m)} \quad i = (1, \dots, r) \quad (13)$$

Use a regularized regression method to fit model (13) and identify the active set, $\hat{\mathcal{S}}$.

3.2.2 Then the model is

$$\log\left(\frac{\pi_i^{(m)}}{1 - \pi_i^{(m)}}\right) = \theta_0^{(m)} + \mathbf{z}_{i,\hat{\mathcal{S}}}^{(m)}\boldsymbol{\theta}^{(m)} \quad i = (1, \dots, r) \quad (14)$$

Approximate the distribution of $(\theta_0, \boldsymbol{\theta})$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_0, \boldsymbol{\theta}) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE}) \quad (15)$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (14) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters using the generalized linear model.

3.2.3 Conduct multiple imputation for $\mathbf{z}_{j,1}$: in the m -th imputation, randomly draw $(\hat{\theta}_0^{(m)}, \hat{\boldsymbol{\theta}}^{(m)})$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and subsequently impute $\mathbf{z}_{j,1}$ with $\mathbf{z}_{j,1}^{(m)}$ by drawing randomly from the predictive distribution $Bernoulli(\hat{\pi}_j^{(m)})$, where $\hat{\pi}_j^{(m)} = \frac{\exp(\hat{\theta}_0^{(m)} + \mathbf{z}_{j,\hat{\mathcal{S}}} \hat{\boldsymbol{\theta}}^{(m)})}{1 + \exp(\hat{\theta}_0^{(m)} + \mathbf{z}_{j,\hat{\mathcal{S}}} \hat{\boldsymbol{\theta}}^{(m)})}$. $j = (r + 1, \dots, n)$

3.3 Poisson

3.3.1 Suppose

$$z_{i,1}^{(m)} \sim \text{Poisson}(\mu_i^{(m)}) \quad (16)$$

$$\log(\mu_i^{(m)}) = \alpha_0^{(m)} + \mathbf{z}_{i,-1}^{(m)} \boldsymbol{\alpha}^{(m)} \quad i = (1, \dots, r) \quad (17)$$

Use a regularized regression method to fit model (17) and identify the active set, $\hat{\mathcal{S}}$.

3.3.2 Then the model is

$$\log(\mu_i^{(m)}) = \theta_0^{(m)} + \mathbf{z}_{i,\hat{\mathcal{S}}}^{(m)} \boldsymbol{\theta}^{(m)} \quad i = (1, \dots, r) \quad (18)$$

Approximate the distribution of $(\theta_0, \boldsymbol{\theta})$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_0, \boldsymbol{\theta}) \sim N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE}) \quad (19)$$

Where $\hat{\boldsymbol{\theta}}_{MLE}$ is the MLE of parameters in model (18) and $\hat{\boldsymbol{\Sigma}}_{MLE}$ is the variance-covariance matrix of the estimated parameters using the generalized linear model.

3.3.3 Conduct multiple imputation for $\mathbf{z}_{j,1}$: in the m -th imputation, randomly draw $(\hat{\theta}_0^{(m)}, \hat{\boldsymbol{\theta}}^{(m)})$ from $N(\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{MLE})$, and subsequently impute $\mathbf{z}_{j,1}$ with $\mathbf{z}_{j,1}^{(m)}$ by drawing randomly from the predictive distribution $\text{Poisson}(\hat{\mu}_j^{(m)})$, where $\hat{\mu}_j^{(m)} = \exp(\hat{\theta}_0^{(m)} + \mathbf{z}_{j,\hat{\mathcal{S}}} \hat{\boldsymbol{\theta}}^{(m)})$. $j = (r + 1, \dots, n)$