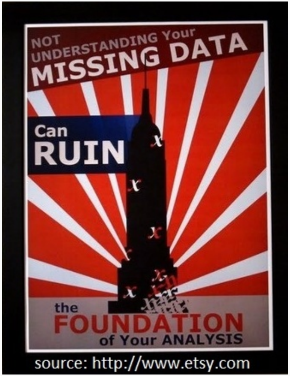
Supervised learning with missing values

Julie Josse
INRIA, Ecole Polytechnique

25 May 2022

Emerging Challenges for Statistics and Data Sciences: Complex Data with Missingness, Measurement Errors, and High Dimensionality 2022



Introduction

Collaborators on supervised learning with missing values

- M. Le Morvan, Junior researcher at INRIA, Paris. Topic: supervised learning.
- E. Scornet, Associate Professor at Ecole Polytechnique, IP Paris.
Topic: random forests.
- G. Varoquaux, Senior researcher at INRIA, Paris.
Topic: machine learning. Creator of Scikit-learn in python.

1. *Consistency of supervised learning with missing values. (2019). Revis.*

2. *Linear predictor on linearly-generated data with missing values: non consistency and solutions. AISTAT2020.*

3. *Neumiss networks: differential programming for supervised learning with missing values. Neurips2020 (Oral).*

4. *What's a good imputation to predict with missing values? Neurips2021 (Oral).*

Traumabase project: decision support for trauma patients

- 30000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 30 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

Traumabase project: decision support for trauma patients

- 30000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 30 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

⇒ **Estimate causal effect**: Administration of the **treatment** "tranexamic acid" (within 3 hours after the accident) on the **outcome** mortality for traumatic brain injury patients. [1]

[1] Mayer, Wager, J. Doubly robust treatment effect estimation with incomplete confounders. *Annals Of Applied Statistics*. 2020.

Traumabase project: decision support for trauma patients

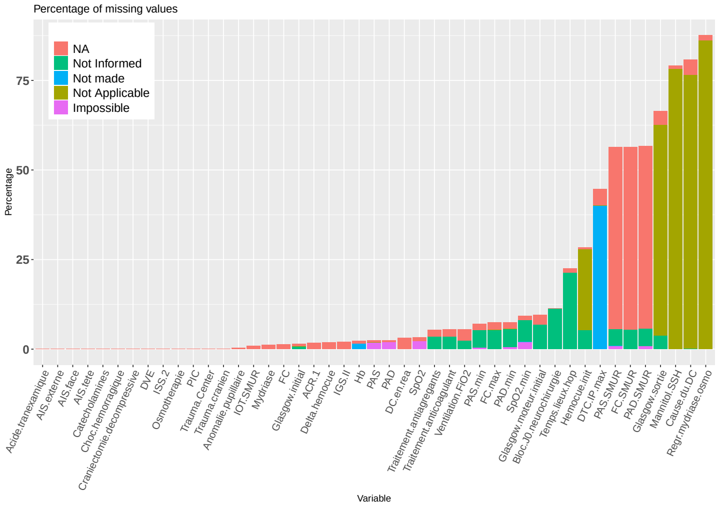
- 30000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 30 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

⇒ **Predict** platelet levels given pre-hospital features

Ex linear regression/ random forests with covariates with missing values

Missing values



Different pattern: sporadic & systematic (missing variable in one hospital)

Different types: informative, non informative

Solutions to handle missing values (in the covariates)

Abundant literature: Rmistatic platform, more than 150 packages

Maximum likelihood (EM + Supplemented EM algorithms): modify the estimation process to deal with missing values

Pros: Tailored toward a specific problem

Cons: One specific algorithm for each statistical method...

Difficult to establish - not many softwares even for simple models [1]

Multiple imputation to get a complete data set

Pros: Any analysis can be performed - mice package

Cons: Generic - Computational issues for large dimensions

[1] Jiang, J. et al. 2019. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*.

Solutions to handle missing values (in the covariates)

Abundant literature: Rmistatic platform, more than 150 packages

Maximum likelihood (EM + Supplemented EM algorithms): modify the estimation process to deal with missing values

Pros: Tailored toward a specific problem

Cons: One specific algorithm for each statistical method...

Difficult to establish - not many softwares even for simple models [1]

Multiple imputation to get a complete data set

Pros: Any analysis can be performed - mice package

Cons: Generic - Computational issues for large dimensions

Inferential aim: **Estimate parameters & their variance**

Three missing data mechanisms: MCAR, MAR, MNAR

[1] Jiang, J. et al. 2019. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*.

Solutions to handle missing values (in the covariates)

Abundant literature: Rmistatic platform, more than 150 packages

Maximum likelihood (EM + Supplemented EM algorithms): modify the estimation process to deal with missing values

Pros: Tailored toward a specific problem

Cons: One specific algorithm for each statistical method...

Difficult to establish - not many softwares even for simple models [1]

Multiple imputation to get a complete data set

Pros: Any analysis can be performed - mice package

Cons: Generic - Computational issues for large dimensions

Inferential aim: **Estimate parameters & their variance**

Three missing data mechanisms: MCAR, MAR, MNAR

Few works on supervised learning with missing values, no theoritical results, whatever the missing data mechanism

[1] Jiang, J. et al. 2019. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*.

Notations

- Random Variables:
  - $X \in \mathbb{R}^d$: the complete unvailable data
  - $\widetilde{X} \in \{\mathbb{R} \cup \{\mathrm{NA}\}\}^d$: incomplete data (observed), NA: Not Available
  - $M \in \{0,1\}^d$: the missing-data pattern, the mask

$obs(M)$ (resp. $\mathrm{mis}(M)$) indices of the observed (resp. missing) entries.

- Realizations:

$$x = (1.1, 2.3, 3.1, 8, 5.27)$$

$$\widetilde{x} = (1.1, \mathrm{NA}, -3.1, 8, \mathrm{NA})$$

$$m = (0, 1, 0, 0, 1)$$

$$x_{\mathrm{obs(m)}} = (1.1, 3.1, 8), \qquad x_{\mathrm{mis}(m)} = (2.3, 5.27)$$

MCAR[2]: For all $m \in \{0,1\}^d, P(M = m \mid X) = P(M = m)$
MAR[3]: For all $m \in \{0,1\}^d, P(M = m \mid X) = P\left(M = m \mid X_{obs(m)}\right)$

[2] Michel, Naf, Spohn, ¨ Meinshausen. 2021. PKLM: a flexible mcar test using classification.
[3] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

Supervised learning with missing values

$\tilde{X} = X \odot (1 - M) + \mathtt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\mathtt{NA}\})^d$.

$$\mathbf{Y} = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{\mathbf{X}} = \begin{pmatrix} 9.1 & \mathtt{NA} & 1 \\ 2.1 & \mathtt{NA} & 3 \\ \mathtt{NA} & 9.6 & 2 \\ \mathtt{NA} & 5.5 & 6 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the expected risk

$$\text{Bayes rule: } f^* \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M)}, M\right]$$
$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d) (Rubin, 1984, generalized propensity score)

Supervised learning with missing values

$\tilde{X} = X \odot (1 - M) + \mathtt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\mathtt{NA}\})^d$.

$$\mathbf{Y} = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{\mathbf{X}} = \begin{pmatrix} 9.1 & \mathtt{NA} & 1 \\ 2.1 & \mathtt{NA} & 3 \\ \mathtt{NA} & 9.6 & 2 \\ \mathtt{NA} & 5.5 & 6 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the expected risk

$$\text{Bayes rule: } f^* \in \underset{f:\, \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M)}, M\right]$$
$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d) (Rubin, 1984, generalized propensity score)

Supervised learning with missing values

- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^* \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right]$ $f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X}]$

- Empirical risk: $\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^n \ell\left(f(\tilde{X}_i), Y_i\right)\right)$

  A new data $\mathcal{D}_{n,\mathrm{test}}$ to estimate the generalization error rate

- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(\tilde{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\tilde{X}), Y)]$

Supervised learning with missing values

- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^* \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right]$ $f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X}]$

- Empirical risk: $\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^n \ell\left(f(\tilde{X}_i), Y_i\right)\right)$

  A new data $\mathcal{D}_{n,\mathrm{test}}$ to estimate the generalization error rate

- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(\tilde{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\tilde{X}), Y)]$

Differences with classical litterature

Aim: target an outcome Y (not estimate parameters and their variance)
Specificities: train & test sets with missing values. If not: distributional shift; data generating process (X, Y, M)

$\Rightarrow$ Is it possible to use previous approaches (EM - impute), consistent?
$\Rightarrow$ Do we need to design new ones?

Impute then Regress procedures

Imputation prior to learning: Impute then Regress

Common practice: use off-the-shelf methods 1) for imputation of missing values and 2) for supervised-learning on the resulting completed data

Separate imputation

Impute train and test separately (with a different model)

Issue: Depends on the size of the test set? one observation?

Group imputation/ semi-supervised

Impute train and test simultaneously but the predictive model is learned only on the training imputed data set

Issue: Sometimes no training set at test time

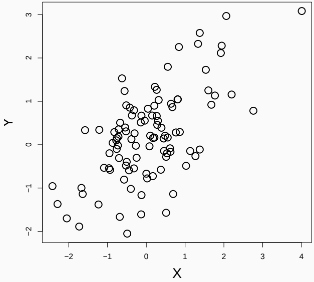
Imputation train and test with the same model

Easy to implement for univariate imputation: compute the means on the observed data $(\hat{\mu}_1, ..., \hat{\mu}_d)$ of each colum of the train set and impute the test set with the same means. (OK for Gaussian imput.)
Issue: Many methods are "black-boxes" and take as an input the incomplete data and output the completed data (`missForest`)

Mean imputation

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	-1.93
-0.86	-1.50
.....	...
2.16	0.7
0.16	0.74



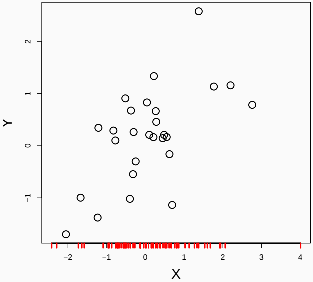
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

$\hat{\mu}_{x_2} = -0.01$
$\hat{\sigma}_{x_2} = 1.01$
$\hat{\rho} = 0.66$

Mean imputation

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on X_2

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	NA
-0.86	NA
.....	...
2.16	0.7
0.16	NA



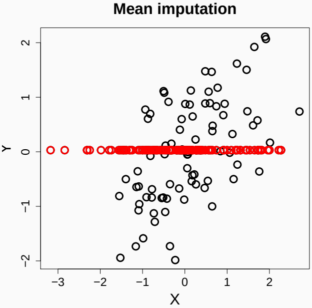
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

$\hat{\mu}_{x_2} = 0.18$
$\hat{\sigma}_{x_2} = 0.9$
$\hat{\rho} = 0.6$

Mean imputation

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on X_2
- Estimate parameters on the mean imputed data

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	**0.01**
-0.86	**0.01**
.....	...
2.16	0.7
0.16	**0.01**

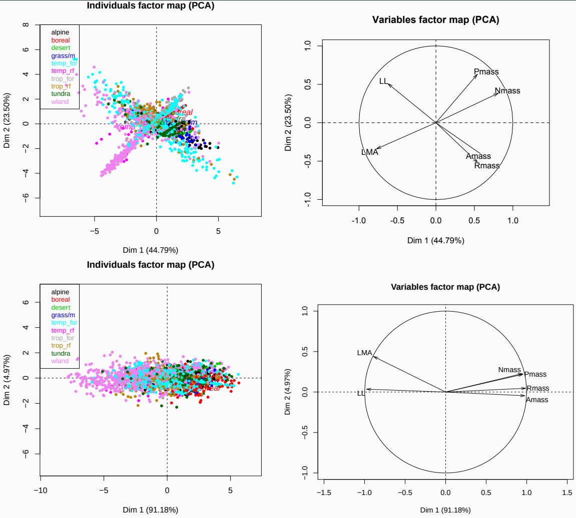


$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

$\hat{\mu}_{x_2} = 0.01$
$\hat{\sigma}_{x_2} = 0.5$
$\hat{\rho} = 0.30$

Mean imputation deforms joint and marginal distributions

Mean imputation is bad for estimation



PCA with mean imputation

```
library(FactoMineR)
PCA(ecolo)
Warning message: Missing
are imputed by the mean
of the variable:
You should use imputePCA
from missMDA
```

EM-PCA

```
library(missMDA)
imp <- imputePCA(ecolo)
PCA(imp$comp)
```

J. (2016). missMDA: Handling Missing Values in Multivariate Data Analysis, JSS.

Ecological data: [4] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

[4] Wright, I. et al. (2004). The worldwide leaf economics spectrum. *Nature*.

Constant (mean) imputation is consistent for prediction

Framework - assumptions

- $Y = f(X) + \varepsilon$
- $X = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0,1]^d$
- $\|f\|_\infty < \infty$
- Missing data MAR on X_1 with $M_1 \perp\!\!\!\perp X_1 | X_2, \ldots, X_d$
- $(x_2, \ldots, x_d) \mapsto \mathbb{P}[M_1 = 1 | X_2 = x_2, \ldots, X_d = x_d]$ is continuous
- ε is a centered noise independent of (X, M_1)

(remains valid when missing values occur for several variables $X_1, \ldots, X_j$)

Constant (mean) imputation is consistent for prediction

Constant imputed entry $x' = (x'_1, x_2, \ldots, x_d)$: $x'_1 = x_1 \mathbb{1}_{M_1=0} + \alpha \mathbb{1}_{M_1=1}$

Theorem. (J. et al. 2019)

$$
\begin{aligned}
f^\star_{impute}(x') =& \mathbb{E}[Y|X_2 = x_2, \ldots, X_d = x_d, M_1 = 1] \\
&\qquad \mathbb{1}_{x'_1=\alpha} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]>0} \\
&\qquad + \mathbb{E}[Y|X = x'] \mathbb{1}_{x'_1=\alpha} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]=0} \\
&\qquad + \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d, M_1 = 0] \mathbb{1}_{x'_1 \neq \alpha}.
\end{aligned}
$$

Prediction with mean is equal to the Bayes function almost everywhere

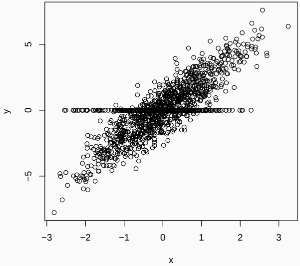
$$f^\star_{impute}(X') = f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X} = \tilde{x}]$$

Rq: pointwise equality if using a constant out of range.

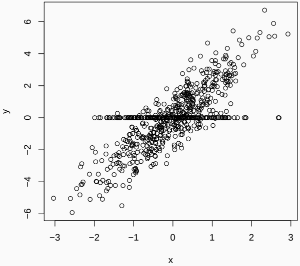
$\Rightarrow$ Learn on the mean-imputed training data, impute the test set with the same means and predict is optimal if the missing data are MAR and the **learning algorithm is universally consistent** (for all distribution)

Consistency of constant imputation: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant:
- Need a lot of data (asymptotic result) and a super powerful learner



Train

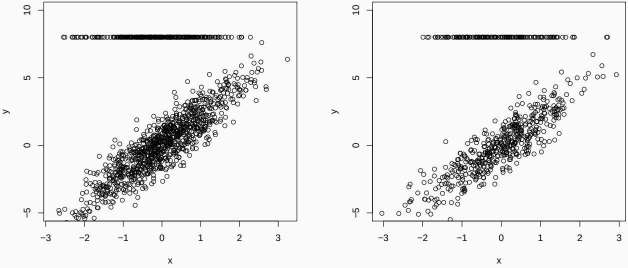


Test

Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

Consistency of constant imputation: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant: out of range
- Need a lot of data (asymptotic result) and a super powerful learner



Train   Test

Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

Bayes optimality of impute-n-regress (Le morvan et al. 2021)

Define Impute-then-Regress procedures as functions of the form: $g \circ \Phi$ where $\Phi \in \mathcal{C}_\infty$ and $g : \mathbb{R}^d \to \mathbb{R}$

Φ is a deterministic imputation, a function of the observed values (Ex: mean imputation, regression imputation, etc.)

Theorem

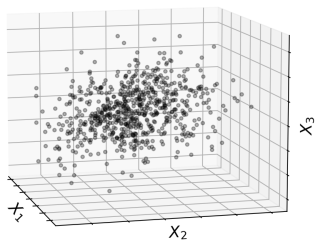
Assume that the response Y satisfies $Y = f^\star(X) + \epsilon$
Let $g^\star_\Phi$ be the minimizer of the risk on the data imputed by Φ. Then,

for all missing data mechanisms & almost all imputation functions, $g^\star_\Phi \circ \Phi$ is Bayes optimal

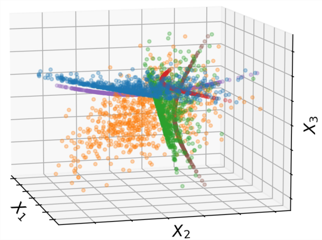
$\Rightarrow$ A universally consistent algorithm trained on the imputed data $\Phi(\widetilde{X})$ is Bayes consistent

Asymptotically, imputing well is not needed to predict well

Bayes optimality of impute-n-regress (Le morvan et al. 2021)



Complete data



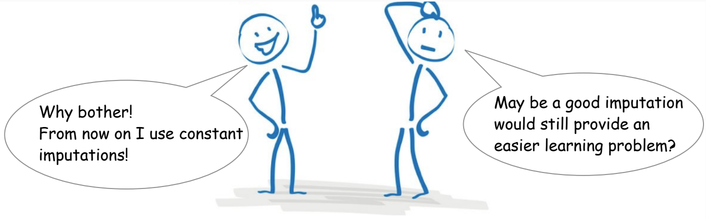
Imputed data (manifolds)

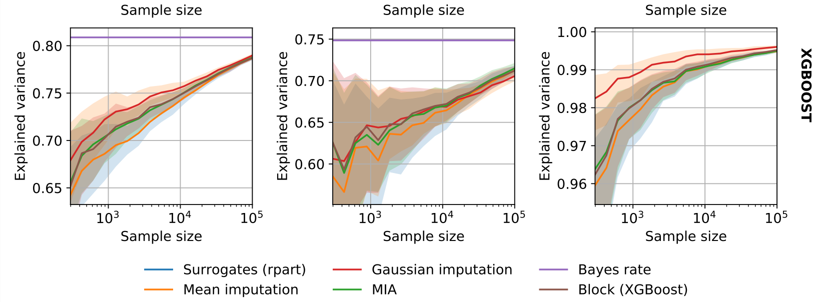
Rationale: Imputation create manifolds to which the learner adapts

1. All data points with a missing data pattern m are mapped to a manifold $\mathcal{M}^{(m)}$ of dimension $|obs(m)|$ (Preimage Theorem)
2. The missing data patterns of imputed data points can almost surely be de-identified (Thom transversality Theorem) [5]
3. Given 2), we can build prediction functions, independent of m, that are Bayes optimal for all missing data patterns

[5] Non transverse: the manifolds on which the data with either x1 missing or x2 missing are projected are exactly the same (the same line)

Which imputation function should one choose?





Constant imputation "breaks" models, introduce strong discontinuities

Which imputation function and predictor should one choose?

- **Chaining oracles**: $f^\star \circ \Phi^{CI}$ with Φ^{CI} the oracle imput $\mathbb{E}[X_{mis}|X_{obs}, M]$

Proposition (excess of risk of chaining oracle)

Assum PSD matrices $\bar{H}^+$ & $\bar{H}^-$ s.t. for all $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$

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

High excess risk if both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high (linear regression consistent)

$\Rightarrow$ Choosing an oracle for one step, imputation or regression, imposes discontinuities on the other step, thus making it harder to learn

Which imputation function and predictor should one choose?

- **Chaining oracles**: $f^\star \circ \Phi^{CI}$ with Φ^{CI} the oracle imput $\mathbb{E}[X_{mis}|X_{obs}, M]$

Proposition (excess of risk of chaining oracle)

Assum PSD matrices $\bar{H}^+$ & $\bar{H}^-$ s.t. for all $X \in \mathcal{S}, \bar{H}^- \preceq H(X) \preceq \bar{H}^+$

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

High excess risk if both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high (linear regression consistent)

- **Learning on Cond. Imput. data (imputing as well as possible before learning)**: Is there a continuous function g, s.t. $g \circ \Phi^{CI}$ is Bayes optimal?

No. Size of the discontinuities are controlled by the variance-curvature tradeoff

⇒ Choosing an oracle for one step, imputation or regression, imposes discontinuities on the other step, thus making it harder to learn

Which imputation function and predictor should one choose?

- **Chaining oracles**: $f^\star \circ \Phi^{CI}$ with Φ^{CI} the oracle imput $\mathbb{E}[X_{mis}|X_{obs}, M]$

Proposition (excess of risk of chaining oracle)

Assum PSD matrices $\bar{H}^+$ & $\bar{H}^-$ s.t. for all $X \in \mathcal{S}$, $\bar{H}^- \preceq H(X) \preceq \bar{H}^+$

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

High excess risk if both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high (linear regression consistent)

- **Learning on Cond. Imput. data (imputing as well as possible before learning)**: Is there a <u>continuous</u> function g, s.t. $g \circ \Phi^{CI}$ is Bayes optimal?

No. Size of the discontinuities are controlled by the variance-curvature tradeoff

- **Optimizing imputations for a fixed regression function.** Keeping $f^\star$, is there a <u>continuous</u> imputation function Φ s.t $f^\star \circ \Phi$ is Bayes optimal?

Sometimes yes and no

$\Rightarrow$ Choosing an oracle for one step, imputation or regression, imposes discontinuities on the other step, thus making it harder to learn

Jointly learn imputation and prediction: Neumiss

Explicit Bayes predictor with missing values

Linear model:

$Y = \beta_0 + \langle X, \beta \rangle + \varepsilon, \quad X \in \mathbb{R}^d, \ \varepsilon$ gaussian.

Bayes predictor for the linear model:

$$\begin{aligned} f^\star(\tilde{X}) &= \mathbb{E}[Y|\tilde{X}] = \mathbb{E}[\beta_0 + \beta^\top X \mid \ M, X_{obs(M)}] \\ &= \beta_0 + \beta_{obs(M)}^\top X_{obs(M)} + \beta_{mis(M)}^\top \ \mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] \end{aligned}$$

Assumptions on covariates and missing values

1. Gaussian pattern mixture model, PMM: $X \mid (M = m) \sim \mathcal{N}(\mu_m, \Sigma_m)$
2. Gaussian assumption $X \sim \mathcal{N}(\mu, \Sigma)$ + MCAR and MAR
3. (Also for Gaussian assumption + MNAR self mask gaussian)

Under Assump. 2 the Bayes predictor is linear per pattern

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

use of *obs* instead of *obs*(M) for lighter notations - Expression for 2.

Linear model with missing values not necessarely linear

Example

Let $Y = X_1 + X_2 + \varepsilon$, where $X_2 = \exp(X_1) + \varepsilon_1$. Now, assume that only X_1 is observed. Then, the model can be rewritten as

$$Y = X_1 + \exp(X_1) + \varepsilon + \varepsilon_1,$$

where $f(X_1) = X_1 + \exp(X_1)$ is the Bayes predictor. In this example, the submodel for which only X_1 is observed is not linear.

$\Rightarrow$ There exists a large variety of submodels for a same linear model. Depend on the structure of X and on the missing-value mechanism.

Neumiss Networks to approximate the covariance matrix

Order-ℓ approx of the Bayes predictor in MAR

$f^\star_\ell(X_{obs}, M) = \langle \beta_{obs}, X_{obs} \rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)} (X_{obs} - \mu_{obs}) \rangle.$

Order-ℓ approx of $(\Sigma^{-1}_{obs(m)})$ for any m defined recursively:

$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)}) S^{(\ell-1)}_{obs(m)} + Id.$$

Neuman Series, $S^{(0)} = Id$, $\ell = \infty$: $(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty} (Id - \Sigma_{obs(m)})^k$

Neumiss Networks to approximate the covariance matrix

Order-ℓ approx of the Bayes predictor in MAR

$$f^\star_\ell(X_{obs}, M) = \langle \beta_{obs}, X_{obs}\rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)}(X_{obs} - \mu_{obs})\rangle.$$

Order-ℓ approx of $(\Sigma^{-1}_{obs(m)})$ for any m defined recursively:

$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)}) S^{(\ell-1)}_{obs(m)} + Id.$$

Neuman Series, $S^{(0)} = Id$, $\ell = \infty$: $(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty} (Id - \Sigma_{obs(m)})^k$

$\Rightarrow$ **Neural network architecture to approximate the Bayes predictor**

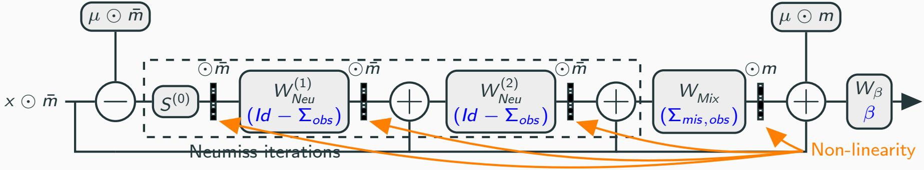


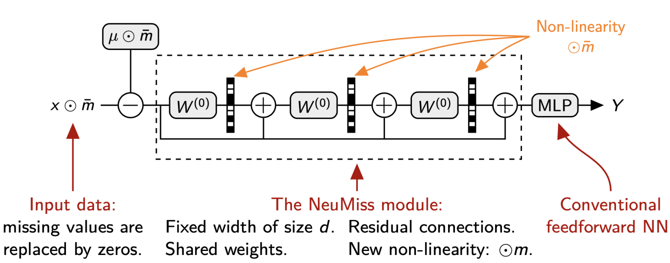
Figure 1: Depth of 3, $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Networks with missing values: $\odot M$ nonlinearity

- Implementing a network with the matrix **weights** $W^{(k)} = (I - \Sigma_{obs(m)})$ **masked differently for each sample** can be challenging
- Masked weights is **equivalent to masking input & output vector.**

Let v a vector, $\bar{m} = 1 - m$. $(W \odot \bar{m}\bar{m}^\top)v = (W(v \odot \bar{m})) \odot \bar{m}$

Classic network with multiplications by the mask nonlinearities $\odot M$

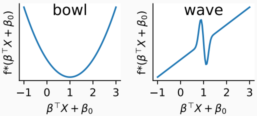


Best imputation is joint learn with regression

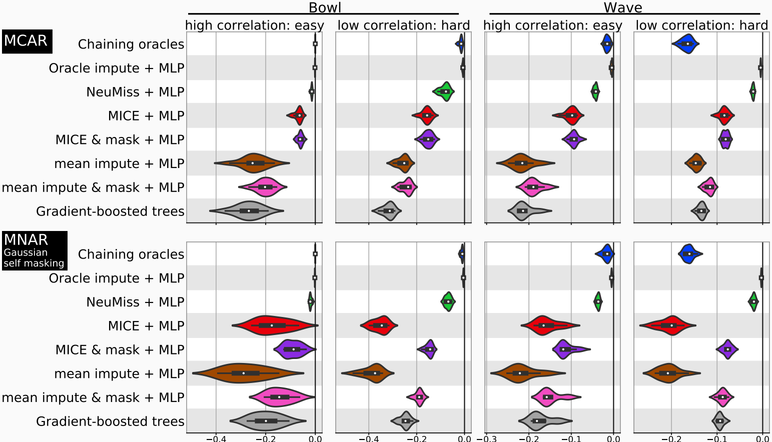
Experimental results

- $Y = f^\star(X) + \epsilon$. $n = 100,000$, $d = 50$, 50% NA

Gaussian X: "high/ low' correlation



- Gradient-Boosted Trees: with Missing Incorporated Attribute strategy
- Concatenating the mask to help for MNAR



Discussion - challenges

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with a powerful learner
- Rather than a sophisticated imputation use rather a powerful learner
- **Rethinking imputation: a good imputation is the one that makes the prediction easy**
- Close to conditional imputation but not CI
- Can even work in MNAR

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with a powerful learner
- Rather than a sophisticated imputation use rather a powerful learner
- **Rethinking imputation: a good imputation is the one that makes the prediction easy**
- Close to conditional imputation but not CI
- Can even work in MNAR

Implicit and jointly learned Impute-then-Regress strategy

- Neumiss network: new architecture $\odot M$ nonlinearity
- Theoritically: differentiable approximation of the cond. expectation
- Tree-based models: Missing Incorporated in Attribute

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with a powerful learner
- Rather than a sophisticated imputation use rather a powerful learner
- **Rethinking imputation: a good imputation is the one that makes the prediction easy**
- Close to conditional imputation but not CI
- Can even work in MNAR

Implicit and jointly learned Impute-then-Regress strategy

- Neumiss network: new architecture $\odot M$ nonlinearity
- Theoritically: differentiable approximation of the cond. expectation
- Tree-based models: Missing Incorporated in Attribute

Causal inference with missing values

Multiple imputation - Superlearner with missing values (aggregation)

Conformal prediction with missing values (conditional by pattern)

Ressources

R-miss-tastic `https://rmisstastic.netlify.com/R-miss-tastic`

J., I. Mayer, N. Tierney & N. Vialaneix

Project funded by the R consortium (Infrastructure Steering Committee)[6]

Aim: a reference platform on the theme of missing data management

- list existing packages
- available literature
- tutorials
- analysis workflows on data
- main actors

⇒ Federate the community

⇒ Contribute!

[6] `https://www.r-consortium.org/projects/call-for-proposals`

Ressources

Examples:

- Lecture [7] - General tutorial : Statistical Methods for Analysis with Missing Data (Mauricio Sadinle)
- Lecture - Multiple Imputation: `mice` by Nicole Erler [8]
- Longitudinal data, Time Series Imputation (Steffen Moritz - very active contributor of r-miss-tastic), Principal Component Methods[9]

---

[7] https://rmisstastic.netlify.com/lectures/

[8] https://rmisstastic.netlify.com/tutorials/erler_course_multipleimputation_2018/erler_practical_mice_2018

[9] https://rmisstastic.netlify.com/tutorials/Josse_slides_imputation_PCA_2018.pdf