SMPGD 2026

Generative Methods for Handling Missing Data

Aude Sportisse
<aude.sportisse@univ-grenoble-alpes.fr>
CNRS researcher
Laboratoire d'Informatique de Grenoble, APTIKAL Team

Plan

Missing values in real applications

Missing values in tabular datasets

Missing values in many variables (quantitative, categorical, binary).

TraumaBase®

Hospital	Heart rate	Death	Anticoagulant therapy
Pitie-Salpêtrière	88	0	No
Beaujon	103	0	NA
Bicêtre	NA	0	No
Lille	NA	0	No

+ 250 clinical variables

+30 000 patients, 30 different hospitals

Formal definition

Definition (Little and Rubin, 2019)

"Missing data are unobserved values that would be **meaningful for analysis** if observed"

$$\underbrace{\begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \\ 25 & 47 & 86 \end{pmatrix}}_{\text{not observed}} \qquad \underbrace{\begin{pmatrix} 30 & \texttt{NA} & 61 \\ \texttt{NA} & \texttt{NA} & 50 \\ 25 & 47 & \texttt{NA} \end{pmatrix}}_{\text{observed}}$$

Classical learning methods only for complete datasets

The mean cannot be calculated:

$$(30 + \texttt{NA} + 25)/3.$$

Formal definition

Definition (Little and Rubin, 2019)

"Missing data are unobserved values that would be **meaningful for analysis** if observed"

$$\underbrace{\begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \\ 25 & 47 & 86 \end{pmatrix}}_{\text{not observed}} \qquad \underbrace{\begin{pmatrix} 30 & \texttt{NA} & 61 \\ \texttt{NA} & \texttt{NA} & 50 \\ 25 & 47 & \texttt{NA} \end{pmatrix}}_{\text{observed}}$$

Classical learning methods only for complete datasets

The mean cannot be calculated:

$$(30 + \texttt{NA} + 25)/3.$$

Can we simply delete NAs?

- Loss of information

  Example (Zhu et al., 2022):
  dataset with d variables, ratio of NA: 1%
  - $d = 5$: 95% of complete rows
  - $d = 300$: 5% of complete rows

- Bias in the analysis

  Kept observations: sub-population **not necessarily representative** of the overall population.

Can we simply delete NAs? ✗ No, in most cases.

Can we simply delete NAs?

- **Loss of information**

  Example (Zhu et al., 2022):
  dataset with d variables, ratio of NA: 1%
  - $d = 5$: 95% of complete rows
  - $d = 300$: 5% of complete rows

- **Bias in the analysis**

  Kept observations: sub-population **not necessarily representative** of the overall population.

Can we simply delete NAs? ✗ No, in most cases.

Can we simply delete NAs?

- **Loss of information**

  Example (Zhu et al., 2022):
  dataset with d variables, ratio of NA: 1%
  - $d = 5$: 95% of complete rows
  - $d = 300$: 5% of complete rows

- **Bias in the analysis**

  Kept observations: sub-population **not necessarily representative** of the overall population.

Can we simply delete NAs? ✗ No, in most cases.

Plan

Informative missing values

Example in the TraumaBase® dataset

In emergency situations, doctors do not fill the form → the missing values are more likely to be high heart rate values

Machine can fail **randomly**.

Complete dataset		**Missing dataset**
Hospital	Heart Rate	Heart Rate
Bicêtre	200	NA
Bicêtre	59	59
Pitié	65	65
Pitié	210	NA

Informative

Missing dataset
Heart Rate
NA
59
NA
210

Not informative

Formalism

$$X^{\text{NA}} = \begin{pmatrix} 30 & \texttt{NA} & 61 \end{pmatrix}$$

$$X = \begin{pmatrix} 30 & 100 & 61 \end{pmatrix}$$

$$X_{\text{obs}(M)} = \begin{pmatrix} 30 & 61 \end{pmatrix} \quad X_{\text{mis}(M)} = \begin{pmatrix} 100 \end{pmatrix}$$

$$M = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

Formalism

$$X^{\text{NA}} = \begin{pmatrix} 30 & \text{NA} & 61 \end{pmatrix}$$

$$X = \begin{pmatrix} 30 & 100 & 61 \end{pmatrix}$$

$$X_{\text{obs}(M)} = \begin{pmatrix} 30 & 61 \end{pmatrix} \quad X_{\text{mis}(M)} = \begin{pmatrix} 100 \end{pmatrix}$$

$$M = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

Formalism

$$X^{\text{NA}} = \begin{pmatrix} 30 & \texttt{NA} & 61 \end{pmatrix}$$

$$X = \begin{pmatrix} 30 & 100 & 61 \end{pmatrix}$$

$$X_{\text{obs}(M)} = \begin{pmatrix} 30 & 61 \end{pmatrix} \quad X_{\text{mis}(M)} = \begin{pmatrix} 100 \end{pmatrix}$$

$$M = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

Formalism

$$X^{\text{NA}} = \begin{pmatrix} 30 & \texttt{NA} & 61 \end{pmatrix}$$

$$X = \begin{pmatrix} 30 & 100 & 61 \end{pmatrix}$$

$$X_{\text{obs}(M)} = \begin{pmatrix} 30 & 61 \end{pmatrix} \quad X_{\text{mis}(M)} = \begin{pmatrix} 100 \end{pmatrix}$$

$$M = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

Type of missing-values

(Rubin, 1976)

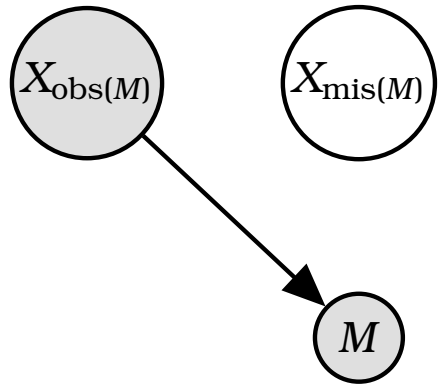
- MCAR: $p(M|X) = p(M)$
- MAR: $p(M|X) = p(M|X_{\text{obs}(M)})$

- MNAR: $p(M|X) = p(M|X_{\text{obs}(M)}, X_{\text{mis}(M)})$ =informative case

Type of missing-values

(Rubin, 1976)

- MCAR: $p(M|X) = p(M)$
- MAR: $p(M|X) = p(M|X_{\text{obs}(M)})$

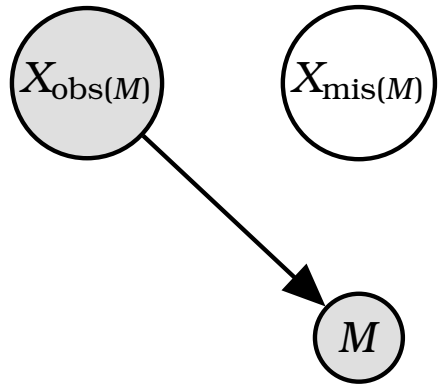


- MNAR: $p(M|X) = p(M|X_{\text{obs}(M)}, X_{\text{mis}(M)})$ =informative case

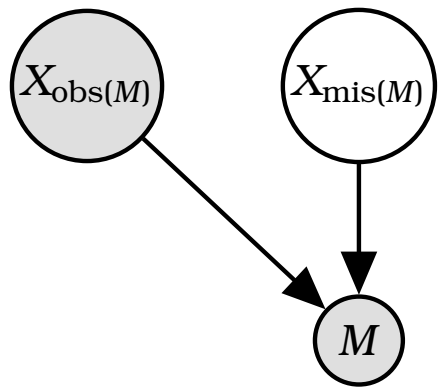
Type of missing-values

(Rubin, 1976)

- MCAR: $p(M|X) = p(M)$
- MAR: $p(M|X) = p(M|X_{\text{obs}(M)})$



- MNAR: $p(M|X) = p(M|X_{\text{obs}(M)}, X_{\text{mis}(M)})$ =informative case



Likelihood-approach

Joint distribution :

$$p(X, M) = p(X)p(M|X)$$

θ: parameter of the data distribution $p(X)$
ϕ: parameter of the missing-data mechanism $p(M|X)$

Classical likelihood (in the classical case):

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full}}(\theta, \phi; X, M)$$

Observed likelihood (when there is missing values):

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)}$$

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M)$$

Likelihood-approach

Joint distribution :

$$p(X, M) = p(X)p(M|X)$$

θ: parameter of the data distribution $p(X)$
ϕ: parameter of the missing-data mechanism $p(M|X)$

Classical likelihood (in the classical case):

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full}}(\theta, \phi; X, M)$$

Observed likelihood (when there is missing values):

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)}$$

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M)$$

Likelihood-approach

Joint distribution :

$$p(X, M) = p(X)p(M|X)$$

θ: parameter of the data distribution $p(X)$
ϕ: parameter of the missing-data mechanism $p(M|X)$

Classical likelihood (in the classical case):

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full}}(\theta, \phi; X, M)$$

Observed likelihood (when there is missing values):

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)}$$

$$\hat{\theta}, \hat{\phi} \in \text{argmax}_{\theta,\phi} L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M)$$

Ignorability

- Parameters: $\theta, \phi \in \Omega_{\theta,\phi}$
- Ignorability: under MAR, $p(M|X;\phi)$ can be ignored

✍ Proof

$$
\begin{aligned}
& L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) \\
&= \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)} \\
&= \int p(X;\theta) p(M|X;\phi) dX_{\text{mis}(M)} \\
&= p(M|X_{\text{obs}(M)};\phi) \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if M(C)AR} \\
&\propto \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if } \phi \text{ nuisance parameter, if } \Omega_{\theta\times\phi} = \Omega_\theta \times \Omega_\phi
\end{aligned}
$$

Ignorability

- Parameters: $\theta, \phi \in \Omega_{\theta,\phi}$
- Ignorability: under MAR, $p(M|X;\phi)$ can be ignored

✍ Proof

$$
\begin{aligned}
& L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) \\
&= \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)} \\
&= \int p(X;\theta) p(M|X;\phi) dX_{\text{mis}(M)} \\
&= p(M|X_{\text{obs}(M)};\phi) \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if M(C)AR} \\
&\propto \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if } \phi \text{ nuisance parameter, if } \Omega_{\theta\times\phi} = \Omega_\theta \times \Omega_\phi
\end{aligned}
$$

Ignorability

- Parameters: $\theta, \phi \in \Omega_{\theta,\phi}$
- Ignorability: under MAR, $p(M|X;\phi)$ can be ignored

✍ Proof

$$
\begin{aligned}
& L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) \\
&= \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)} \\
&= \int p(X;\theta) p(M|X;\phi) dX_{\text{mis}(M)} \\
&= p(M|X_{\text{obs}(M)};\phi) \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if M(C)AR} \\
&\propto \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if } \phi \text{ nuisance parameter, if } \Omega_{\theta\times\phi} = \Omega_\theta \times \Omega_\phi
\end{aligned}
$$

Ignorability

- Parameters: $\theta, \phi \in \Omega_{\theta,\phi}$
- Ignorability: under MAR, $p(M|X;\phi)$ can be ignored

✍ Proof

$$
\begin{aligned}
& L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) \\
&= \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)} \\
&= \int p(X;\theta) p(M|X;\phi) dX_{\text{mis}(M)} \\
&= p(M|X_{\text{obs}(M)};\phi) \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if M(C)AR} \\
&\propto \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if } \phi \text{ nuisance parameter, if } \Omega_{\theta\times\phi} = \Omega_\theta \times \Omega_\phi
\end{aligned}
$$

Ignorability

- Parameters: $\theta, \phi \in \Omega_{\theta,\phi}$
- Ignorability: under MAR, $p(M|X;\phi)$ can be ignored

✍ Proof

$$
\begin{aligned}
& L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) \\
&= \int L_{\text{full}}(\theta, \phi; X, M) dX_{\text{mis}(M)} \\
&= \int p(X;\theta) p(M|X;\phi) dX_{\text{mis}(M)} \\
&= p(M|X_{\text{obs}(M)};\phi) \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if M(C)AR} \\
&\propto \int p(X;\theta) dX_{\text{mis}(M)} \qquad \text{if } \phi \text{ nuisance parameter, if } \Omega_{\theta\times\phi} = \Omega_\theta \times \Omega_\phi
\end{aligned}
$$

Biased results when ignoring MNAR

- Consider a bivariate Gaussian variable $X \sim \mathcal{N}(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} 5 \\ -1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

.

- Missing values in $X_{\cdot 2}$ (30% MCAR or MNAR)
- Maximum likelihood estimates of the mean of $X_{\cdot 2}$ **if the missing-data mechanism is ignored**:

MCAR	MNAR
-1.002	-1.570

$$\hat{\mu}, \hat{\Sigma} \in \text{argmax}_{\mu,\Sigma} \int p(X; \mu, \Sigma) dX_{\text{mis}(M)}$$

Summary

Missing-data mechanism

- MCAR: $p(M|X) = p(M)$
- MAR: $p(M|X) = p(M|X_{\text{obs}(M)})$
- MNAR: $p(M|X) = p(M|X_{\text{obs}(M)}, X_{\text{mis}(M)})$

	MCAR & MAR	MNAR
	Ignorable	**Not-ignorable**
Modelisation	$p(X)$	$p(X)p(M\|X)$

First, we will focus on MCAR & MAR data.

Summary

Missing-data mechanism

- MCAR: $p(M|X) = p(M)$
- MAR: $p(M|X) = p(M|X_{\text{obs}(M)})$
- MNAR: $p(M|X) = p(M|X_{\text{obs}(M)}, X_{\text{mis}(M)})$

	MCAR & MAR	MNAR
	Ignorable	**Not-ignorable**
Modelisation	$p(X)$	$p(X)p(M\|X)$

First, we will focus on MCAR & MAR data.

Plan

Different learning tasks

Imputation

= predict the missing values to get a **complete dataset**

Estimation

= **estimate a quantity** of interest (e.g. a covariance matrix) despite missing values

Prediction

= from a new observation, **predict a target variable**

Different learning tasks

Imputation

= predict the missing values to get a **complete dataset**

Estimation

= **estimate a quantity** of interest (e.g. a covariance matrix) despite missing values

Prediction

= from a new observation, **predict a target variable**

Different learning tasks

Imputation

= predict the missing values to get a **complete dataset**

Estimation

= **estimate a quantity** of interest (e.g. a covariance matrix) despite missing values

Prediction

= from a new observation, **predict a target variable**

Naive imputation methods

- **Mean imputation:** impute by the mean of the observed values in each variable.
- **k-nearest neighbors imputation:** (Troyanskaya et al., 2001) For each missing individual: (i) find the k-nearest neighbors and (ii) predict the missing values with the mean value from the k-nearest neighbors

$$d(X_{i.}, X_{\ell.}) = \sqrt{\frac{d}{|S_{i\ell}|} \sum_{j \in S_{i\ell}} (X_{ij} - X_{\ell j})^2},$$

with $S_{i\ell}$ = set of variables for which i and ℓ are observed

Naive imputation methods

- **Mean imputation:** impute by the mean of the observed values in each variable.
- **k-nearest neighbors imputation:** (Troyanskaya et al., 2001) For each missing individual: (i) find the k-nearest neighbors and (ii) predict the missing values with the mean value from the k-nearest neighbors

$$d(X_{i.}, X_{\ell.}) = \sqrt{\frac{d}{|S_{i\ell}|} \sum_{j \in S_{i\ell}} (X_{ij} - X_{\ell j})^2},$$

with $S_{i\ell}$ = set of variables for which i and ℓ are observed

Towards more complex methods

Joint or conditional model

Joint model

Specify $p(X)$ and derive $p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$

- PCA methods, Generative (deep) learning methods

Conditional model

Directly specify $p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$

- Iterative methods

Towards more complex methods

Joint or conditional model

Joint model

Specify $p(X)$ and derive $p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$

- PCA methods, Generative (deep) learning methods

Conditional model

Directly specify $p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$

- Iterative methods

Iterative imputation

Iterative methods: directly specify the conditional models

$$p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$$

- Initial **naive** imputation
- Repeat until convergence, iterating through variables
  - **Learn a prediction model** of the variable X_j on the other $X_{\cdot -j}$, using the individuals for which X_j is observed
  - **Predict the missing values** of X_j

The model can be...

- a random forest: `missForest` (Stekhoven and Bühlmann, 2012)
- a stochastic regression: `mice` (Van Buuren and Groothuis-Oudshoorn, 2011)

Iterative imputation

Iterative methods: directly specify the conditional models

$$p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$$

- Initial **naive** imputation
- Repeat until convergence, iterating through variables
  - **Learn a prediction model** of the variable $X_{.j}$ on the other $X_{.-j}$, using the individuals for which $X_{.j}$ is observed
  - **Predict the missing values** of $X_{.j}$

The model can be...

- a random forest: `missForest` (Stekhoven and Bühlmann, 2012)
- a stochastic regression: `mice` (Van Buuren and Groothuis-Oudshoorn, 2011)

Iterative imputation

Iterative methods: directly specify the conditional models

$$p(X_{\text{mis}(M)}|X_{\text{obs}(M)})$$

- Initial **naive** imputation
- Repeat until convergence, iterating through variables
  - **Learn a prediction model** of the variable $X_{.j}$ on the other $X_{.-j}$, using the individuals for which $X_{.j}$ is observed
  - **Predict the missing values** of $X_{.j}$

The model can be...

- a random forest: `missForest` (Stekhoven and Bühlmann, 2012)
- a stochastic regression: `mice` (Van Buuren and Groothuis-Oudshoorn, 2011)

Iterative imputation

Example

Matrix with NA

$$\begin{pmatrix} 30 & \text{NA} & 61 \\ \text{NA} & \text{NA} & 50 \\ 25 & 47 & \text{NA} \end{pmatrix}$$

Initialization with mean imputation

$$\begin{pmatrix} 30 & 40 & 61 \\ 27 & 40 & 50 \\ 25 & 47 & 55 \end{pmatrix}$$

Learn a model $X_{\text{obs}(M_{\cdot 1})1} | X_{\text{obs}(M_{\cdot 1})2}, X_{\text{obs}(M_{\cdot 1})3}$

$$\begin{pmatrix} 30 & 40 & 61 \\ \text{NA} & 40 & 50 \\ 25 & 47 & 55 \end{pmatrix}$$

Predict the missing values of $X_{\cdot 1}$ using this model

$$\begin{pmatrix} 30 & 40 & 61 \\ 29 & 40 & 50 \\ 25 & 47 & 55 \end{pmatrix}$$

Likelihood-based methods

If NA's are MCAR or MAR: **observed likelihood**

$$L_{\text{obs}}(\theta; X_{\text{obs}(M)}) = \int p(X; \theta) dX_{\text{mis}(M)} \quad \text{[ignorability]}$$

Dual objectives:

- Estimating the parameter θ of the data distribution
- Predicting missing values

Imputation

- Find an estimator $\hat{\theta}$
- Use the conditional distribution if tractable

$$X_{\text{imputed}} \sim p(X_{\text{mis}(M)} | X_{\text{obs}(M)}; \hat{\theta})$$

Likelihood-based methods

If NA's are MCAR or MAR: **observed likelihood**

$$L_{\text{obs}}(\theta; X_{\text{obs}(M)}) = \int p(X; \theta) dX_{\text{mis}(M)} \quad \text{[ignorability]}$$

Dual objectives:

- Estimating the parameter θ of the data distribution
- Predicting missing values

Imputation

- Find an estimator $\hat{\theta}$
- Use the conditional distribution if tractable

$$X_{\text{imputed}} \sim p(X_{\text{mis}(M)} | X_{\text{obs}(M)}; \hat{\theta})$$

Low-rank methods

- Low-rank approximation :

$$\underbrace{\mathbb{X}}_{\text{data matrix}} = \underbrace{\Theta}_{\text{low-rank matrix}} + \underbrace{\epsilon}_{\text{noise}}$$

Typically: $\epsilon \sim \mathcal{N}(0_d, \sigma^2 I_{d\times d})$

- **Predict missing values in $\mathbb{X}$ = estimate Θ**
- Individual profiles can be summarized into a limited number of general profiles
- `softImpute` (Hastie and Mazumder, 2015)
- `missMDA` (Josse and Husson, 2016)

Low-rank methods

- Low-rank approximation :

$$\underbrace{\mathbb{X}}_{\text{data matrix}} = \underbrace{\Theta}_{\text{low-rank matrix}} + \underbrace{\epsilon}_{\text{noise}}$$

Typically: $\epsilon \sim \mathcal{N}(0_d, \sigma^2 I_{d\times d})$

- **Predict missing values in $\mathbb{X}$ = estimate Θ**
- Individual profiles can be summarized into a limited number of general profiles
- `softImpute` (Hastie and Mazumder, 2015)
- `missMDA` (Josse and Husson, 2016)

Low-rank methods

- Low-rank approximation :

$$\underbrace{\mathbb{X}}_{\text{data matrix}} = \underbrace{\Theta}_{\text{low-rank matrix}} + \underbrace{\epsilon}_{\text{noise}}$$

  Typically: $\epsilon \sim \mathcal{N}(0_d, \sigma^2 I_{d\times d})$
- **Predict missing values in $\mathbb{X}$ = estimate Θ**
- Individual profiles can be summarized into a limited number of general profiles
- `softImpute` (Hastie and Mazumder, 2015)
- `missMDA` (Josse and Husson, 2016)

Low-rank methods

- Low-rank approximation :

$$\underbrace{\mathbb{X}}_{\text{data matrix}} = \underbrace{\Theta}_{\text{low-rank matrix}} + \underbrace{\epsilon}_{\text{noise}}$$

  Typically: $\epsilon \sim \mathcal{N}(0_d, \sigma^2 I_{d\times d})$
- **Predict missing values in $\mathbb{X}$ = estimate Θ**
- Individual profiles can be summarized into a limited number of general profiles
- `softImpute` (Hastie and Mazumder, 2015)
- `missMDA` (Josse and Husson, 2016)

Low-rank methods

- Low-rank approximation :

$$\underbrace{\mathbb{X}}_{\text{data matrix}} = \underbrace{\Theta}_{\text{low-rank matrix}} + \underbrace{\epsilon}_{\text{noise}}$$

  Typically: $\epsilon \sim \mathcal{N}(0_d, \sigma^2 I_{d\times d})$
- **Predict missing values in $\mathbb{X}$ = estimate Θ**
- Individual profiles can be summarized into a limited number of general profiles
- `softImpute` (Hastie and Mazumder, 2015)
- `missMDA` (Josse and Husson, 2016)

Change of basis with SVD

Truncated Singular Value Decomposition (SVD) with rank $\mathbf{r}$

$$\mathrm{SVD}_{\mathbf{r}}(\mathbb{X}) = \underbrace{\left[\begin{array}{c|c|c} \mathbf{U}_1 & \cdots & \mathbf{U}_{\mathbf{r}} \end{array}\right]}_{\text{new coordinates}} \underbrace{\begin{bmatrix} \sigma_1(\mathbb{X}) & 0 & \cdots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sigma_{\mathbf{r}}(\mathbb{X}) \end{bmatrix}}_{\text{singular values}} \underbrace{\begin{bmatrix} \mathbf{V}_1^\top \\ \vdots \\ \mathbf{V}_{\mathbf{r}}^\top \end{bmatrix}}_{\text{new basis}}$$

$\mathbb{R}^{n \times d}$ $\mathbb{R}^{n \times \mathbf{r}}$ $\mathbb{R}^{\mathbf{r} \times \mathbf{r}}$ $\mathbb{R}^{\mathbf{r} \times d}$

Iterative PCA with NA

Algorithm Iterative Principal Component Analysis (simplest version)

Input: **r** (number of kept dimensions)
Initialisation: $X^{(0)}$ (mean imputation)
Iteratively do:

Estimation step:
Compute the SVD with **r** dim. on the completed dataset:

$$\mathbf{SVD_r}(X^{(t)}) = U_{\mathbf{r}} D_{\mathbf{r}} V_{\mathbf{r}}^t,$$

Imputation step:
Predict the missing values with the SVD

$$X^{(t+1)} = \underbrace{X \odot (\mathbf{1}_{n\times d} - M)}_{\text{on observed values}} + \underbrace{\mathbf{SVD_r}(X^{(t)}) \odot M}_{\text{on missing values}}$$

Plan

Introduction on missing data

Informative missing values

Classical methods for handling missing values

Generative (deep) methods

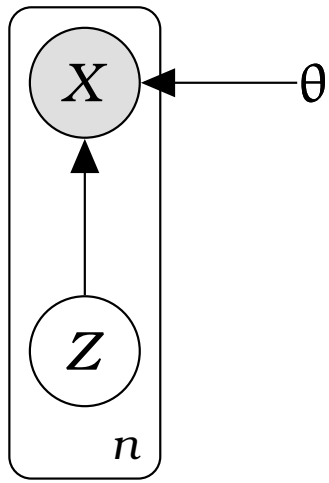
Methods for informative missing values

Change of basis with deep learning

Deep latent variable model

$$\begin{cases} Z \sim p(Z) \\ X \sim p_\theta(X|Z) = \Phi(X|f_\theta(Z)) \quad \text{(decoder)} \end{cases}$$

Consider $\Phi(.;\tau)$ the Gaussian distribution of parameter τ.
Typically: $f_\theta(Z)$ is a deep neural network



DLVM with NA: MIWAE (Mattei and Frellsen, 2019)

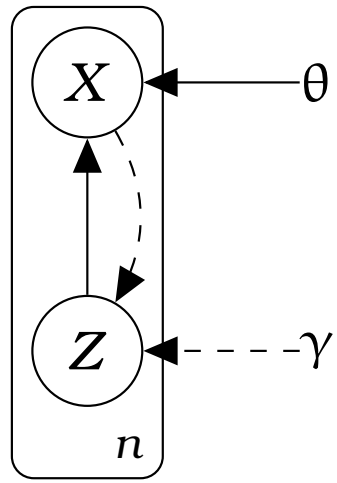
$$\begin{cases} z \sim q_\gamma(z|x_{\text{obs}(m)}) = \Phi(z|g_\gamma(x_{\text{obs}(m)})) & \text{(encoder)} \\ x \sim p_\theta(x|z) = \Phi(x_{\text{obs}(m)}|f_\theta(z)) & \text{(decoder)} \end{cases}$$

$g_\gamma(x)$ is also a deep neural network

DLVM with NA: MIWAE (Mattei and Frellsen, 2019)

$$\begin{cases} z \sim q_\gamma(z|x_{\text{obs}(m)}) = \Phi(z|g_\gamma(x_{\text{obs}(m)})) & \text{(encoder)} \\ x \sim p_\theta(x|z) = \Phi(x_{\text{obs}(m)}|f_\theta(z)) & \text{(decoder)} \end{cases}$$

$g_\gamma(x)$ is also a deep neural network



Impute with DLVM (Mattei and Frellsen, 2019)

$$
\begin{aligned}
&\mathbb{E}[X_{\text{mis}(M)}|X_{\text{obs}(M)}] \\
&= \int X_{\text{mis}(M)} p_\theta(X_{\text{mis}(M)}|X_{\text{obs}(M)})\, dX_{\text{mis}(M)} \\
&= \int\int X_{\text{mis}(M)} p_\theta(X_{\text{mis}(M)}|X_{\text{obs}(M)}, Z)\, p_\theta(Z|X_{\text{obs}(M)})\, dZ dX_{\text{mis}(M)} \\
&= \int\int X_{\text{mis}(M)} \frac{p_\theta(Z|X_{\text{obs}(M)})}{q_\gamma(Z|X_{\text{obs}(M)})} p_\theta(X_{\text{mis}(M)}|X_{\text{obs}(M)}, Z)\, q_\gamma(Z|X_{\text{obs}(M)})\, dZ dX_{\text{mis}(M)}
\end{aligned}
$$

Estimate with self-normalised importance sampling:

$$
\sum_{l=1}^{L} w_l X_{\text{mis}(M)}^{(l)}, \quad \text{(Section 9.2 of (?))}
$$

with $w_l = \frac{r_l}{\sum_{l=1}^{L} r_l}$, $r_l = \frac{p_\theta(X_{\text{obs}(M)}|Z^{(l)})p(Z^{(l)})}{q_\gamma(Z^{(l)}|X_{\text{obs}(M)})}$, and

$(X_{\text{mis}(M)}^{(l)}, Z^{(l)}) \overset{\text{iid}}{\sim} p_\theta(X_{\text{mis}(M)}|X_{\text{obs}(M)}, Z)\, q_\gamma(Z|X_{\text{obs}(M)})$

Plan

Introduction on missing data

Informative missing values

Classical methods for handling missing values

Generative (deep) methods

Methods for informative missing values

Challenge with MNAR data

Here, we have to model $p(X, M)$

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int p(X, M; \theta, \phi) dX_{\text{mis}(M)}$$

with $p(X, M; \theta, \phi) = p(X; \theta)\ p(M|X; \phi)$.

How to estimate ϕ ?

Challenge with MNAR data

Here, we have to model $p(X, M)$

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int p(X, M; \theta, \phi) dX_{\text{mis}(M)}$$

with $p(X, M; \theta, \phi) = p(X; \theta)\ p(M|X; \phi)$.

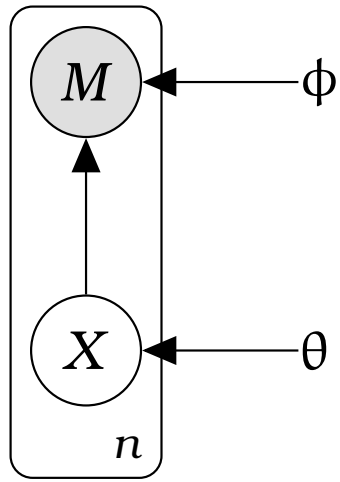
How to estimate ϕ ?

Challenge with MNAR data

Here, we have to model $p(X, M)$

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int p(X, M; \theta, \phi) dX_{\text{mis}(M)}$$

with $p(X, M; \theta, \phi) = p(X; \theta)\ p(M|X; \phi)$.



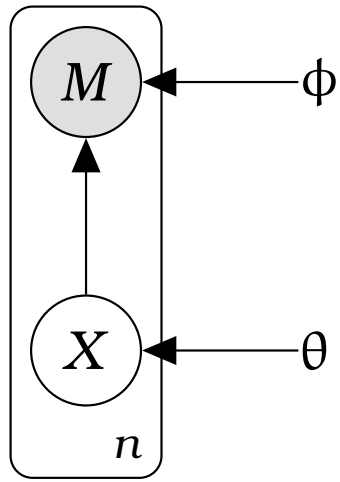
How to estimate ϕ ?

Challenge with MNAR data

Here, we have to model $p(X, M)$

$$L_{\text{full,obs}}(\theta, \phi; X_{\text{obs}(M)}, M) = \int p(X, M; \theta, \phi) dX_{\text{mis}(M)}$$

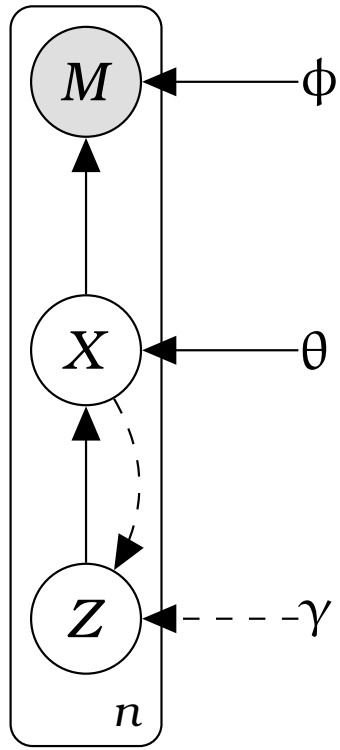
with $p(X, M; \theta, \phi) = p(X; \theta)\ p(M|X; \phi)$.



How to estimate ϕ ?

Generative methods for MNAR data

- For DLVMs: not-MIWAE (Ipsen et al., 2022)



Generative methods for MNAR data

- For DLVMs: not-MIWAE (Ipsen et al., 2022)

For MCAR or MAR data:

$$\mathbb{E}_{z_{i1},\dots,z_{iK}\sim q_\gamma(z|x_{\text{obs}(M)})}\left[\log\frac{1}{K}\sum_{k=1}^{K}\frac{p(x_{\text{obs}(M)}|z_{ik})p(z_{ik})}{q_\gamma(z_{ik}|x_{\text{obs}(M)})}\right]$$

For MNAR data:

$$\mathbb{E}_{(z_1,x_1^m),\dots,(z_K,x_K^m)}\left[\log\frac{1}{K}\sum_{k=1}^{K}\frac{p_\phi(M|x_{\text{obs}(M)},x_k^m)p(x_{\text{obs}(M)}|z_{ik})p(z_{ik})}{q_\gamma(z_{ik}|x_{\text{obs}(M)})}\right]$$

with $p_\phi(M|x_{\text{obs}(M)},x_k^m)=\text{Bernoulli}(M|\pi_\phi(x_{\text{obs}(M)},,x_k^m))$.

Form of the MNAR model?

- logit self-masked: $(\pi_\phi(x_{\text{obs}(M)},x_k^m))_j=1/(1+e^{-(ax_j+b)})$
- but can be more general...

Conclusion

- More **effective methods** than classical mean imputation or KNN-imputation exist at a **low cost**.
- **Generative (deep) models** can be used to fill in missing data.
- Choosing the right imputation method **depends on the nature of the missing data** and on **the learning task**.

Conclusion

- More **effective methods** than classical mean imputation or KNN-imputation exist at a **low cost**.
- **Generative (deep) models** can be used to fill in missing data.
- Choosing the right imputation method **depends on the nature of the missing data** and on **the learning task**.

Conclusion

- More **effective methods** than classical mean imputation or KNN-imputation exist at a **low cost**.
- **Generative (deep) models** can be used to fill in missing data.
- Choosing the right imputation method **depends on the nature of the missing data** and on **the learning task**.

References I

- Hastie, T. and Mazumder, R. (2015). softimpute: Matrix completion via iterative soft-thresholded svd. *R package version 1.4*.
- Ipsen, N. B., Mattei, P.-A., and Frellsen, J. (2022). not-miwae: Deep generative modelling with missing not at random data. *ICLR*.
- Josse, J. and Husson, F. (2016). missmda: a package for handling missing values in multivariate data analysis. *Journal of statistical software*, 70:1–31.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Mattei, P.-A. and Frellsen, J. (2019). Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR.

References II

- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

References III

- Zhu, Z., Wang, T., and Samworth, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031.