Employing Machine Learning to Handle Missing Data

Presented by: Anindro Bhattacharya | March 23, 2022

Agenda

- Mechanisms of Missingness
- NeuMiss Networks
- Doubly Robust Estimators
- Next Steps

Mechanisms of Missingness

Notation

- V: set of observable variables
  - V_o: set of variables that are observed in all records in the population
  - V_m: set of variables missing in at least one record
- U: set of unobserved variables (latent variables)
- R: missingness pattern

Notations in Action

U

F_1	F_2	F_3	F_4
1	2	3	NA
423	2	NA	10
34	32	42	NA

V_o V_m

V

Missing Completely at Random (MCAR)

- Probability that V_m is missing is independent of V_m or any other variable in the study
- e.g., deciding to reveal income levels based on coin flips

Missing at Random (MAR)

- For all data cases Y, $P(R|Y_{obs}, Y_{mis}) = P(R|Y_{obs})$
  - Y_{obs}: observed component of Y
  - Y_{mis}: observed component of Y
- e.g., patients under the age of 65 are less likely to have bone density tests taken

Missing Not at Random (MNAR)

- Neither MCAR or MAR
- Missingness is related to factors not measured in the study
    - Probability that V_m is missing is dependent on some $u \in U$
- e.g., smoking status is not recorded in patients admitted as an emergency

NeuMiss Networks

Morvan, M. L., Josse, J., Moreau, T., Scornet, E., & Varoquaux, G. (2020). Neumann networks: differential programming for supervised learning with missing values. *CoRR, abs/2007.01627*. Opgehaal van https://arxiv.org/abs/2007.01627

Notation

- $X \in \mathbb{R}^d$: set of features
- $Y \in \mathbb{R}$: outcomes
- $M \in \{0,1\}^d$
  - $\forall\, 1 \leq j \leq d, M_j = 1 \Leftrightarrow X_j$ is not observed
- For realizations m of M,
  - $obs(m)$: indices of zero entries of m
  - $mis(m)$: indices of non-zero entries of m

Notation - Example

- Suppose:
  - $x = (1.1,\ 2.3,\ -3.1,\ 8,\ 5.27)$
  - $m = (0,\ 1,\ 0,\ 0,\ 1)$
- then:
  - $\tilde{x} = (1.1,\ \mathrm{NA},\ -3.1,\ 8,\ \mathrm{NA})$
  - $obs(m) = \{0,\ 2,\ 3\}$
  - $mis(m) = \{1,\ 4\}$
  - $x_{obs(m)} = (1.1,\ -3.1,\ 8)$
  - $x_{mis(m)} = (2.3,\ 5.27)$

Bayes Predictor

$$f^\star(X_{obs(M)}, M) = \mathbb{E}\left[Y | X_{obs(M)}, M\right]$$

Assumption 1 (Gaussian data). *The distribution of X is Gaussian, that is, $X \sim \mathcal{N}(\mu, \Sigma)$.*

Assumption 2 (MCAR mechanism). *For all $m \in \{0,1\}^d$, $P(M = m|X) = P(M = m)$.*

Assumption 3 (MAR mechanism). *For all $m \in \{0,1\}^d$, $P(M = m|X) = P(M = m|X_{obs(m)})$.*

Proposition 2.1 (MAR Bayes predictor). *Assume that the data are generated via the linear model defined in equation (1) and satisfy Assumption 1. Additionally, assume that either Assumption 2 or Assumption 3 holds. Then the Bayes predictor $f^\star$ takes the form*

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle, \quad (4)$$

where we use obs (resp. mis) instead of obs(M) (resp. $mis(M)$) for lighter notations.

Deriving the Bayes Predictor

$$f^{\star}_{\widetilde{X}}(\widetilde{X}) = \mathbb{E}[Y|\widetilde{X}]$$

$$= \mathbb{E}[\beta_0^{\star} + \langle \beta^{\star}, X\rangle \mid M, X_{obs(M)}]$$, by linear model

$$= \beta_0^{\star} + \langle \beta^{\star}_{obs(M)}, X_{obs(M)}\rangle + \langle \beta^{\star}_{mis(M)}, \mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}]\rangle.$$

If missingness is MAR or MCAR,

$$\mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] = \mathbb{E}[X_{mis(M)} \mid X_{obs(M)}].$$

Since $X \sim N(\mu, \Sigma)$,

$$\mathbb{E}\left[X_{mis(m)} \mid X_{obs(m)}\right] = \mu_{mis(m)} + \Sigma_{mis(m),obs(m)} \left(\Sigma_{obs(m)}\right)^{-1} \left(X_{obs(m)} - \mu_{obs(m)}\right)$$

Therefore,

$$f^{\star}(X_{obs}, M) = \beta_0^{\star} + \langle \beta^{\star}_{obs}, X_{obs}\rangle + \langle \beta^{\star}_{mis}, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})\rangle$$

The Neumann Series

$$(I - S)^{-1} = \sum_{j=0}^{\infty} S^j = I + S + S^2 + \cdots$$

Example with Neumann Series

Estimate inverse of $\boldsymbol{T} = \begin{bmatrix} \mathbf{1} & \mathbf{0.5} \\ \mathbf{0.25} & \mathbf{1} \end{bmatrix}$. Actual value of $T^{-1} = \begin{bmatrix} 1.14285714 & -0.57142857 \\ -0.28571429 & 1.14285714 \end{bmatrix}$.

1st order approximation of $T^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

3rd order approximation of $T^{-1} = \begin{bmatrix} 1.125 & -0.5 \\ -0.25 & 1.125 \end{bmatrix}$.

10th order approximation of $T^{-1} = \begin{bmatrix} 1.14282227 & -0.57141113 \\ -0.28570557 & 1.14282227 \end{bmatrix}$.

20th order approximation of $T^{-1} = \begin{bmatrix} 1.14285714 & -0.57142857 \\ -0.28571429 & 1.14285714 \end{bmatrix}$.

Approximating Inverse Covariances with Neumann Series

- $S^{(0)}$: starting point for approximation (*d* x *d*)
- $S^{(0)}_{obs(m)}$: submatrix of $S^{(0)}$ obtained by selecting components for which $m = 0$
  - Order-0 approximation of $\left(\Sigma_{obs(m)}\right)^{-1}$

Approximating Inverse Covariances with Neumann Series

- For all $m \in \{0,1\}^d$, define the order-l approximation $S_{obs(m)}^{(l)}$ of $\left(\Sigma_{obs(m)}\right)^{-1}$ via:

$$S_{obs(m)}^{(\ell)} = \left(Id - \Sigma_{obs(m)}\right) S_{obs(m)}^{(\ell-1)} + Id$$

Order-ℓ Approximation of Bayes Predictor

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs}) \rangle$$

$$\approx$$

$$f_\ell^\star(X_{obs}, M) = \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs} S_{obs(m)}^{(\ell)} (X_{obs} - \mu_{obs}) \rangle$$

NeuMiss Architecture

- Approximates the Bayes predictor
- Inverses $\left(\Sigma_{obs(m)}\right)^{-1}$ are computed using an unrolled version of the iterative algorithm

NeuMiss Architecture

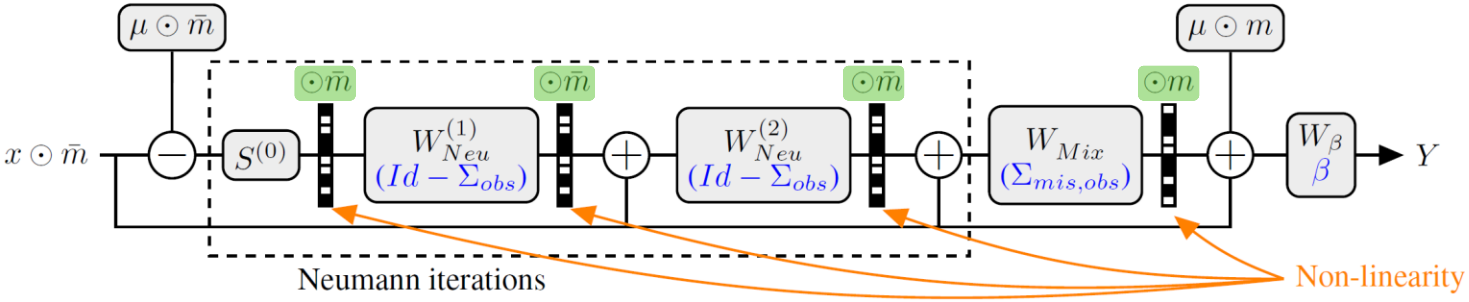


Figure 1: **NeuMiss network architecture with a depth of 4** — $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Results from NeuMiss

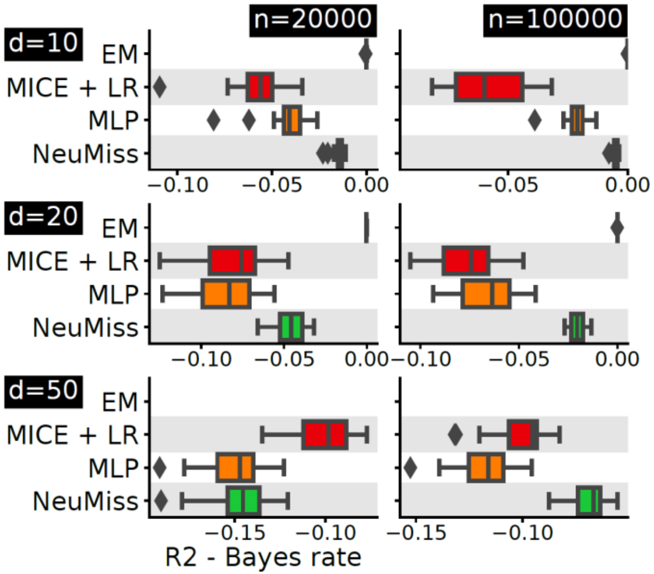
- Generated synthetic data according to multivariate Gaussian distribution
  - $\Sigma = UU^{\mathrm{T}} + \mathrm{diag}(\epsilon)$
  - $U \in \mathbb{R}^{d \times d/2}$ and entries of U drawn from N(0, 1)
  - ϵ: vector of entries drawn uniformly in [0.01, 0.1]

Results from NeuMiss

- Compared performance of NeuMiss to 3 other algorithms
  - **EM:** Expectation-Maximization algorithm run to estimate parameters of $P(X, Y)$ with missing values
    - Predictions made using E[X|Y]
  - **MICE+LR:** data is imputed using conditional imputation (scikit-learn IterativeImputer), LR fit on imputed data
  - **MLP:** multilayer perceptron with 1 hidden layer followed by ReLU nonlinearity; data imputed with 0

Results from NeuMiss

- EM gives best results when it can be run
  - Cannot be run when $d \geq 50$
- NeuMiss is second best
  - Except when d=50, n=20 000
- NeuMiss works best for high $\frac{\text{\# of samples}}{\text{\# of parameters}}$ ratio



Doubly Robust Estimators

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Doubly Robust (DR) Estimators

- Remain consistent when either:
  - A model for the treatment assignment mechanism is correctly specified
  - A model for counterfactual data is correctly specified

Notation

- **Full data: L** = (**V**', Y)'
  - **V**: always observed baseline variables
  - Y: scalar outcome which is missing on some subjects
- **Observed data: O** = (Δ, $\mathbf{L}_{\mathbf{obs}}$)
  - $\mathbf{L}_{\mathbf{obs}}$ = **L** when Δ = 1
  - $\mathbf{L}_{\mathbf{obs}}$ = **V** when Δ = 0

GOAL

- Estimate unconditional mean μ of Y based on n i.i.d. copies of $\mathbf{O}_i$, where i = 1, …, n
- *Assumptions:*
  - Y is MAR
  - $P(\Delta = 1 \mid Y, \boldsymbol{V}) = P(\Delta = 1 \mid \boldsymbol{V}) \equiv \pi(\boldsymbol{V}) > 0$
  - $\mu = E(Y) = E\{E(Y \mid \boldsymbol{V})\}$

Propensity Score

Mean of Y in Terms of Observed Data Distribution

$$
\begin{aligned}
\mu &= E(Y) \\
&= E\{E(Y \mid \boldsymbol{V})\} \\
&= E\{E(Y \mid \Delta = 1, \boldsymbol{V})\} \star \\
&= E\left(\frac{\Delta Y}{\pi(\boldsymbol{V})}\right) \star
\end{aligned}
$$

Method 1

1. Fit model for PS $\pi(\boldsymbol{V})$ based on parametric model $\pi(\boldsymbol{V};\boldsymbol{\alpha})$
2. Estimate μ with Horvitz-Thompson (HT) estimator

$$\hat{\mu}_{HT} = n^{-1}\Sigma_i \frac{\Delta_i Y_i}{\pi(\boldsymbol{V}_i;\widehat{\boldsymbol{\alpha}})}, \text{where } \widehat{\boldsymbol{\alpha}} \text{ is the MLE of } \boldsymbol{\alpha}$$

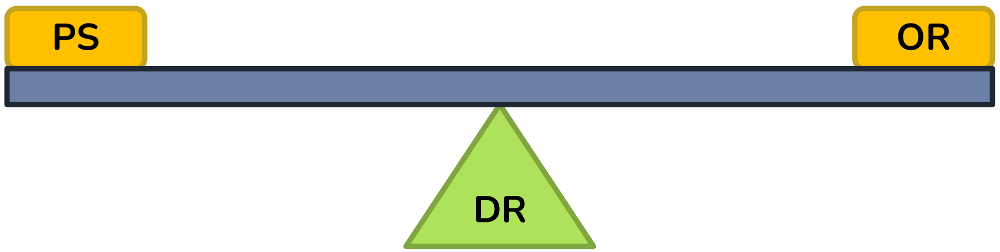
Method 2

1. Fit model for $\Psi\{s(\boldsymbol{V};\boldsymbol{\beta})\}$ for $E(Y \mid \Delta = 1, \boldsymbol{V})$
   - Ψ^{-1}: known link function
   - $s(\boldsymbol{V};\boldsymbol{\beta})$: known regression function ($\boldsymbol{\beta}$ is an unknown finite-dimensional parameter)
2. Estimate μ by OR estimator

$$\hat{\mu}_{OR} = n^{-1}\Sigma_i \Psi\{s(\boldsymbol{V}_i;\tilde{\boldsymbol{\beta}})\}$$

DR Estimator (A Combination of Methods 1 and 2)

- Model $E(Y \mid \Delta = 1, \boldsymbol{V})$ as $e(\boldsymbol{V}; \boldsymbol{\beta}, \phi) = \Psi\{(s(\boldsymbol{V}; \beta) + \phi\pi^{-1}(\boldsymbol{V}; \widehat{\boldsymbol{\alpha}})\}$
- $\widehat{\mu_{dr}} = n^{-1}\Sigma_i \Psi\{(s\left(\boldsymbol{V}_i; \widehat{\boldsymbol{\beta}}\right) + \phi\pi^{-1}(\boldsymbol{V}_i; \widehat{\boldsymbol{\alpha}})\}$
  - $\phi = \Delta_i(\mathrm{Y_i} - s\left(\boldsymbol{V}_i; \widehat{\boldsymbol{\beta}}\right))$



Simulation Study Results

The naively calculated mean on the complete data is 0.03116.

Estimator	$\hat{E(Y)}$	Bias
Horvitz-Thompson	0.07259	-0.04143
Outcome Regression	0.03214	-0.00098
Doubly Robust	0.03043	0.00072

Table 5: Estimates of $E(Y)$ with correctly specified π and s models

Models are incorrectly specified by fitting them on the original dataset masked with MCAR missingness and missingness rate of 0.8.

Estimator	$\hat{E(Y)}$	Bias
Horvitz-Thompson	0.05721	-0.02605
Doubly Robust	0.03043	0.00072

Table 6: Estimates of $E(Y)$ with incorrectly specified π and correctly specified s models

Estimator	$\hat{E(Y)}$	Bias
Outcome Regression	0.00970	0.02145
Doubly Robust	0.06804	-0.01688

Table 7: Estimates of $E(Y)$ with correctly specified π and incorrectly specified s models

Estimator	$\hat{E(Y)}$	Bias
Doubly Robust	0.06804	-0.01688

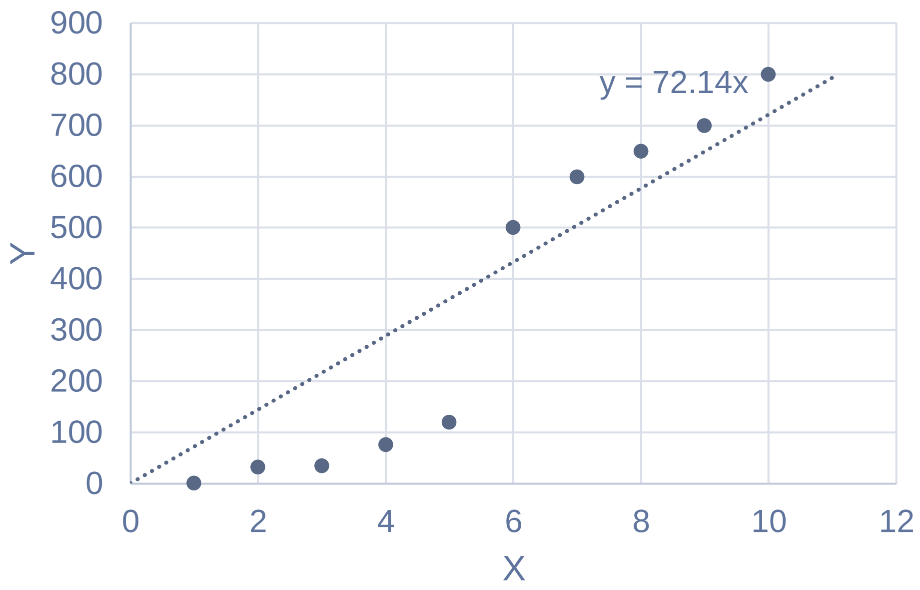
Table 8: Estimates of $E(Y)$ with incorrectly specified π and incorrectly specified s models

Next Steps

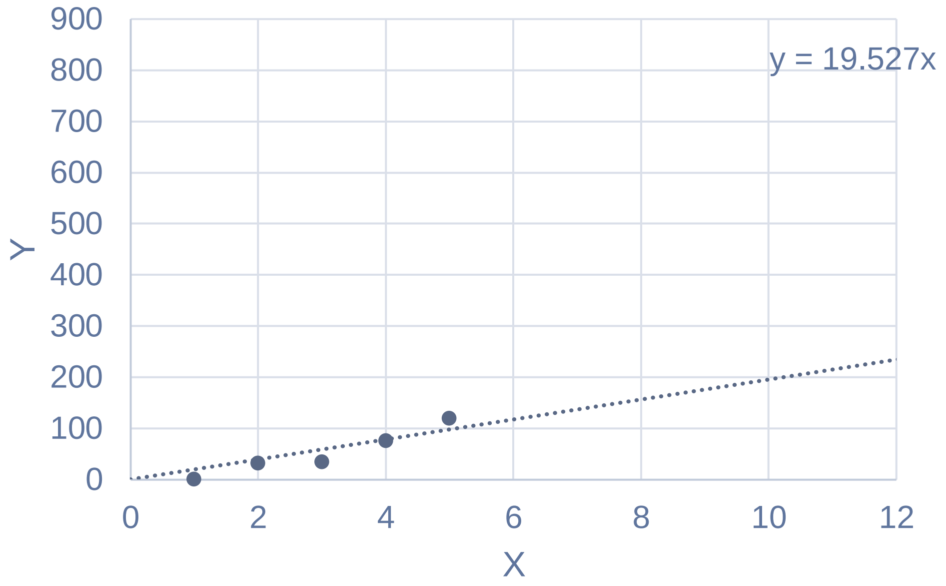
Combining NeuMiss and DR Estimators

- Use propensity score model to decide how likely a subject's outcome is to be missing
- Use propensity score model to modify Bayes' predictor and over-compensate for subjects that are more likely to be missing
- Use updated Bayes' predictor to impute data

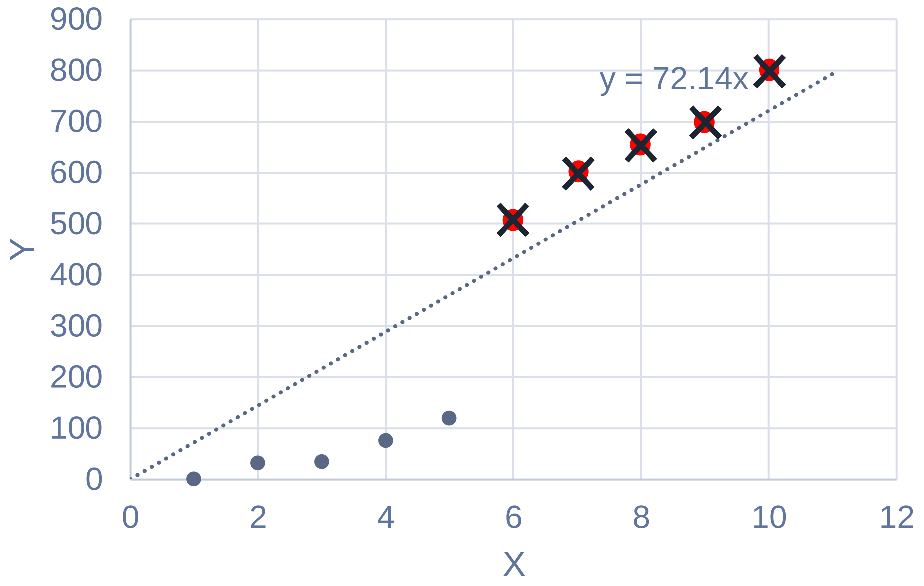
Visual Example



VISUAL EXAMPLE



Visual Example



Any suggestions?