What’s a good imputation to predict with missing values?

Marine Le Morvan – Soda, INRIA
Incomplete data is ubiquitous in many fields.

Sources of missingness:

- Survey nonresponse.
- Sensor failure.
- Changing data gathering procedure.
- Database join.
- ...

Missing data is frequent in economics, social, political or health sciences.

Traumabase clinical health records.
Since the 70s, an abundant literature on missing data has flourished.

- Missing data mechanisms are usually divided into 3 categories:
  - MCAR (Missing Completely at Random)
  - MAR (Missing at Random)
  - MNAR (Missing Non At Random)

- The literature has been mainly focused on inference and imputation tasks:
  - Likelihood based methods under MAR.
  - Multiple imputation under MAR.
  - Inverse probability weighting under MAR.

But very few works have addressed supervised learning with missing values, whatever the missing data mechanism.
Challenges of supervised learning with missing values

- Supervised learning with missing values:
  - Scarce literature (vs Inference, Imputation)
Challenges of supervised learning with missing values

- Supervised learning with missing values:
  - Scarce literature (vs Inference, Imputation)
  - Arbitrary subsets of variables.

Samples

\[
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3
\end{bmatrix}
\]
Challenges of supervised learning with missing values

- Supervised learning with missing values:
  - Scarce literature (vs Inference, Imputation)
  - Arbitrary subsets of variables.

Samples

\[
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3 \\
0.1 & \text{na} & 0.2 & 1.4
\end{bmatrix}
\]

Diagram:

- $x_1 ightarrow 0.1$
- na $ightarrow$
- $x_3 ightarrow 0.2$
- $x_4 ightarrow 1.4$
- $y$
Challenges of supervised learning with missing values

- Supervised learning with missing values:
  
  • Scarce literature (vs Inference, Imputation)
  
  • Arbitrary subsets of variables.

\[
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3 \\
0.1 & \text{na} & 0.2 & 1.4 \\
\text{na} & 0.7 & 1.2 & \text{na}
\end{bmatrix}
\]

\[
\text{na} \rightarrow \text{red circle} \rightarrow 0.7 \rightarrow 1.2 \rightarrow y
\]
Supervised learning with missing values:

- Scarce literature (vs Inference, Imputation)
- Arbitrary subsets of variables.
- Computational and statistical challenge.

Ex: \( d = 50 \implies 2^{50} \approx 10^{15} \) possible missing data patterns.

Samples

\[
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3 \\
0.1 & \text{na} & 0.2 & 1.4 \\
\text{na} & 0.7 & 1.2 & \text{na}
\end{bmatrix}
\]
Challenges of supervised learning with missing values

- Supervised learning with missing values:
  - Scarce literature (vs Inference, Imputation)
  - Arbitrary subsets of variables.
  - Computational and statistical challenge.  
    Ex: \( d = 50 \implies 2^{50} \approx 10^{15} \) possible missing data patterns.
  - Widespread current practice: Impute-then-Regress. Very little theoretical foundation.

\[
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3 \\
0.1 & \text{na} & 0.2 & 1.4 \\
\text{na} & 0.7 & 1.2 & \text{na}
\end{bmatrix}
\xrightarrow{\text{Impute}}
\begin{bmatrix}
1.1 & 0.2 & 0.5 & 1.3 \\
0.1 & 0.5 & 0.2 & 1.4 \\
0.6 & 0.7 & 1.2 & 1.3
\end{bmatrix}
\xrightarrow{\text{Regress}}
\begin{bmatrix}
\text{Predictions } y
\end{bmatrix}
\]
Impute-then-Regress procedures

Questions

▶ Can Impute-then-Regress procedures be Bayes optimal?
▶ How should we choose the imputation function?
▶ What if the data is Missing Not At Random (MNAR)?

Define Impute-then-Regress procedures as functions of the form:

\[ g \circ \Phi, \text{ where } \Phi \in \mathcal{F}^I, \ g : \mathbb{R}^d \mapsto \mathbb{R}. \]

where imputation functions \( \Phi \in \mathcal{F}^I \) are of the form:

\[ \phi_1^{(m)}(x_2, x_3) \]
\[ \phi_4^{(m)}(x_2, x_3) \]
Formalizing the problem

- **Assumption** - The response $Y$ is a function of the (unavailable) complete data plus some noise:

  \[ Y = f^*(X) + \epsilon, \quad X \in \mathbb{R}^d, \quad Y \in \mathbb{R}. \]

- Optimization problem:

  \[
  \min_{f: (\mathbb{R} \cup \{NA\})^d \mapsto \mathbb{R}} \mathcal{R}(f) := \mathbb{E} \left[ (Y - f(\tilde{X}))^2 \right]
  \]

- A Bayes predictor is a minimizer of the risk. It is given by:

  \[
  \tilde{f}^*(\tilde{X}) := \mathbb{E} \left[ Y | X_{\text{obs}}(M), M \right] = \mathbb{E} \left[ f(X) | X_{\text{obs}}(M), M \right]
  \]

  where $M \in \{0, 1\}^d$ is the missingness indicator.

  The Bayes rate $\mathcal{R}^*$ is the risk of the Bayes predictor: $\mathcal{R}^* = \mathcal{R}(\tilde{f}^*)$.

  A Bayes optimal function $f$ achieves the Bayes rate, i.e., $\mathcal{R}(f) = \mathcal{R}^*$.
Can Impute-then-Regress procedures be Bayes optimal?

Yes, they can!
In fact, they almost always are...

Theorem (Bayes optimality of Impute-then-Regress procedures)
Let $g^* \Phi$ be the minimizer of the risk on the data imputed by $\Phi$. Assume that i) $\Phi \in F_I\infty$ ii) the response $Y$ is generated as $Y = f^*(X) + \epsilon$. Then, for:
• all missing data mechanisms,
• almost all imputation functions,
$g^* \Phi \circ \Phi$ is Bayes optimal.
In other words, for almost all imputation functions $\Phi \in F_I\infty$, a universally consistent algorithm trained on the imputed data $\Phi(\tilde{X})$ is Bayes consistent.
Can Impute-then-Regress procedures be Bayes optimal?

Yes, they can! In fact, they almost always are...
Can Impute-then-Regress procedures be Bayes optimal?

Yes, they can! In fact, they almost always are...

**Theorem (Bayes optimality of Impute-then-Regress procedures)**

Let $g^*_\Phi$ be the minimizer of the risk on the data imputed by $\Phi$. Assume that i) $\Phi \in \mathcal{F}_\infty^l$, ii) the response $Y$ is generated as $Y = f^*(X) + \epsilon$. Then, for:

- all missing data mechanisms,
- almost all imputation functions,

$g^*_\Phi \circ \Phi$ is Bayes optimal.

In other words, for almost all imputation functions $\Phi \in \mathcal{F}_\infty^l$, a universally consistent algorithm trained on the imputed data $\Phi(\tilde{X})$ is Bayes consistent.
Sketch of the proof: arguments from differential topology

Sketch of the proof:

1. All data points with a missing data pattern $m$ are mapped to a manifold $\mathcal{M}^{(m)}$ of dimension $|\text{obs}(m)|$ (Preimage Theorem).

![Complete data](image1.png)

![Imputed data (manifolds)](image2.png)
Sketch of the proof: arguments from differential topology

Sketch of the proof:

1. All data points with a missing data pattern $m$ are mapped to a manifold $\mathcal{M}^{(m)}$ of dimension $|\text{obs}(m)|$ (Preimage Theorem).

2. The missing data patterns of imputed data points can almost surely be de-identified (Thom transversality Theorem).

Complete data

Imputed data (manifolds)
Sketch of the proof: arguments from differential topology

Sketch of the proof:

1. All data points with a missing data pattern \( m \) are mapped to a manifold \( \mathcal{M}^{(m)} \) of dimension \( |\text{obs}(m)| \) (Preimage Theorem).

2. The missing data patterns of imputed data points can almost surely be de-identified (Thom transversality Theorem).

3. Given 2), we can exhibit a \( g^{\Phi}_{\star} \), which does not depend on \( m \), and which for each manifold equals the Bayes predictor except on a set of measure 0.
Which imputation function should one choose?

Why bother!
From now on I use constant imputations!

May be a good imputation would still provide an easier learning problem?

Question

Are there *continuous* Impute-then-Regress decompositions of Bayes predictors?

From now on, we suppose $f^*$ is smooth.
We will denote by $\Phi^{CI}$ the imputation by the conditional expectation.
Chaining oracles

**Question**

What is the risk of chaining oracles: $f^* \circ \Phi^{CI}$?

**Assumption (i):** there exists positive semi-definite matrices $\bar{H}^+$ and $\bar{H}^-$ such that for all $X \in S$, $\bar{H}^- \leq H(X) \leq \bar{H}^+$.

**Proposition (Non consistency of chaining oracles)**

Under assumption (i), the excess risk of chaining oracles compared to the Bayes risk $\mathcal{R}^*$ is upper-bounded by:

$$
\mathcal{R}(f^* \circ \Phi^{CI}) - \mathcal{R}^* \leq \frac{1}{4} \mathbb{E}_M \left[ \max \left( tr \left( \bar{H}^-_{mis,mis} \Sigma_{mis|obs,M} \right)^2, tr \left( \bar{H}^+_{mis,mis} \Sigma_{mis|obs,M} \right)^2 \right) \right]
$$
Chaining oracles

Question

What is the risk of chaining oracles: $f^* \circ \Phi^C_l$?

$$\mathcal{R}(f^* \circ \Phi^C_l) - \mathcal{R}^* \leq \frac{1}{4} \mathbb{E}_M \left[ \max \left( \text{tr} \left( \bar{H}_{mis,mis} \Sigma_{mis|obs,M} \right)^2, \text{tr} \left( \bar{H}_{mis,mis}^+ \Sigma_{mis|obs,M} \right)^2 \right) \right]$$

$f^*(w^\top X)$

High variance

Low curvature

$$\text{var}_{X_{mis|X_{obs}}} [w^\top X] = w_{mis}^\top \Sigma_{mis|obs} w_{mis}$$
Chaining oracles

Question

What is the risk of chaining oracles: $f^* \circ \Phi^{CI}$?

$$\mathcal{R}(f^* \circ \Phi^{CI}) - \mathcal{R}^* \leq \frac{1}{4} \mathbb{E}_M \left[ \max \left( tr \left( \bar{H}_{mis,mis} \Sigma_{mis|obs,M} \right)^2, tr \left( \bar{H}_{mis,mis} \Sigma_{mis|obs,M} \right)^2 \right) \right]$$

$f^*(w^\top X)$

Low variance

High curvature

$$\text{var}_{X_{mis}|X_{obs}}[w^\top X] = w_{mis}^\top \Sigma_{mis|obs} w_{mis}$$
Chaining oracles

**Question**

What is the risk of chaining oracles: $f^* \circ \Phi^{Cl}$?

$$\mathcal{R}(f^* \circ \Phi^{Cl}) - \mathcal{R}^* \leq \frac{1}{4} \mathbb{E}_M \left[ \max \left( \text{tr} \left( \bar{H}^-_{mis,mis} \Sigma_{mis|obs,M} \right)^2, \text{tr} \left( \bar{H}^+_{mis,mis} \Sigma_{mis|obs,M} \right)^2 \right) \right]$$

$$f^*(w^T X)$$

High variance

High curvature

$$\text{var}_{X_{mis} \mid X_{obs}} [w^T X] = w_{mis}^T \Sigma_{mis|obs} w_{mis}$$
Learning on conditionally imputed data

**Question**

What can we say about the optimal predictor on the conditionally imputed data: $g_{\Phi CI}^* \circ \Phi CI$?

**Proposition (Regression function discontinuities)**

Suppose that $f^* \circ \Phi CI$ is not Bayes optimal, and that the probability of observing all variables is strictly positive, i.e., for all $x$,

$$P(M = (0, \ldots, 0), X = x) > 0.$$  

Then there is no continuous function $g$ such that $g \circ \Phi CI$ is Bayes optimal.

Note: The size of the discontinuities are also controlled by the variance-curvature tradeoff.
If the predictor is fixed as $f^*$, can we find a continuous imputation function $\Phi$ so that $f^* \circ \Phi$ is Bayes optimal?
If the predictor is fixed as $f^*$, can we find a continuous imputation function $\Phi$ so that $f^* \circ \Phi$ is Bayes optimal?

Not always...
Question

If the predictor is fixed as $f^*$, can we find a continuous imputation function $\Phi$ so that $f^* \circ \Phi$ is Bayes optimal?

But sometimes yes!
Optimizing imputations for a fixed regression function

**Question**

If the predictor is fixed as \( f^* \), can we find a *continuous* imputation function \( \Phi \) so that \( f^* \circ \Phi \) is Bayes optimal?

**Proposition (Existence of continuous corrected imputations)**

Assume that \( f^* \) is uniformly continuous, twice continuously differentiable and that, for all missing patterns \( m \) and all \( x_{obs} \), the support of \( X_{mis} | X_{obs} = x_{obs}, M = m \) is connected.

Additionally, assume that for all missing patterns \( m \), and all \((x_{obs}, x_{mis})\), the gradient of \( f^* \) with respect to the missing coordinates is nonzero:

\[
\nabla_{x_{mis}} f^*(x_{obs}, x_{mis}) \neq 0.
\]

Then, for all \( m \), there exist continuous imputation functions \( \phi^*(m) : \mathbb{R}^{|obs(m)|} \rightarrow \mathbb{R}^{|mis(m)|} \) such that \( f^* \circ \Phi \) is Bayes optimal.

**Proof**: based on a Global Implicit Function theorem.
Simulations

- $X \sim \mathcal{N}(X|\mu, \Sigma)$ with two covariance settings: 'high' and 'low'.

- $Y = f^*(X) + \epsilon$.
  - Two settings: 'bowl' and 'wave'.
  - $\beta$ chosen so that $\beta^T X$ centered on 1 with variance 1.
  - Signal-to-noise ratio of 10.

- Two missing data mechanisms: MCAR and Gaussian self-masking (MNAR). 50% missing values.
Baseline methods benchmarked

**Oracles and semi-oracles**
- Bayes predictors.
- Chaining oracles: \( f^* \circ \Phi^{CI} \)
- Oracle Impute + MLP: Imputation by \( \Phi^{CI} \) followed by regression with a MultiLayer Perceptron.

**Impute-then-Regress predictors**
- Mean Impute + MLP
- MICE + MLP: MICE implements a conditional imputation, but only valid under MAR.
- Gradient-Boosted Regression Trees: with Missing Incorporated Attribute strategy.
- NeuMiss + MLP

We also try concatenating the mask after mean or MICE imputation to help handle the MNAR case.
NeuMiss: a neural network for missing values

Input data:
missing values are replaced by zeros.

The NeuMiss layer:
Fixed width of size $d$. Shared weights. Residual connections. New non-linearity: $\odot m$.

Conventional feedforward NN

Theoretically grounded: differentiable approximation of the conditional expectation.

Impute-then-Regress architecture.

Non-linearity $\odot m$
Gaussian data hypothesis: $X \sim \mathcal{N}(X|\mu, \Sigma)$

Conditional expectation:

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

Approximation of $(\Sigma_{obs})^{-1}$ by a truncated Neumann series:

$$(\Sigma_{obs})^{-1} = \frac{1}{L} \sum_{k=0}^{\infty} (Id_{obs} - \frac{1}{L} \Sigma_{obs})^k$$

Order-$\ell$ approximation of $(\Sigma_{obs})^{-1}$ (for any $\text{obs}$):

$$S_{obs}^{(\ell)} = (Id_{obs} - \frac{1}{L} \Sigma_{obs}) S_{obs}^{(\ell-1)} + \frac{1}{L} Id.$$
NeuMiss: a neural network for missing values

- Gaussian data hypothesis: \( X \sim \mathcal{N}(X|\mu, \Sigma) \)

- Conditional expectation:

\[
\mathbb{E}[X_{mis}|X_{obs}] = \mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs})
\]

- Approximation of \((\Sigma_{obs})^{-1}\) by a truncated Neumann series:

\[
(\Sigma_{obs})^{-1} = \frac{1}{L} \sum_{k=0}^{\infty} (Id_{obs} - \frac{1}{L} \Sigma_{obs})^k
\]

- Order-\(\ell\) approximation of \((\Sigma_{obs})^{-1}\) (for any obs):

\[
S^{(\ell)}_{obs}(x_{obs} - \mu_{obs}) = (Id_{obs} - \frac{1}{L} \Sigma_{obs}) S^{(\ell-1)}_{obs}(x_{obs} - \mu_{obs}) + \frac{1}{L} (x_{obs} - \mu_{obs}).
\]
NeuMiss: a neural network for missing values

\[
\begin{align*}
S^{(\ell)}_{obs}(x_{obs} - \mu_{obs}) &= \left( I_{d_{obs}} - \frac{1}{L} \sum_{obs} \right) S^{(\ell-1)}_{obs}(x_{obs} - \mu_{obs}) + \frac{1}{L} (x_{obs} - \mu_{obs}). 
\end{align*}
\]
Experimental results

**Chaining oracles**
- Oracle impute + MLP
- NeuMiss + MLP
- MICE + MLP
- MICE & mask + MLP
- mean impute + MLP
- mean impute & mask + MLP
- Gradient-boosted trees

**MCAR**

**MNAR**
- Gaussian
- Self masking

Drop in R2 compared to Bayes predictor
Takeaway

- A theoretical foundation for Impute-then-Regress procedures
  Impute-then-Regress procedures are Bayes optimal for all missing data mechanisms and almost all imputation functions.

- NeuMiss + MLP: a powerful predictor in the presence of missing values.

Thank you for your attention!