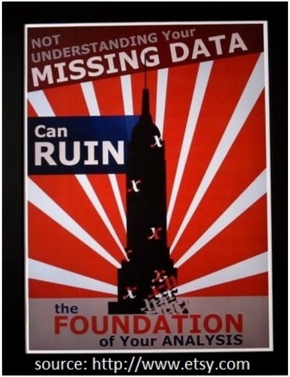
A missing value tour

Julie Josse

Ecole Polytechnique, INRIA

26 january 2020

Workshop of the Applied Machine Learning Days 2020, Lausanne



Overview

1. Introduction
2. Handling missing values (inferential framework)
3. Supervised learning with missing values
4. Discussion - challenges

Introduction

Collaborators

- PhD students - postdocs: W. Jiang, M. Le Morvan, I. Mayer, G. Robin (former), A. Sportisse
- Colleagues: C. Boyer (LPSM), G. Bogdan (Wroclaw), F. Husson (Agrocampus) - (package `missMDA`), J-P Nadal (EHESS), E. Scornet (X), G. Varoquaux (INRIA), S. Wager (Stanford)
- Traumabase (hospital): T. Gauss, S. Hamada, J-D Moyer/ Capgemini

Traumabase

- 20000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

Traumabase

- 20000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

$\Rightarrow$ **Estimate causal effect**: Administration of the **treatment** "tranexamic acid" (within 3 hours after the accident) on the **outcome** mortality for traumatic brain injury patients

Traumabase

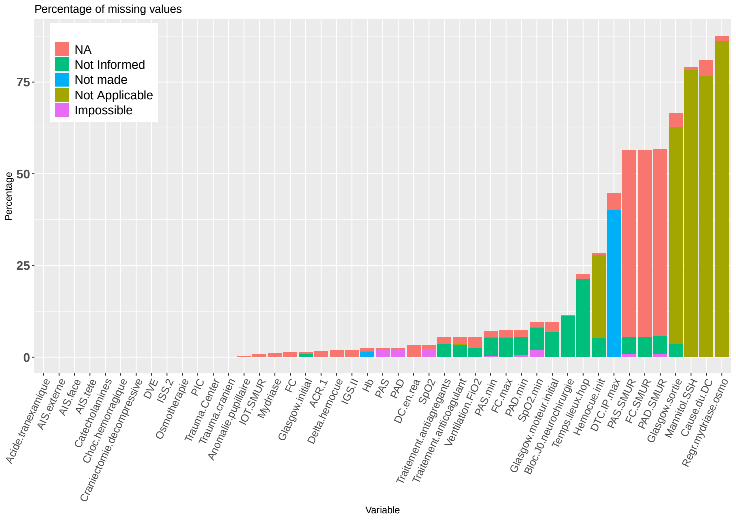
- 20000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

$\Rightarrow$ **Predict** the risk of hemorrhagic shock given pre-hospital features

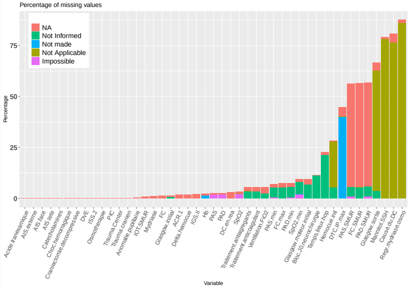
Ex random forests/logistic regression with covariates with missing values

Missing values



Multilevel data/ data integration: Systematic missing variable in one hospital

Complete-case analysis



`?lm, ?glm, na.action = na.omit`

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

An $n \times p$ matrix, each entry is missing with probability 0.01

$p = 5 \quad \Longrightarrow \approx 95\%$ of rows kept

$p = 300 \Longrightarrow \approx \ \ 5\%$ of rows kept

Handling missing values (inferential framework)

Solutions to handle missing values

Books: Schafer (2002), Little & Rubin (2002); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability

Ex logistic regression: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Aim: **Estimate parameters & their variance** from an incomplete data
$\Rightarrow$ Inferential framework

Solutions to handle missing values

Books: Schafer (2002), Little & Rubin (2002); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability

Ex logistic regression: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Cons: Difficult to establish - not many softwares even for simple models
One specific algorithm for each statistical method...

Aim: **Estimate parameters & their variance** from an incomplete data
$\Rightarrow$ Inferential framework

Solutions to handle missing values

Books: Schafer (2002), Little & Rubin (2002); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability

Ex logistic regression: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Cons: Difficult to establish - not many softwares even for simple models

One specific algorithm for each statistical method...

Imputation (multiple) to get a complete data set

Any analysis can be performed

Ex logistic regression: Impute and apply logistic model to get $\hat{\beta}$, $\hat{V}(\hat{\beta})$

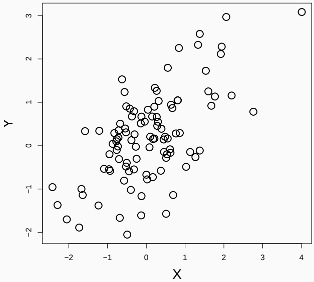
Aim: **Estimate parameters & their variance** from an incomplete data

$\Rightarrow$ Inferential framework

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$

X	Y
-0.56	-1.93
-0.86	-1.50
.....	...
2.16	0.7
0.16	0.74



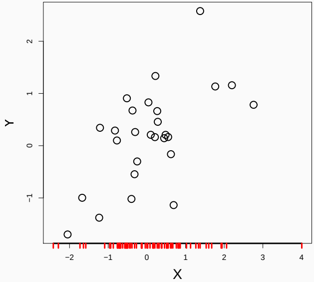
$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

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

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y

X	Y
-0.56	NA
-0.86	NA
.....	...
2.16	0.7
0.16	NA



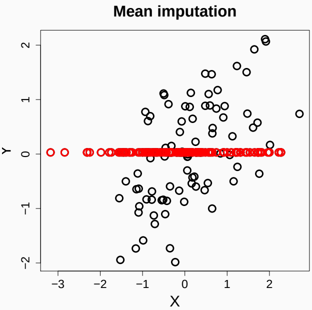
$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

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

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y
- Estimate parameters on the mean imputed data

X	Y
-0.56	**0.01**
-0.86	**0.01**
.....	...
2.16	0.7
0.16	**0.01**



$\mu_y = 0$

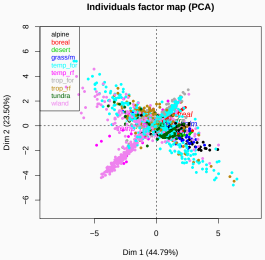
$\sigma_y = 1$

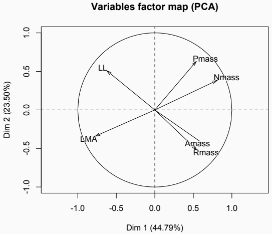
$\rho = 0.6$

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

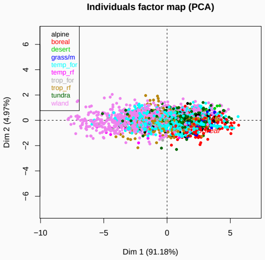
Mean imputation deforms joint and marginal distributions

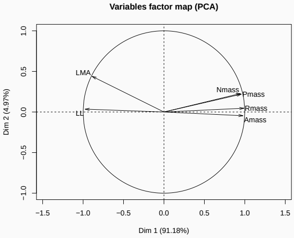
Mean imputation is bad for estimation






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






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

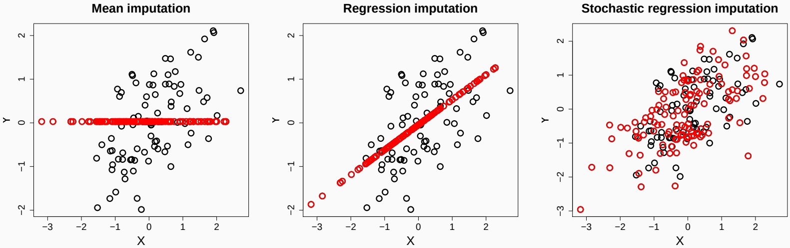
Ecological data: [1] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

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

Imputation methods

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

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



	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Imputation methods for multivariate data

Assuming a joint model

- Gaussian distribution: $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$ (`Amelia` Honaker, King, Blackwell)
- low rank: $X_{n \times d} = \mu_{n \times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k
  (`softimpute` Hastie & Mazuder; `missMDA` J. & Husson)
- latent class - nonparametric Bayesian (`dpmpm` Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018)

Using conditional models (joint implicitly defined)

- with logistic, multinomial, poisson regressions (`mice` van Buuren)
- iterative impute each variable by random forests (`missForest` Stekhoven)

Imputation for categorical, mixed, blocks/multilevel data [2], etc.
$\Rightarrow$ Missing values taskview[3] J., Mayer., Tierney, Vialaneix

[2] J., Husson, Robin & Narasimhan. (2018). Imputation of mixed data with multilevel SVD.
[3] https://cran.r-project.org/web/views/MissingData.html

Random forests versus 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

Single imputation: Underestimation of 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: Underestimation of 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: Underestimation of the variability

⇒ 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

⇒ 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

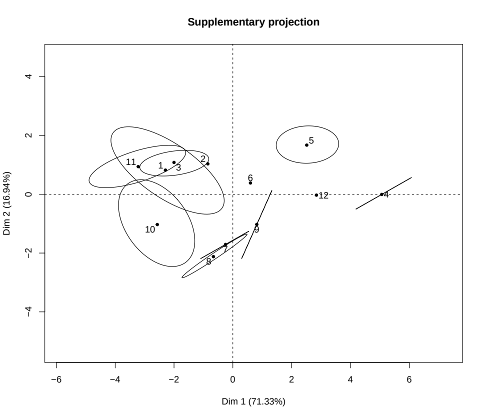
```
library(mice); mice(traumadata)
library(missMDA); MIPCA(traumadata)
```

Visualization of the imputed values

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)
library(Amelia)
?compare.density
```

Percentage of NA?

Multiple imputation

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

Supervised learning with missing values

On the consistency of supervised learning with missing values. (2019). J., Prost, Scornet & Varoquaux

- A feature matrix $\mathbf{X}$ and a response vector Y
- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^\star \in \underset{f:\,\mathcal{X}\rightarrow\mathcal{Y}}{\arg\min}\, \mathbb{E}\left[\ell(f(\mathbf{X}), Y)\right]; \quad f^\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$
- Empirical risk: $\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\mathcal{X}\rightarrow\mathcal{Y}}{\arg\min}\, \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(f(\mathbf{X}_i), Y_i\right)\right)$

  A new data $\mathcal{D}_{n,\mathrm{test}}$ to estimate the generalization error rate
- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(\mathbf{X}), Y)] \xrightarrow[n\rightarrow\infty]{} \mathbb{E}[\ell(f^\star(\mathbf{X}), Y)]$

On the consistency of supervised learning with missing values. (2019). J., Prost, Scornet & Varoquaux

- A feature matrix $\mathbf{X}$ and a response vector Y
- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^\star \in \underset{f:\,\mathcal{X}\to\mathcal{Y}}{\arg\min}\ \mathbb{E}\left[\ell(f(\mathbf{X}), Y)\right]; \quad f^\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$

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

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

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

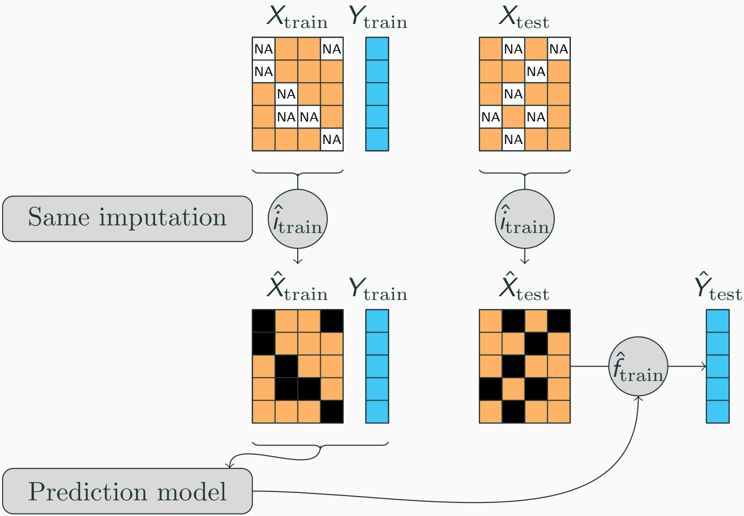
Differences with classical litterature

- explicitely consider the response variable Y - Aim: Prediction
- two data sets (out of sample) with missing values: Train & test sets

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

Imputation prior to learning

Impute the train with $\hat{\imath}_{train}$ learn a model $\hat{f}_{train}$ with $\hat{X}_{\mathrm{train}}, Y_{\mathrm{train}}$
Impute the test with the same imputation $\hat{\imath}_{train}$ - predict $\hat{X}_{\mathrm{test}}$ with $\hat{f}_{train}$



Imputation prior to learning

Imputation with the same model

Easy to implement for univariate imputation: The means $(\hat{\mu}_1, ..., \hat{\mu}_d)$ of each colum of the train. Also OK for Gaussian imputation.

Issue: Many methods are "black-boxes" and take as an input the incomplete data and output the completed data (`mice`, `missForest`)

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 with the same model: Mean imputation consistent

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**

Framework - assumptions

- $Y = f(\mathbf{X}) + \varepsilon$
- $\mathbf{X} = (X_1, \ldots, X_d)$ 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 $(\mathbf{X}, M_1)$

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

Imputation with the same model: Mean imputation consistent

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**

Mean imputed entry $\mathbf{x}' = (x'_1, x_2, \ldots, x_d)$: $x'_1 = x_1 \mathbb{1}_{M_1=0} + \mathbb{E}[X_1]\mathbb{1}_{M_1=1}$

Note the data: $\widetilde{\mathbf{X}} = \mathbf{X} \odot (\mathbf{1} - \mathbf{M}) + \mathtt{NA} \odot \mathbf{M}$ (takes value in $\mathbb{R} \cup \{\mathtt{NA}\}$)

Theorem

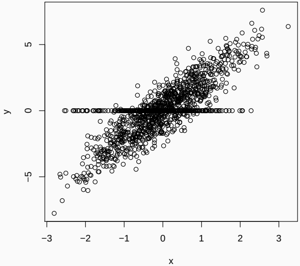
Prediction with mean is equal to the Bayes function almost everywhere

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

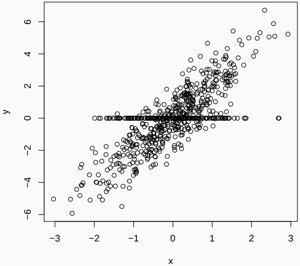
Other values than the mean are OK but use the same value for the train and test sets, otherwise the algorithm may fail as the distributions differ

Consistency of supervised learning with NA: 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

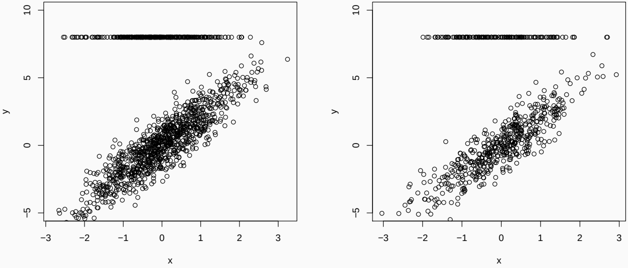


Test

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

Consistency of supervised learning with NA: Rationale

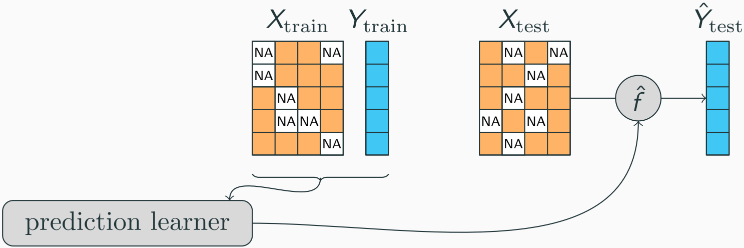
- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant: out of range
- Need a lot of data (asymptotic result) and a super powerful learner



Train Test

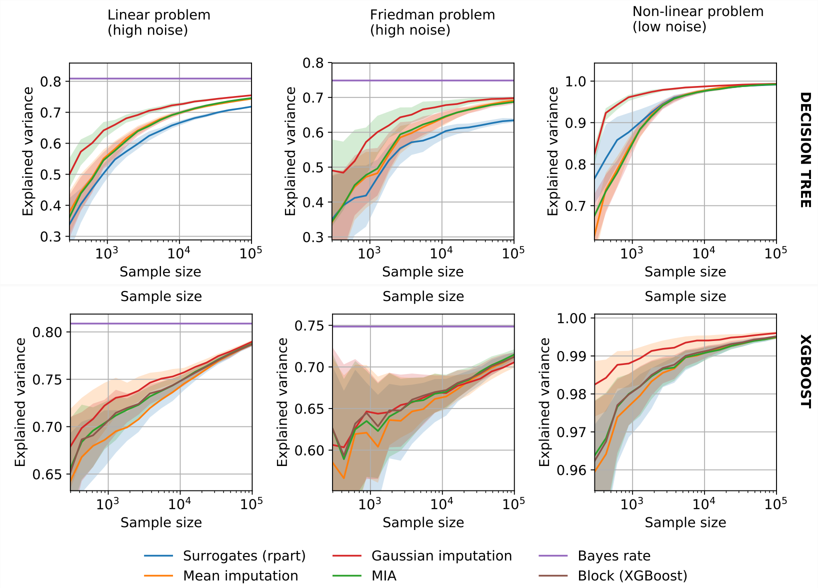
Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

End-to-end learning with missing values



- Trees well suited for empirical risk minimization with missing values: Handle half discrete data $\tilde{\mathbf{X}}$ that takes values in $\mathbb{R} \cup \{\mathtt{NA}\}$
- Random forests powerful learner

Consistency: 40% missing values MCAR



Discussion - challenges

Take home message EM/imputation

- Few implementation of EM strategies

"**The idea of imputation is both seductive and dangerous**". *It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the imputed data have substantial biases." (Dempster & Rubin, 1983)*

- Single imputation aims at completing a dataset as best as possible
- **Multiple imputation** aims at estimating the parameters and their variability taking into account the uncertainty of the missing values
- Single imputation can be appropriate for point estimates
- Both % of NA & structure matter (5% of NA can be an issue)

Principal component methods powerful for single & multiple imputation of quanti & categorical data: Dimensionality reduction and capture similarities between observations and variables. `missMDA package`

Take-home message supervised learning

- Incomplete train and test $\rightarrow$ **same imputation model**
- **Single mean imputation is consistent given a powerful learner**
- Empirically, good imputation methods reduce the number of samples required to reach good prediction

Tree-based models :

- **Missing Incorporated in Attribute** optimizes not only the split but also the handling of the missing values
- Informative missing data: **Adding the mask** helps imputation - MIA

To be done

- Nonasymptotic results
- Uncertainty associated with the prediction
- Distributional shift: No missing values in the test set?
- Prove the usefulness of methods in MNAR

Still an active area of research! Join this exciting field!

Current works

- Variable selection in high dimension Adaptive bayesian SLOPE with missing values. 2019. Jiang, Bogdan, J., Miasojedow, Rockova & TraumaBase
- **MNAR missing values**
  - Contribution of causality for missing data
  - Graphical Models for Processing Missing Data. 2019. Mohan, Pearl.
  - Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data. 2019. Sportisse, Boyer, J.
  - Contribution of neural nets J., Prost, Scornet, Varoquaux

Other challenges

- MI theory: Good theory for regression parameters but others? Theory with other asymptotic small n, large p ?, etc.
- Practical imputation issues: Imputation not in agreement (X & X^2), imputation out of range? problems of logical bounds (> 0), etc.

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)[4]

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!

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

Ressources

Examples:

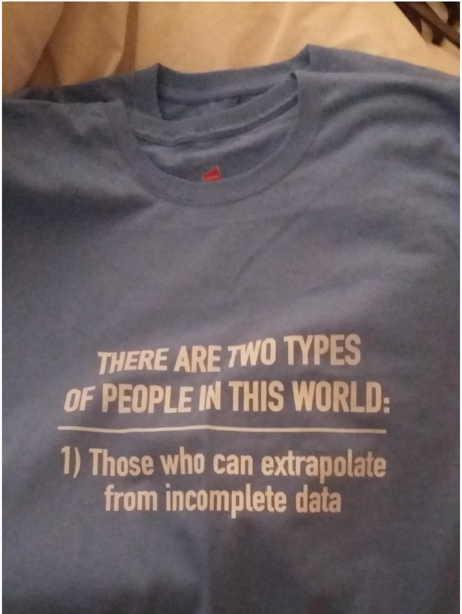
- Lecture [5] - General tutorial : Statistical Methods for Analysis with Missing Data (Mauricio Sadinle)
- Lecture - Multiple Imputation: `mice` by Nicole Erler [6]
- Longitudinal data, Time Series Imputation (Steffen Moritz - very active contributor of r-miss-tastic), Principal Component Methods[7]

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

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

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

Thank you


THERE ARE TWO TYPES
OF PEOPLE IN THIS WORLD:
1) Those who can extrapolate
from incomplete data