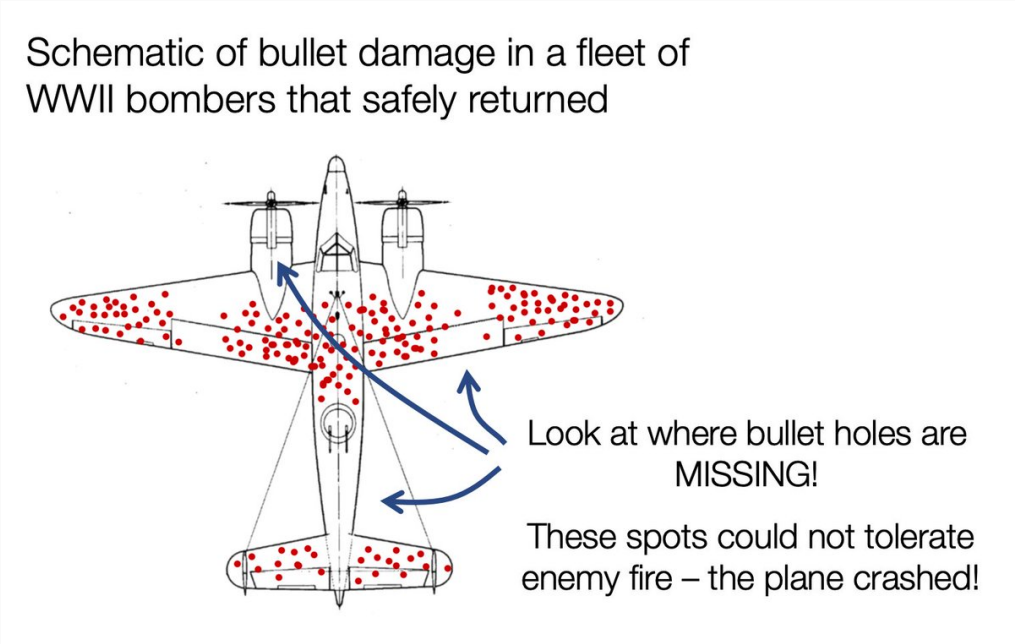
Missing data: from inference to imputation & prediction; Is there a one for all solution?

Julie Josse
Inria, Ecole Polytechnique
Head of Premedical (Precision medicine by data integration & causal learning) Inria-Inserm team

27 July 2022

AutoML 2022



Traumabase project: decision support for trauma patients

- 30000 French trauma patients
- 250 features from the accident site to the hospital discharge
- 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 French trauma patients
- 250 features from the accident site to the hospital discharge
- 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" on the **outcome** mortality for trauma brain patients.
Causal inference with covariates with missing values [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

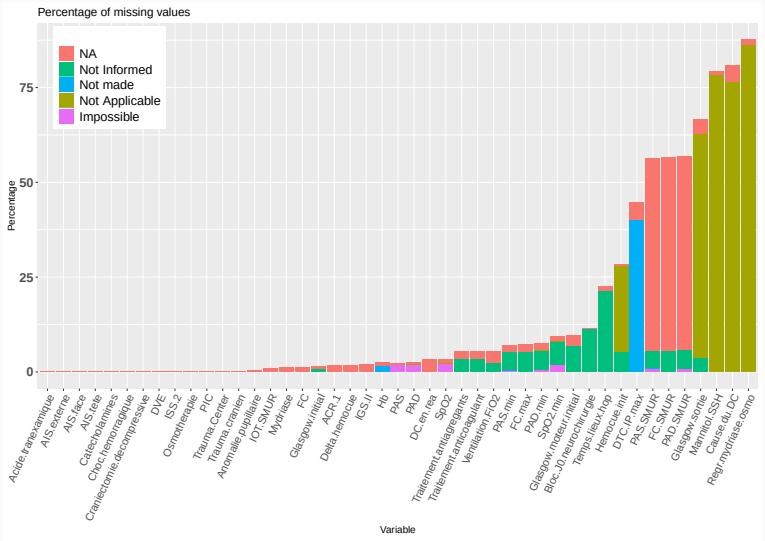
- 30000 French trauma patients
- 250 features from the accident site to the hospital discharge
- 30 hospitals
- 4000 new patients/ year

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	
⋮								⋱

$\Rightarrow$ **Explain and Predict** hemorrhagic shock given pre-hospital features.
Ex logistic regression/ random forests with covariates with missing values

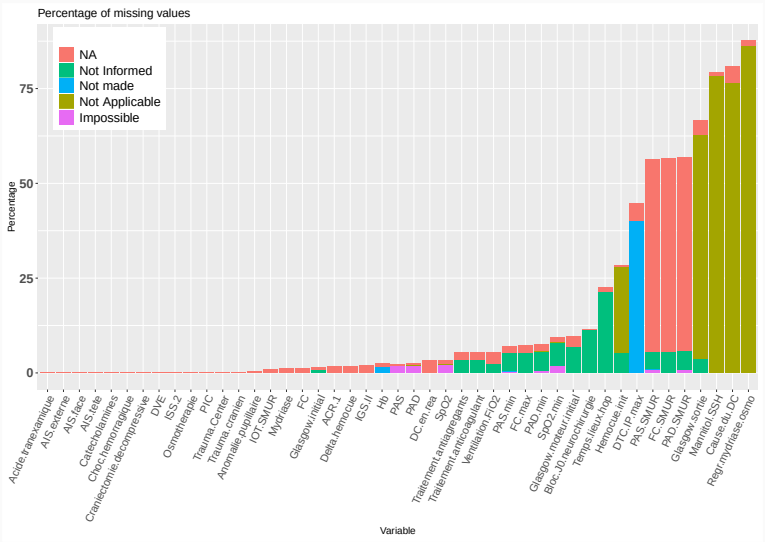
Prospective study: real-time testing of models in the ambulance via a mobile data collection application

Missing data: important bottleneck in data science



Sporadic & systematic (missing variable in one hospital). Due to the pandemic, many patients did not complete their tests

Missing data: important bottleneck in data science



Sporadic & systematic (missing variable in one hospital). Due to the pandemic, many patients did not complete their tests

"One of the ironies of Big Data is that missing data play an ever more significant role" (R. Samworth, 2019)

Complete case analysis (deletion):

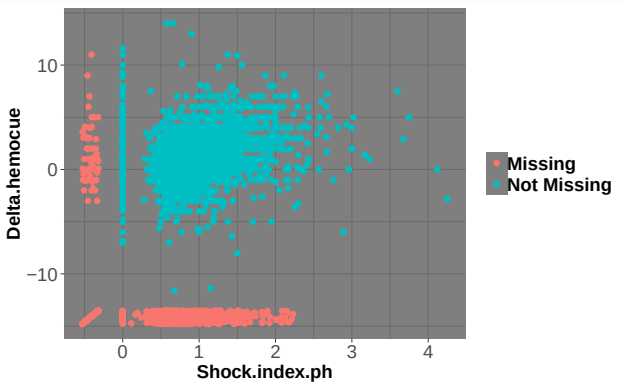
- Bias: Resulting sample not representative of the target population
- Loss of information: An $n \times d$ matrix, each entry is missing with probability 0.01. $d = 5 \implies \approx 95\%$ of rows kept; $d = 300 \implies \approx 5\%$ of rows kept

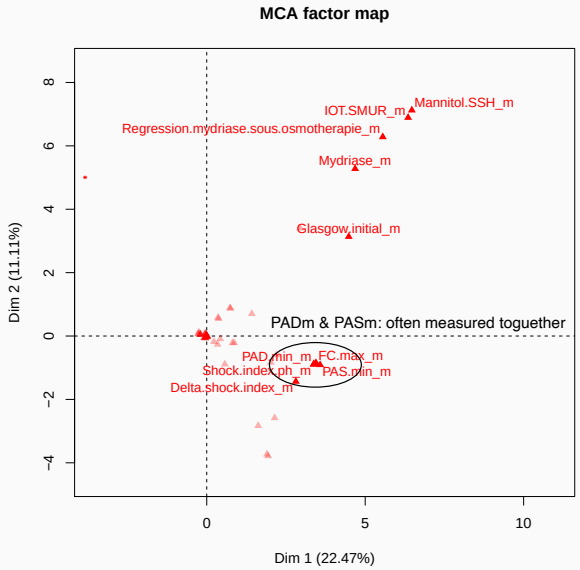
Inference with missing values

What is a 'true' missing value?

The first thing to do with missing values (as with any analysis) is descriptive statistics: Visualize their patterns to get clues about how and why they occured

R packages: `VIM`, `naniar`, `FactoMineR`



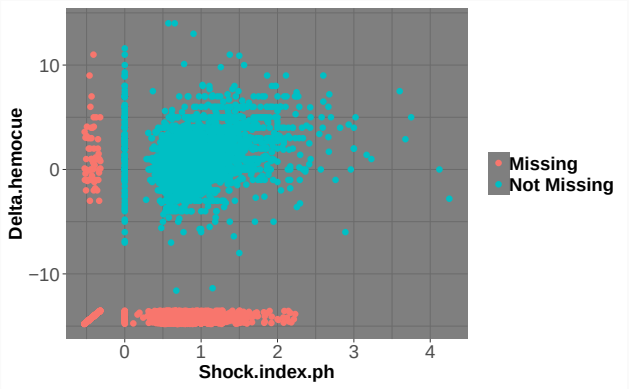


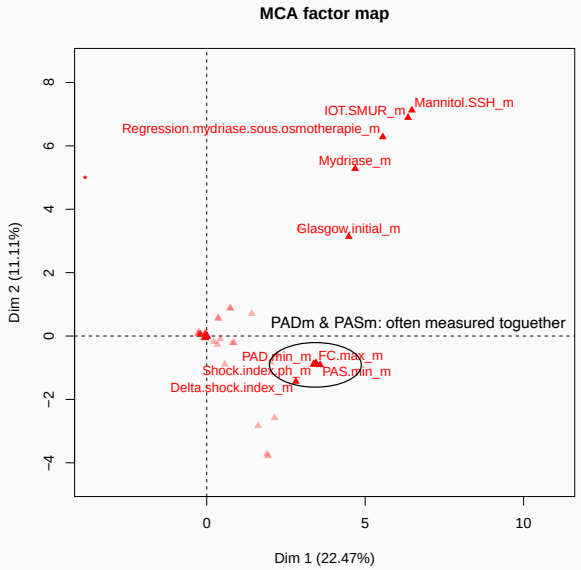
Right plot: clustering of the missingness matrix (with m for miss & o for obs.)

What is a 'true' missing value?

The first thing to do with missing values (as with any analysis) is descriptive statistics: Visualize their patterns to get clues about how and why they occured

R packages: `VIM`, `naniar`, `FactoMineR`





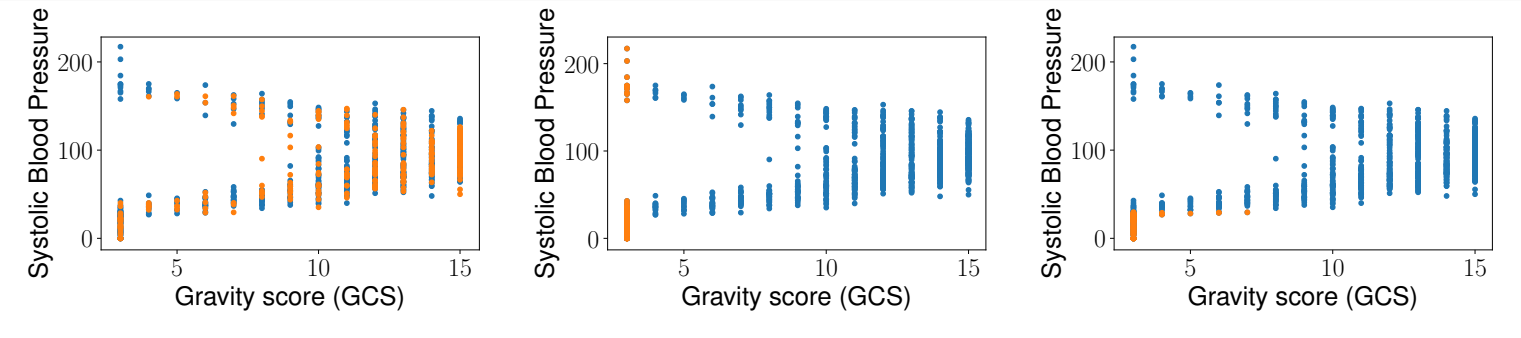
Right plot: clustering of the missingness matrix (with m for miss & o for obs.)

Detect nested variables: Test1: yes/no, if yes Test2 (a, b), if no Test2 'missing'

- Not a 'true' missing value, does not **mask an underlying value**
- Solution: recoding with a new variable with 3 categories 'yes a', 'yes b', 'no'

$\Rightarrow$ Feedbacks on data collection/encoding process

Missing values mechanism: Rubin's taxonomy [1] [2]



MCAR - MAR - MNAR

Orange: missing values for SBP - GCS is always observed

- MCAR: Proba of having missing values does not depend on the observed or the missing values
- MAR: Proba of having missing values depends on the observed values
- MNAR: Proba of having missing values depends on the missing values

Data distribution $f_\theta(X)$ for the complete data; Missingness distribution $g_\phi(M)$
Under M(C)AR, $g_\phi(M)$ can be **ignored** while performing inference for θ

[1] Rubin, 1976. Inference and missing data. *Biometrika*
[2] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

Solutions to handle M(C)AR values (in the covariates)

Abundant literature: Rmistatic platform [3], more than 150 packages

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

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

Pros: Tailored toward a specific problem

Cons: Difficult to establish, few softwares even for simple models [4]

One specific algorithm for each statistical/ML method...

Multiple imputation to get a complete data set

Pros: Any analysis can be performed - `mice R package`

Cons: Generic - current implementation have computational issues for large dimensions

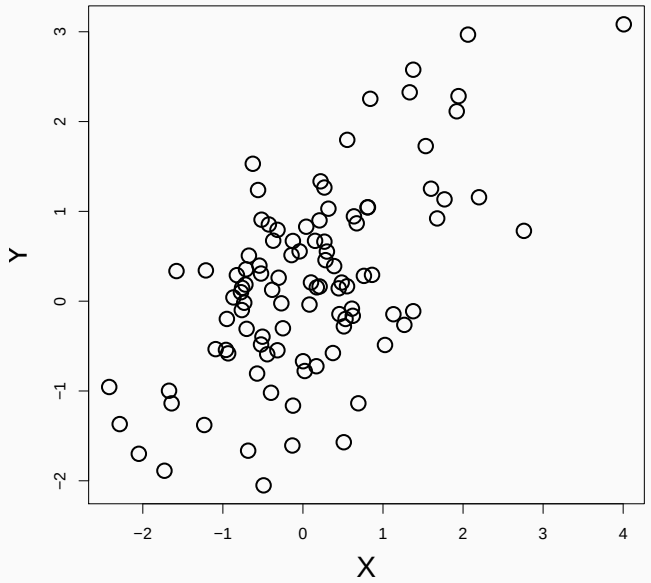
[3] Mayer, et al. A unified platform for missing values methods and workflows. *R journal*. 2022.

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

Single imputation by the mean

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

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



$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

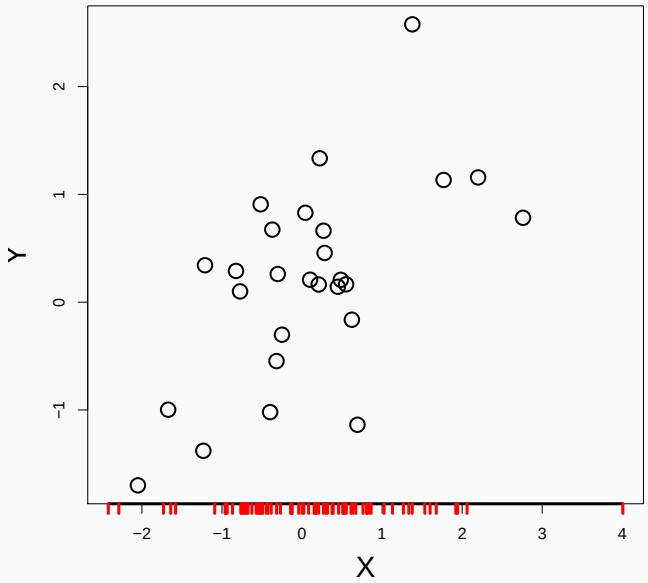
$\rho = 0.6$

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

Single imputation by the mean

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



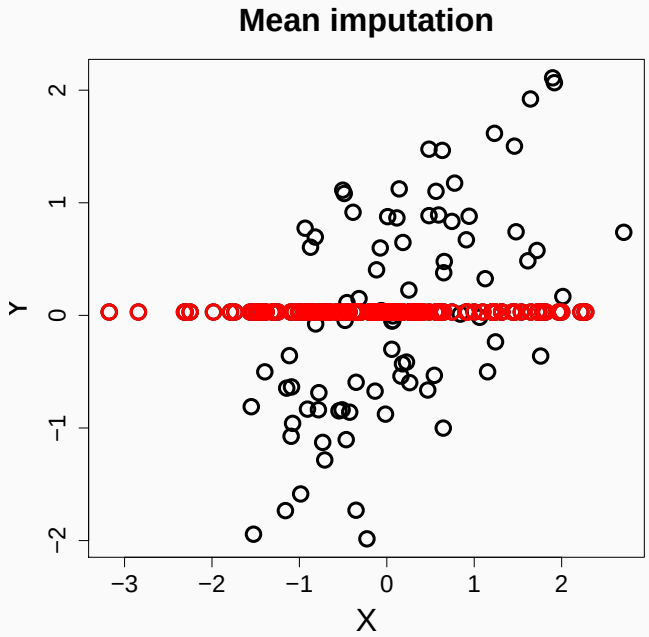
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

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

Single imputation by the mean

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

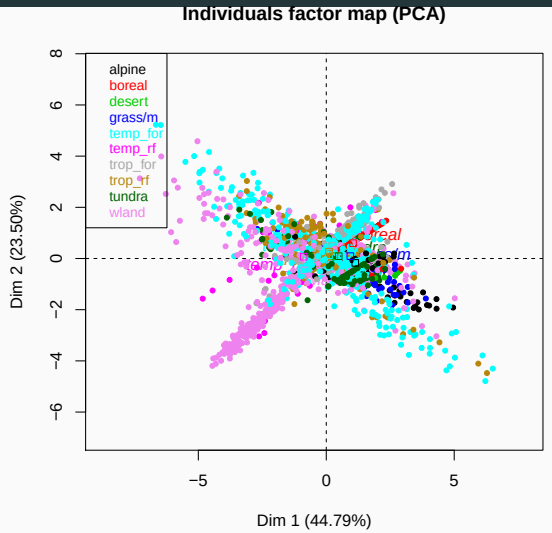


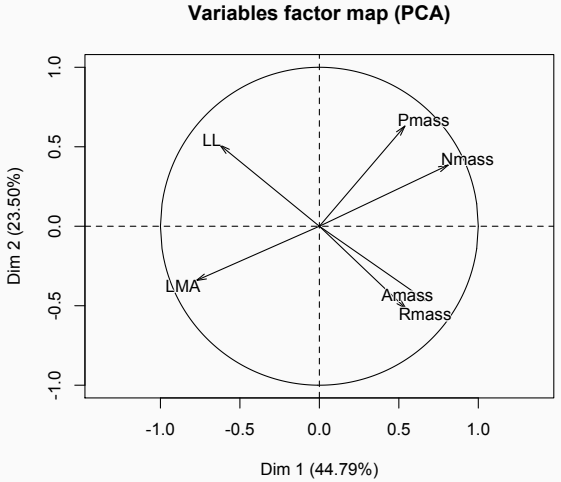
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

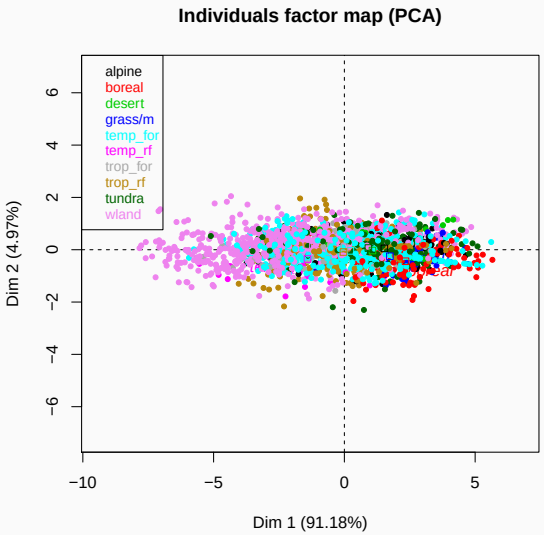
$\hat{\mu}_{x_2} = 0.01$
$\hat{\sigma}_{x_2} = 0.5$
$\hat{\rho} = 0.30$

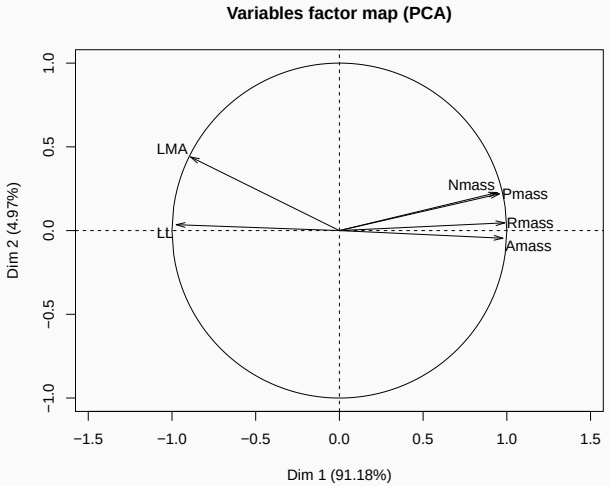
Mean imputation deforms joint and marginal distributions

Mean imputation is to be avoided 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. missMDA: Handling Missing Values in Multivariate Data Analysis, *JSS*. 2016.

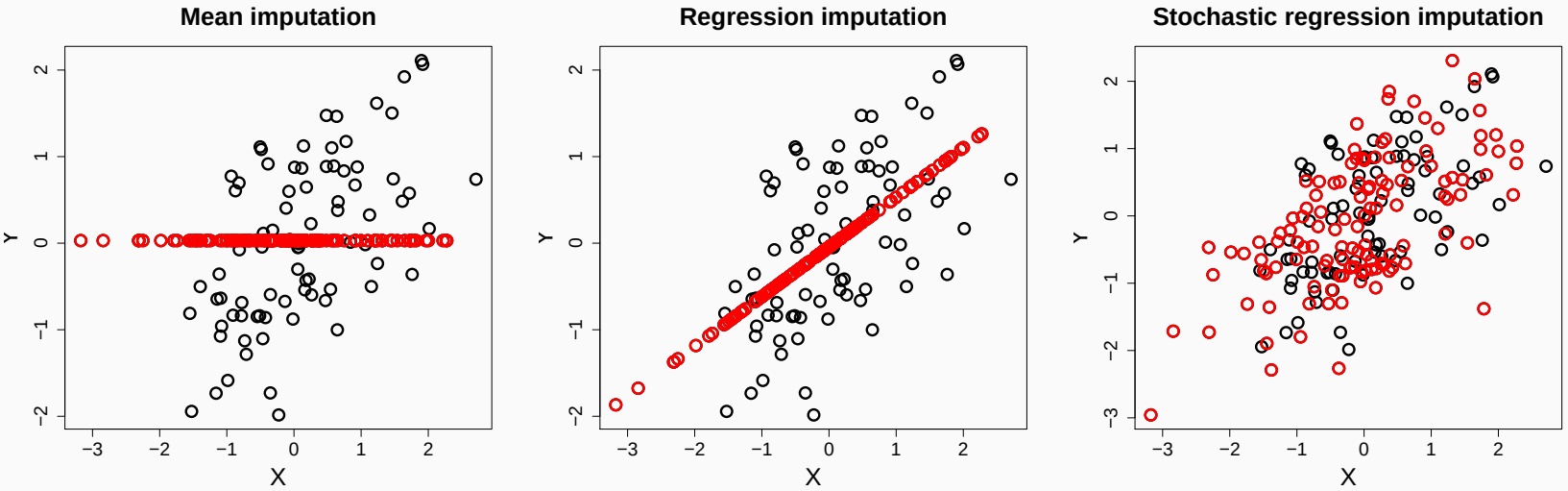
Ecological data: [5] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

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

Objective: to impute while preserving distribution

- by regression takes into account the relationship: Estimate β - impute $\hat{x}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \Rightarrow$ variance underestimated and correlation overestimated
- by stochastic reg: Estimate β and σ - impute from the predictive $\hat{x}_{i2} \sim \mathcal{N}\left(\beta_0 + \hat{\beta}_1 x_{i1}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions

Here $\hat{\beta}, \hat{\sigma}^2$ estimated with complete data, but MLE can be obtained with EM



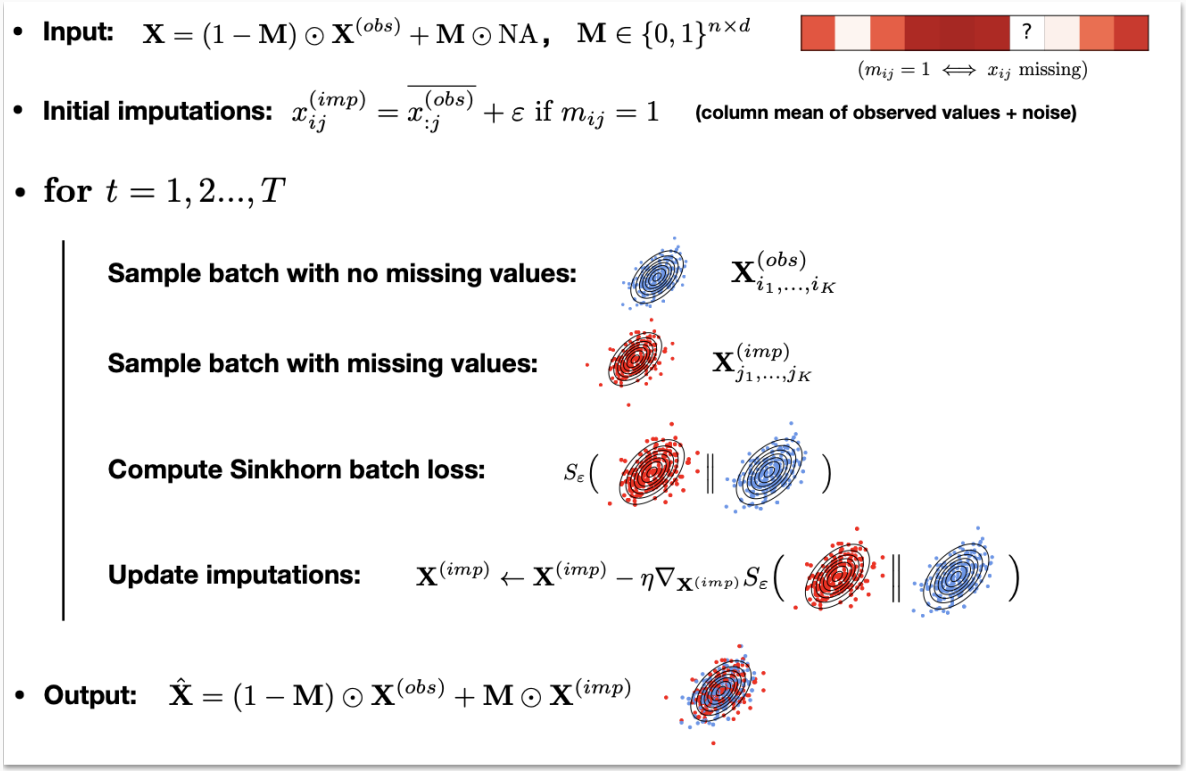
$\mu_{x_2} = 0$	0.01	0.01	0.01
$\sigma_{x_2} = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Impute while preserving distribution. Multivariate case

$\Rightarrow$ Parametric: assuming a joint model, Gaussian $z_i \sim \mathcal{N}(\mu, \Sigma)$

$\Rightarrow$ Nonparametric: using optimal transport [6]

- Two batches from the same dataset should have similar distributions
- Measure this with Sinkhorn divergence: differentiable & fast



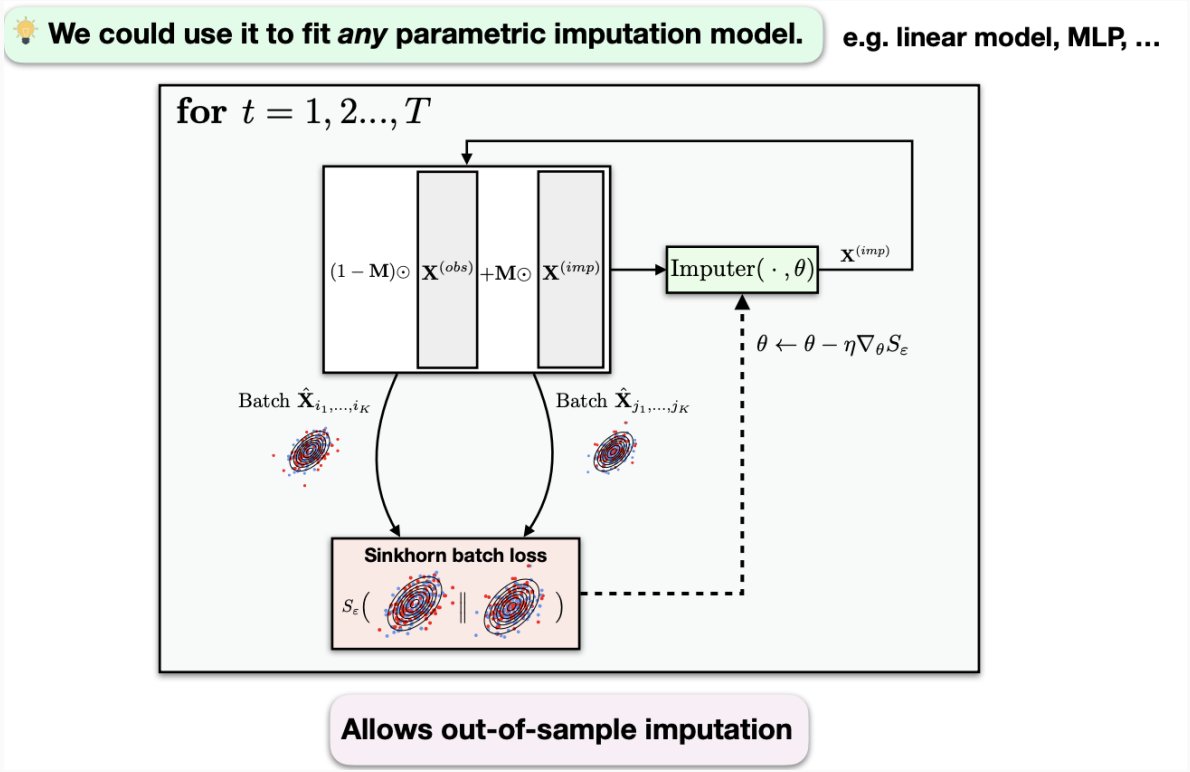
[6] Muzelec, Cuturi, Boyer, J. Missing Data Imputation using Optimal Transport. *ICML*. 2020.

Impute while preserving distribution. Multivariate case

$\Rightarrow$ Parametric: assuming a joint model, Gaussian $z_i \sim \mathcal{N}(\mu, \Sigma)$

$\Rightarrow$ Nonparametric: using optimal transport [6]

- Two batches from the same dataset should have similar distributions
- Measure this with Sinkhorn divergence: differentiable & fast



[6] Muzelec, Cuturi, Boyer, J. Missing Data Imputation using Optimal Transport. *ICML*. 2020.

Single imputation is not enough: Underestimate the variability

$\Rightarrow$ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

Single imputation is not enough: Underestimate the variability

$\Rightarrow$ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

$\Rightarrow$ Completed Traumabase

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

Single imputation is not enough: Underestimate the variability

$\Rightarrow$ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

$\Rightarrow$ Completed Traumabase

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

A single value can't reflect the uncertainty of prediction

Multiple impute 1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
-2	75	12	no s
1	63	40	s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
-2	10	12	no s
1	63	40	s

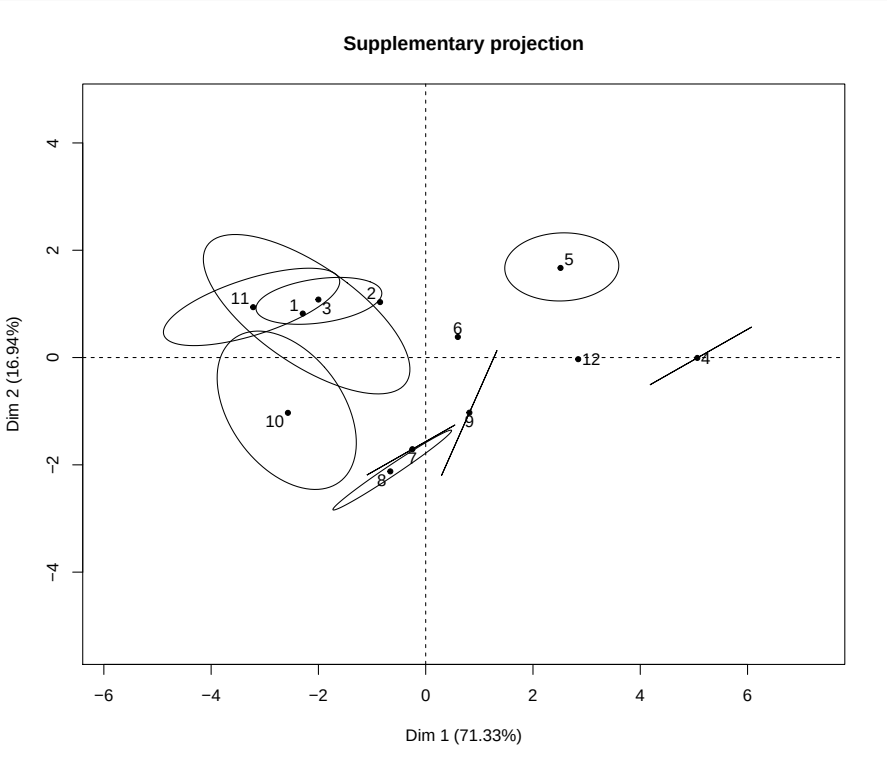
X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s

Visualization of the imputed values [7]

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
-2	15	12	no s
1	63	40	s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
-2	10	12	no s
1	63	40	s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s




```
library(missMDA)
MIPCA(traumadata)
```

Percentage of NA?

Projection of the M imputed data on a 'compromise' subspace (PCA with missing values)

[7] J. et al. Multiple imputation in principal component analysis. *ADAC*. 2011.

Multiple imputation: standard errors are not underestimated

1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
1	63	40	s
-2	15	12	no s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
1	63	40	s
-2	10	12	no s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
1	63	40	s
-2	20	12	no s

2) Perform the analysis on each imputed data set: $\hat{\beta}_m, \widehat{Var}\left(\hat{\beta}_m\right)$

3) Combine the results (Rubin's rules):

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$$

$$T = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

```
imp.mice <- mice(traumadata)
lm.mice.out <- with(imp.mice, glm(Y ~ ., family = "binomial"))
```

$\Rightarrow$ Variability of missing values taken into account. **Metric: coverage.**

Multiple imputation by chained equations [9]

- Impute variables 1 by 1 using all other variables as inputs (round-robin)
- One model/variable: flexible for categorical, ordinal variables
- Cycle through variables: iteratively refine the imputation

1. Initial imputation: mean imputation
2. For a variable j

   2.2 Imputation of the missing values in variable j with a model of X_j on the other X_{-j}: stochastic regression imput. $\sim \mathcal{N}\left((x_{i,-j})'\hat{\beta}^{-j}, \hat{\sigma}^{-j}\right)$
3. Cycling through variables

$\Rightarrow$ Imputed values are draws from an (implicit) joint distribution

Implemented in R package `mice` and IterativeImputer from `scikitlearn` [8]

[8] IterativeImputer by default does single imputation with iterative ridge regression

[9] van Buuren. 2018. Flexible Imputation of Missing Data. Second Edition. CRC Press

Multiple imputation by chained equations [9]

- Impute variables 1 by 1 using all other variables as inputs (round-robin)
- One model/variable: flexible for categorical, ordinal variables
- Cycle through variables: iteratively refine the imputation

1. Initial imputation: mean imputation
2. For a variable j
   - 2.1 $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})$ drawn from a Bootstrap: $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})^1, ..., (\hat{\beta}^{-j}, \hat{\sigma}^{-j})^M$
   - 2.2 Imputation of the missing values in variable j with a model of X_j on the other X_{-j}: stochastic regression imput. $\sim \mathcal{N}\left((x_{i,-j})'\hat{\beta}^{-j}, \hat{\sigma}^{-j}\right)$
3. Cycling through variables

$\Rightarrow$ Variance of prediction = variance of estimation + noise

$\Rightarrow$ Imputed values are draws from an (implicit) joint distribution

Implemented in R package `mice` and IterativeImputer from `scikitlearn` [8]

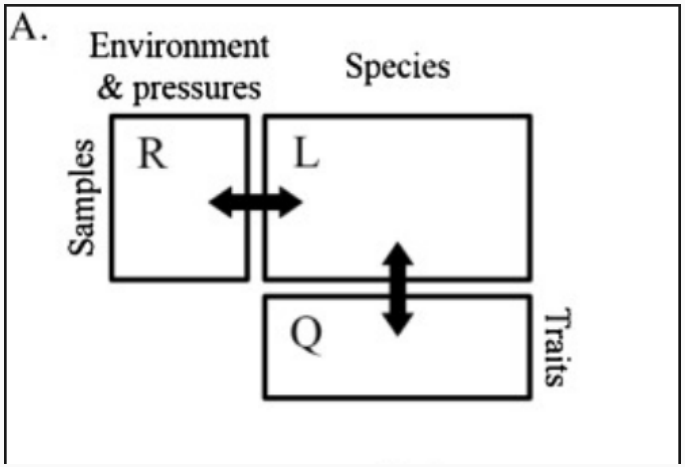
[8] IterativeImputer by default does single imputation with iterative ridge regression

[9] van Buuren. 2018. Flexible Imputation of Missing Data. Second Edition. CRC Press

Matrix completion/Single imputation

Monitor population & assess wetlands conservation policies

- National agency for wildlife and hunting management (ONCFS) data
- Contingency tables: Water (722 wetland sites) - bird (species) count data, from 1990-2016 in 5 countries in North Africa
- Side info: Additional sites & years info: meteo, geographical (altitude, etc.)



$\Rightarrow$ Aims: Assess the effect of time on species abundances

$\Rightarrow$ 70% of missing values in contingency tables (drough, war, etc.) [10] [11]

[10] Robin, J, Moulines Sardy. 2019. Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis.*

[11] Robin, Klopp, J, Moulines Tibshirani. Main effects and interactions in mixed and incomplete data frames. 2019. *JASA*.

Predicting as well as possible the missing values

Assuming a joint model

- low rank[12] : $Z_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k

$\Rightarrow$ Powerful in recommandation system: Netflix prize 90% of missing

$\Rightarrow$ Use similarities between rows & links between variables + reduct. of dim.

$\Rightarrow$ Different regularization depending on noise regime [13], [14], [15]

$\Rightarrow$ Count data, ordinal data [16], categorical data [17], blocks/multilevel data [18]

[12] Udell & Townsend. Why Are Big Data Matrices Approximately Low Rank? *SIAM*. 2019.
[13] J. & Sardy. Adaptive Shrinkage of singular values. *Stat & Computing*. 2015.
[14] J. & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*.. 2016.
[15] Hastie et al. Matrix completion & low-rank SVD via alternating least squares. *JMLR*. 2015.
[16] Zhao, Udell. Matrix completion with uncertainty through low rank copula. *Neurips. 2020*
[17] J. et al. Main effects and interactions in mixed and incomplete data frames. *JASA*. 2018.
[18] J. et al. Imputation of mixed data with multilevel SVD. *JCGS*. 2018.

Predicting as well as possible the missing values

Assuming a joint model

- low rank : $Z_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k

$\Rightarrow$ Powerful in recommandation system: Netflix prize 90% of missing

$\Rightarrow$ Use similarities between rows & links between variables + reduct. of dim.

$\Rightarrow$ Different regularization depending on noise regime

$\Rightarrow$ Count data, ordinal data , categorical data , blocks/multilevel data

- deep generative models: GAIN [12], VAEAC [13], MIWAE, [14]

$\Rightarrow$ challenging optimization, some require complete data, or MCAR

[12] Yoon et al. Gain: Missing data imputation using generative adversarial nets. *ICML*. 2018.

[13] Ivanov et al. Variational autoencoder with arbitrary conditioning. arXiv.

[14] Mattei & Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. *ICML*. 2018.

Predicting as well as possible the missing values

Assuming a joint model

- low rank : $Z_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k

$\Rightarrow$ Powerful in recommandation system: Netflix prize 90% of missing

$\Rightarrow$ Use similarities between rows & links between variables + reduct. of dim.

$\Rightarrow$ Different regularization depending on noise regime

$\Rightarrow$ Count data, ordinal data , categorical data , blocks/multilevel data

- deep generative models: GAIN [12], VAEAC [13], MIWAE, [14]

$\Rightarrow$ challenging optimization, some require complete data, or MCAR

Using conditional models (joint implicitly defined)

- with multinomial, poisson regressions (ICE: Imputation by Chained Equations)
- iterative impute each variable by random forests R package `missForest`

[12] Yoon et al. Gain: Missing data imputation using generative adversarial nets. *ICML*. 2018.

[13] Ivanov et al. Variational autoencoder with arbitrary conditioning. arXiv.

[14] Mattei & Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. *ICML*. 2018.

Iterative imputation by random forests versus by low rank (PCA)

	Feat1	Feat2	Feat3	Feat4	Feat5...
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	NA	NA	8	8
Frank	8	NA	NA	8	8
Bertrand	9	NA	NA	9	9
Alex	9	NA	NA	9	9
Yohann	10	NA	NA	10	10
Jean	10	NA	NA	10	10

Missing

Feat1	Feat2	Feat3	Feat4	Feat5
1	1.0	1.00	1	1
1	1.0	1.00	1	1
2	2.0	2.00	2	2
2	2.0	2.00	2	2
3	3.0	3.00	3	3
3	3.0	3.00	3	3
4	4.0	4.00	4	4
4	4.0	4.00	4	4
5	5.0	5.00	5	5
5	5.0	5.00	5	5
6	6.0	6.00	6	6
6	6.0	6.00	6	6
7	7.0	7.00	7	7
7	7.0	7.00	7	7
8	6.87	6.87	8	8
8	6.87	6.87	8	8
9	6.87	6.87	9	9
9	6.87	6.87	9	9
10	6.87	6.87	10	10
10	6.87	6.87	10	10

`missForest`

Feat1	Feat2	Feat3	Feat4	Feat5
1	1	1	1	1
1	1	1	1	1
2	2	2	2	2
2	2	2	2	2
3	3	3	3	3
3	3	3	3	3
4	4	4	4	4
4	4	4	4	4
5	5	5	5	5
5	5	5	5	5
6	6	6	6	6
6	6	6	6	6
7	7	7	7	7
7	7	7	7	7
8	8	8	8	8
8	8	8	8	8
9	9	9	9	9
9	9	9	9	9
10	10	10	10	10
10	10	10	10	10

`imputePCA`

$\Rightarrow$ Imputation inherits from the method: RF (computationaly costly) good for non linear relationships / PCA good for linear relationships

MNAR data: identifiability issues, few solutions in practice

Before estimation, we should prove the identifiability of the parameters

Example: Credit: Ilya Shpitser $X^{\mathrm{NA}} = [1, \mathrm{NA}, 0, 1, \mathrm{NA}, 0]$.

- **Case 1:** X missing only if $X = 1$.

$$X = [1, 1, 0, 1, 1, 0],\ \mathbb{P}(X = 1) = 2/3.$$

- **Case 2:** X missing only if $X = 0$.

$$X = [1, 0, 0, 1, 0, 0],\ \mathbb{P}(X = 1) = 1/3.$$

$\Rightarrow$ Start from 2 equal observed distribution. It leads to different parameters of the data distribution $\mathbb{P}(X = 1)$.
Identifiability: the parameters of (X, M) are uniquely determined from available information $(X, M = 0)$.

Estimation: restrictive setting (few variables, only missing values on the outcome, simple models) [15] [16] [17]

[15] Ibrahim, et al. Missing covariates in glm when the mechanism is non-ignorable. *JRSSB*. 1999.
[16] Tang. Statistical inference for nonignorable missing-data. *Statistic. theory & rel. fields*. 2018.
[17] Mohan, Thoemmes, Pearl. Estimation with incomplete data: The linear case. *IJCAI*. 2018.

Low rank estimation/imputation with MNAR data [19], [20]

MAR (ignorable): maximize the observed penalized log-likelihood

$$\hat{\mu} \in \operatorname{argmin}_{\mu} \|(X - \mu) \odot M\|_2^2 + \lambda \|\mu\|_\star,$$

Algo: iterative soft-thresholding SVD (ISTA), accelerated version: FISTA

MNAR (non ignorable) $L(\mu, \phi; x_{\text{obs}}, m) = \int p(x; \mu) p(m|x; \phi) dx_{\text{mis}}$.

MNAR missing-data mechanism via a Logistic Model

$$p(M_{ij}|x_{ij}; \phi) = [(1 + e^{-\phi_{1j}(x_{ij} - \phi_{2j})})^{-1}]^{(1-M_{ij})} [1 - (1 + e^{-\phi_{1j}(x_{ij} - \phi_{2j})})^{-1}]^{M_{ij}}$$

$\rightsquigarrow$ self-masked MNAR : the lack only depends on the value itself.

- **E-step:** Monte-Carlo approximation and SIR algorithm.
- **M-step:** μ: `softImpute`, FISTA, ϕ: Newton-Raphson algorithm.

Not MIWAE [18]

[18] Ipsen et al. not-MIWAE: Deep Generative Modelling with MNAR Data *ICLR2021*.

[19] Sportisse, J. Low-rank estimation with missing non at random data. *Stat. & Computing*.2018.

[20] Sportisse, Boyer, J. Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data. *Neurips2020.*

Take home message inference & imputation

- Few implementation of EM strategies

- **"Imputation is both seductive & dangerous** *(Dempster & Rubin, 1983). Seductive because it can lull the user into the pleasant state of believing that the data are complete after all & dangerous because it lumps together situations where the problem is minor enough to be handled in this way & situations where estimators applied to the imputed data have substantial biases."*

- **Multiple imputation** aims at estimating the parameters and their variability taking into account **the uncertainty of the missing values**

- **Single imputation** aims to complete data as best as possible.
⇒ Principal components/**low rank powerful** for heterogeneous data; useful for clustering, exploratory multivariate analysis (correspondence analysis with NA)
⇒ Sustained implementations (R `missMDA`, python (Udell): GLRM, gcimpute)
  - Single imputation can be appropriate for point estimates

- **Both % of NA & structure matter (5% of NA can be an issue)**

Challenges and on-going works in inference & imputation

The methods used are methods implemented in a sustainable way

⇒ Challenges with multiple imputation

- Selecting one model/variable [21],[22]
- Aggregating lasso regressions. Alternatives EM [23]
- Theory with other asymptotics, i.e. small n, large p?, MNAR
- High dimension? **Computational costly** [24]: Multitask reg. (Jeff. Näf)

⇒ What to do when you have both MCAR, MAR, MNAR in the data?

⇒ Federated learning with missing values

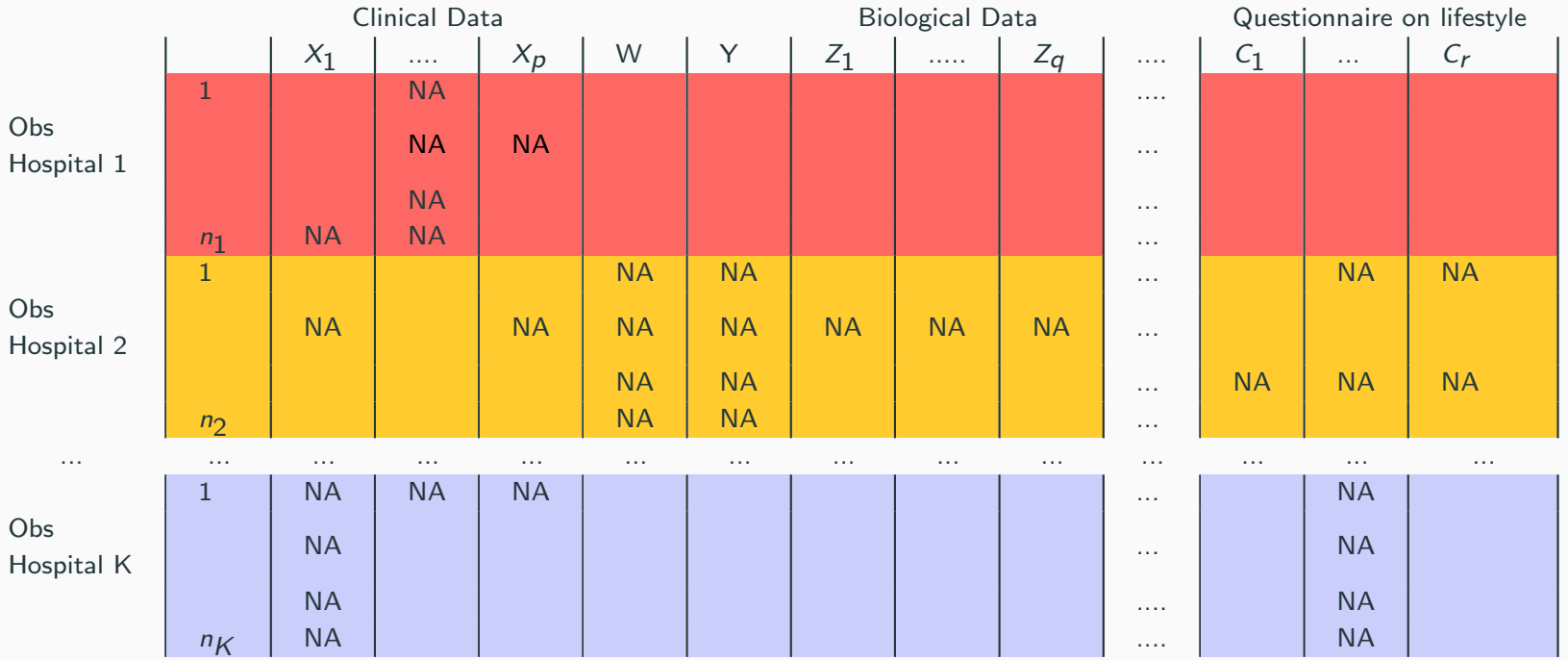
[21] Laqueur et al. SuperMICE: An Ensemble Machine Learning Approach to MICE. *Am J Epidemiol*. 2022.

[22] Jarret et al. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. *ICML*. 2022.

[23] Bogdan, J. et al. Adaptive Bayesian SLOPE - High dimensional Model Selection with Missing Values. *JCGS*. 2020.

[24] Improvement on mice pmm for large sample size, see mice github repo - still costly for large d

Challenges with heterogeneous sources and missing data



Classical methodologies are not designed to handle high-dimensional data with selection biais and informative missing data.

Challenges with heterogeneous sources and missing data

Ex: Predict the treatment effect from an RCT to a target population (distributional shift). [25], [26]

RCTs $\mathcal{R}$ & observational data $\mathcal{O}$ with different covariates: separate MIs, Joint MIs ?

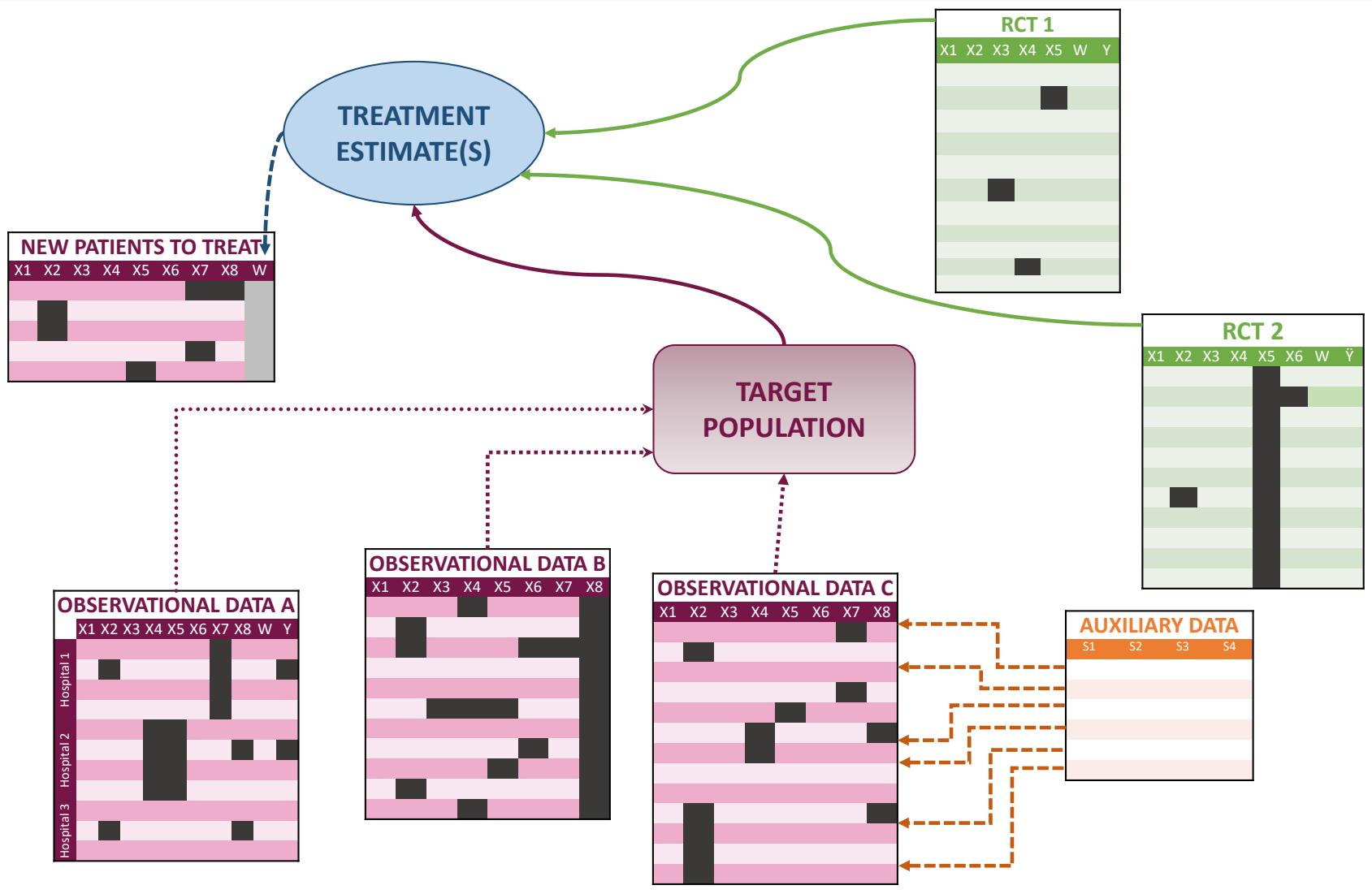
	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{R}$	1	1.1	20	5.4	1	24.1
...	$\mathcal{R}$	1		...		...	...
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3
n	$\mathcal{R}$	1	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	NA	-2	52	7.1	NA	NA
$n+2$	$\mathcal{O}$	NA	-1	35	2.4	NA	NA
...	$\mathcal{O}$	NA		...		NA	NA
$n+m$	$\mathcal{O}$	NA	-2	22	3.4	NA	NA

Data with observed treatment W and outcome Y only in the RCT.

[25] Mayer & J. Generalizing treatment effects with incomplete covariates. Archiv. 2022.

[26] Colnet, J. et al. Generalizing a causal effect: sensitivity analysis and missing covariates. *In revision in journal of causal inference*. 2022.

Challenges with heterogeneous sources and missing data



Supervised learning with missing values

Supervised learning with missing values

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

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

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

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

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

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

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?

Supervised learning with missing values

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?

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 completed data

- Separate imputat. Impute train & test separately (with a different model)
- Group imputation/ semi-supervised Impute train & test simultaneously but the predictive model is learned only on the training imputed data
- Imputation train & test with the same model. For instance, compute the means on the observed data $(\hat{\mu}_1, ..., \hat{\mu}_d)$ of each column of the train set & impute the test set with the same means

Bayes optimality of impute-n-regress [27]

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,

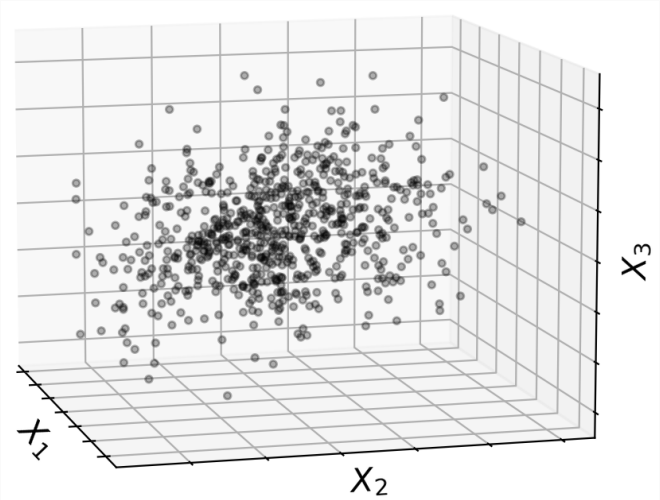
for all missing data mechanisms & almost all imputation functions,
$g^\star_\Phi \circ \Phi$ is Bayes optimal

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

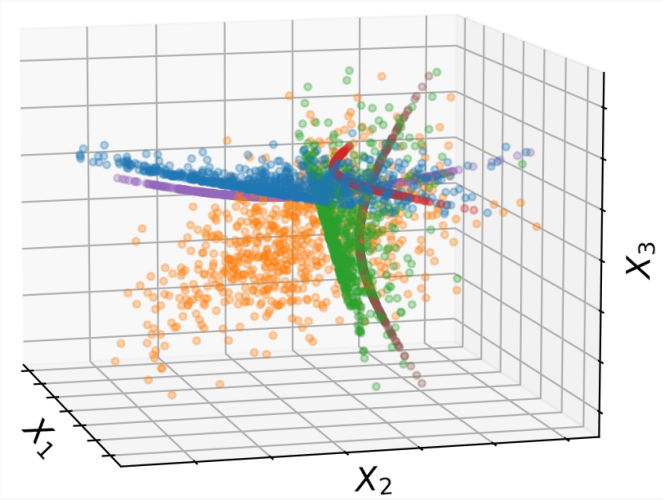
Asymptotically, imputing well is not needed to predict well

[27] Le morvan, J. et al. What's a good imputation to predict with missing values? Neurips2021 (Oral).

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



Complete data



Imputed data (manifolds)

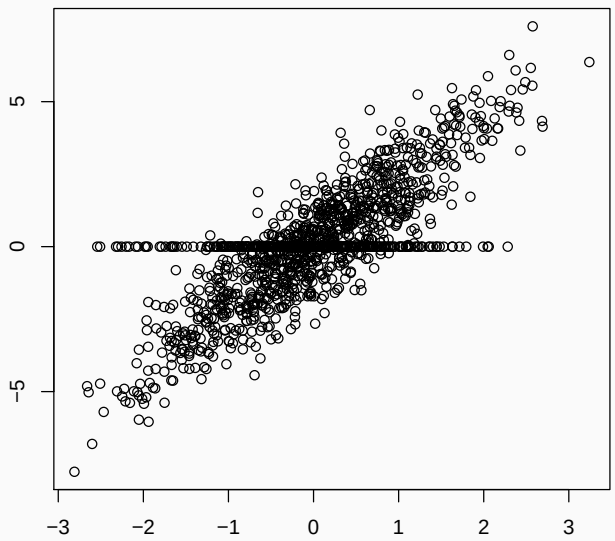
Rationale: Imputation create manifolds to which the learner adapts

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) [28]
3. Given 2), we can build prediction functions, independent of m, that are Bayes optimal for all missing data patterns

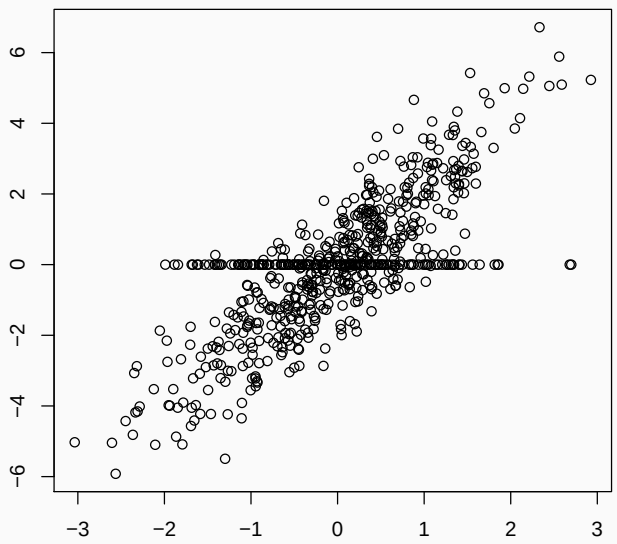
[28] Non transverse: the manifolds on which the data with either x1 missing or x2 missing are projected are exactly the same (the same line)

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

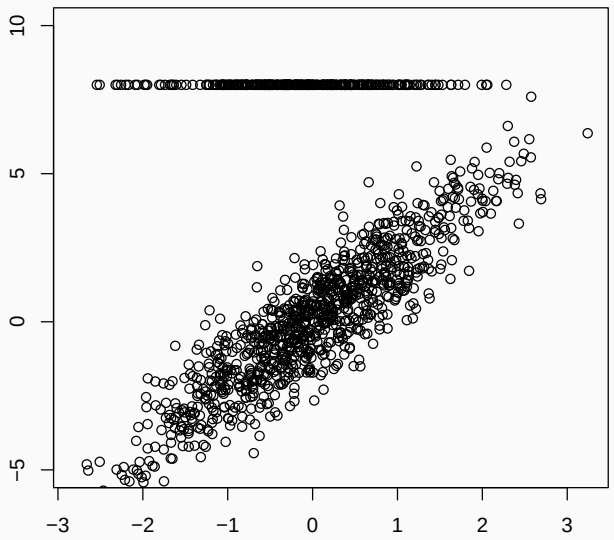


Test

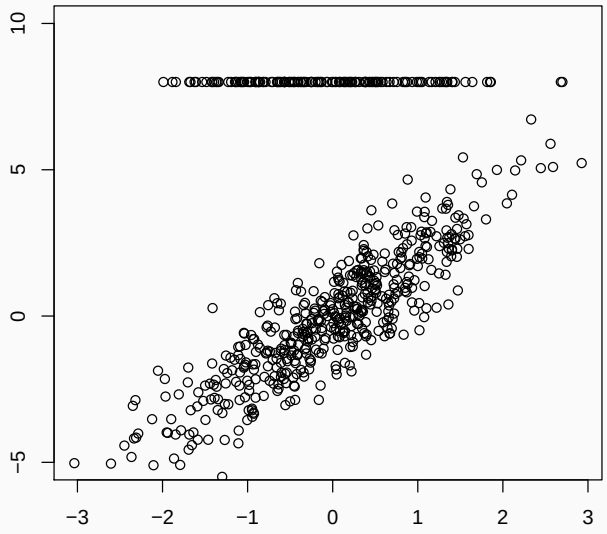
Imputing both train and test with the mean of train is consistent ie it converges to the best possible prediction, despite its drawbacks for estimation - Useful in practice!

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



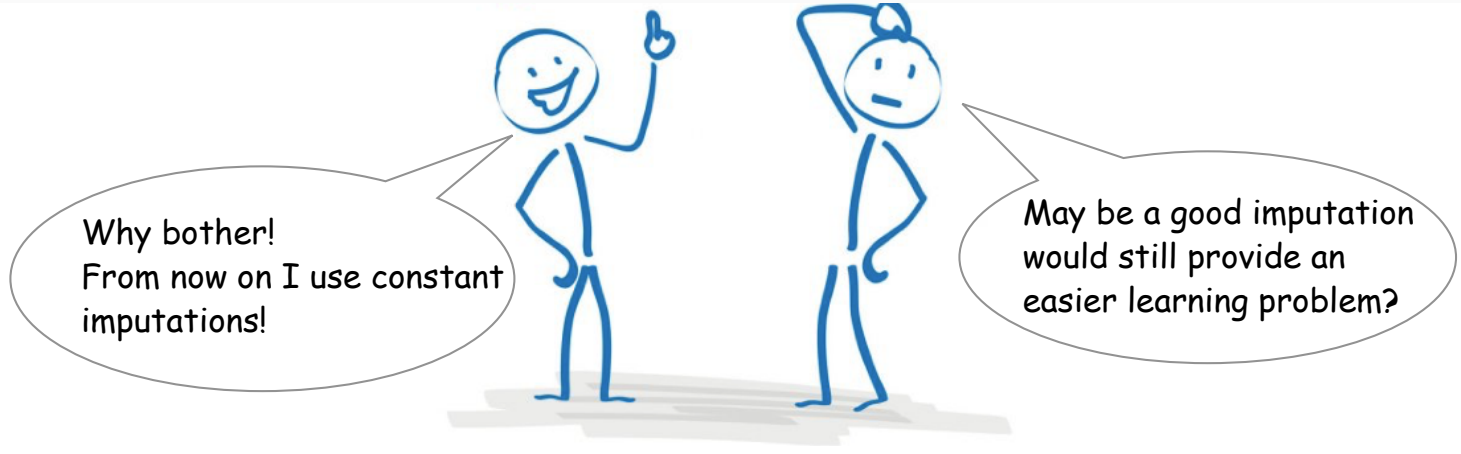
Train

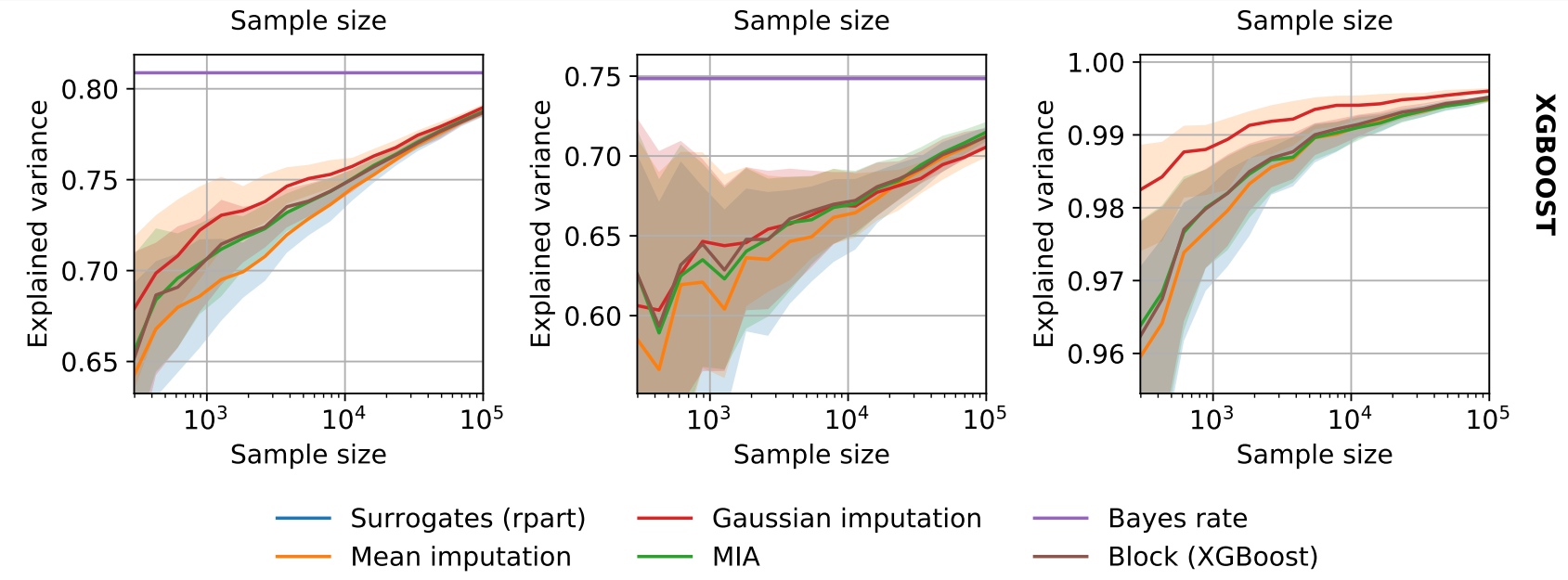


Test

Imputing both train and test with the mean of train is consistent ie it converges to the best possible prediction, despite its drawbacks for estimation - Useful in practice!

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 $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$

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

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 $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$

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

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

$\Rightarrow$ 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 $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$

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

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

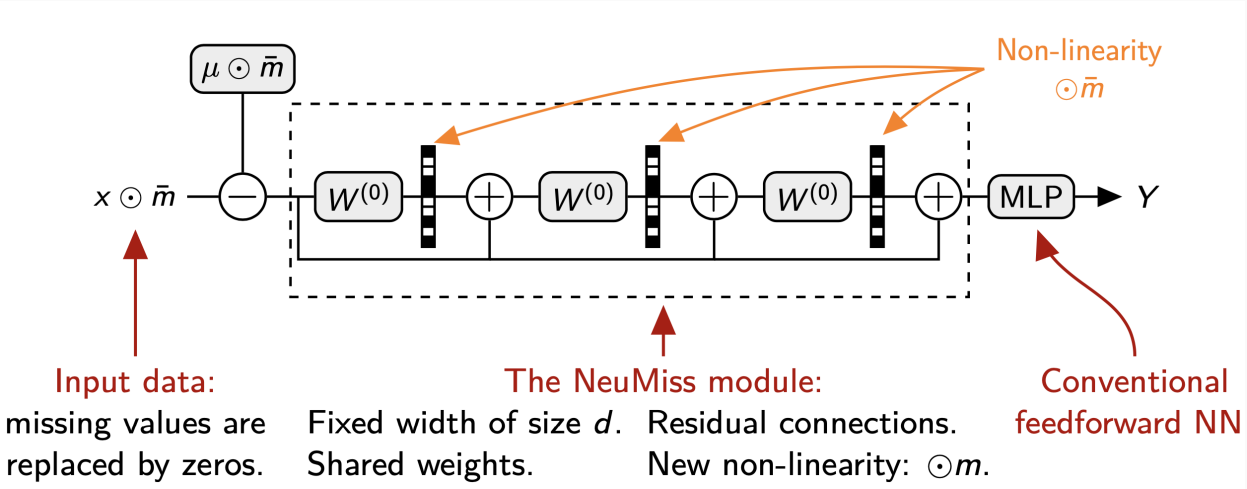
• **Optimizing imputations for a fixed regression function.** Keeping $f^\star$, is there a continuous imputation function Φ s.t $f^\star \circ \Phi$ is Bayes optimal?

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

Best imputation is joint learn with regression

- Neumiss network: [29], [30]
  - Motivated by linear regression with missing values in the covariates
  - Theoritically grounded: approximation of the Bayes predictor (truncated neumiss series to approximate inverses of covariance matrices)
  - Classic network with multiplications by the mask nonlinearities $\odot M$
- Couple Neumiss and MLP to jointly learn imputation and regression



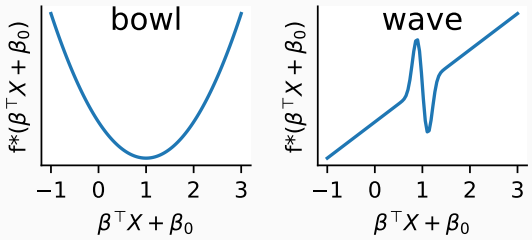
[29] Le morvan, J. et al. Linear predictor on linearly-generated data with missing values: non consistency and solutions. *AISTAT2020*.

[30] Le morvan, J. et al. Neumiss networks: differential programming for supervised learning with missing values. *Neurips2020 (Oral)*.

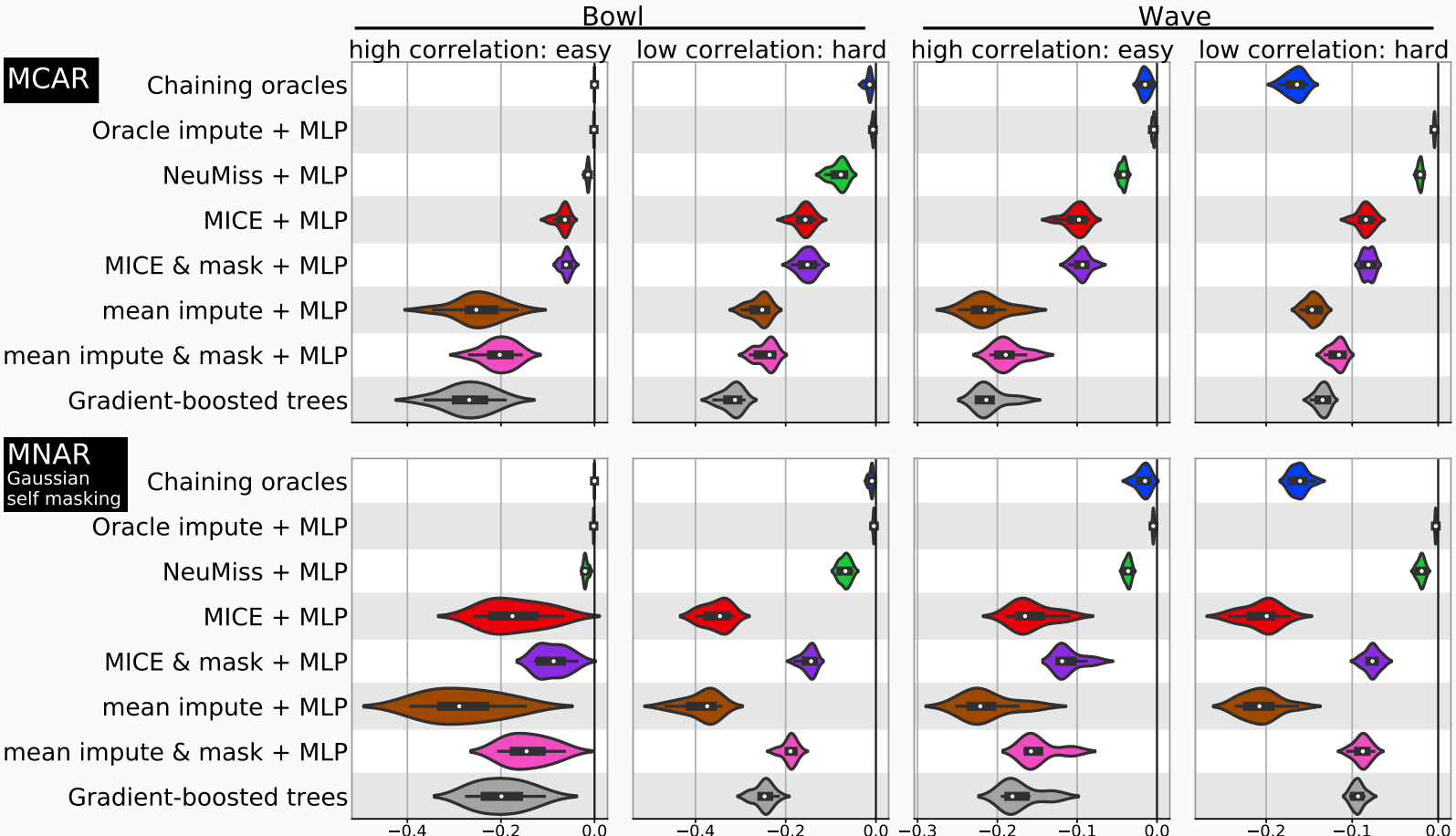
Experimental results

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

Gaussian X: "high/ low' correlation



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



Take home message in supervised learning with missing values

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with 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

Take home message in supervised learning with missing values

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with 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**

Take home message in supervised learning with missing values

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with 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 (+identifiability issues)

Challenges and on-going works with missing values

$\Rightarrow$ On-going works

- Superlearner (aggregation)
- Optimal policy (best dose of Fresh Frozen Plasma for each patient)
- Dynamic treatment regimes (who to treat & when)
- Confidence in machine learning algorithms

$\Rightarrow$ Challenges

- Distributional shifts in the missing values
- SGD with NA under MAR and MNAR in logistic regression? [31]
- Times series with MNAR (predict intubation given online monitoring, features measured each 15 minutes/1 hour + clinical data

- No benchmark datasets
- Devils in the details: scaling?

[31] Sportisse, J. et al. Debiasing Stochastic Gradient Descent to handle missing values. *Neurips2020*.

Collaborators on missing values

- F. Husson, Professor Agronomy University. (package `missMDA`, `FactoMineR`)
- Gosia Bogdan, Professor Wroclaw. High dimensional regression
- Claire Boyer, Associate Professor Sorbonne. Signal, missing values
- Imke Mayer, Postdoc Charité Institute, Berlin. Causal inference
- Aude Sportisse, Postdoc Inria Nice. Missing values
- Marine Le Morvan, Junior researcher at INRIA, Paris. Supervised learning
- Erwan Scornet, Asso. Prof. at Ecole Polytechnique, Paris. Random forests
- Gael Varoquaux, Senior researcher at INRIA, Paris. ML, Scikit-learn