A missing values tour in R
with a special focus on parameter estimation

Wei Jiang
Ecole Polytechnique

R meetup @ Wroclaw

December 18, 2019

Missing values

When we attempt to explore data as a source of knowledge, **missing values** lies in the process of obtaining, recording, and preparing the data.

- Unanswered questions in a survey
- loss of data
- machines that fail

"We should be suspicious of any dataset (large or small) which appears perfect." – David J. Hand

Paris Hospitals - TraumaBase dataset

20 000 severely traumatised patients + 250 measurements

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	....
1	97	35.6	NA	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NA	14.4	15	no	
7	100	36.6	NA	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NA	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

....

Paris Hospitals - TraumaBase dataset

20 000 severely traumatised patients + 250 measurements

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	....
1	97	35.6	NA	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NA	14.4	15	no	
7	100	36.6	NA	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NA	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

....

⇒ Predict the Glasgow score, whether to start a blood transfusion, etc...

⇒ Linear regression / **Logistic regression** /Random Forests with missing covariates

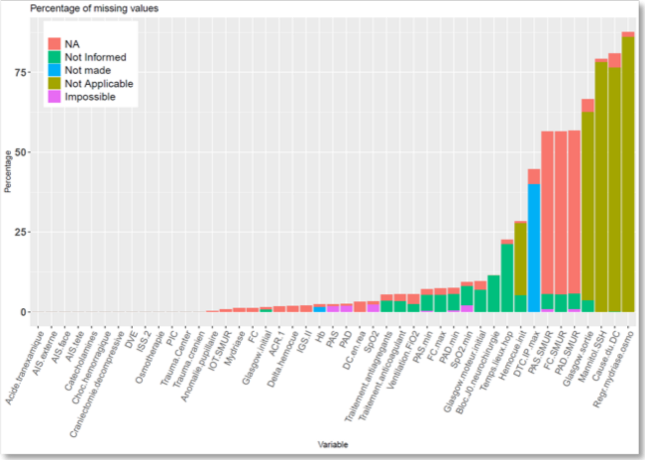
Missing values problematic

List-wise deletion (default `lm` function in R)
$\Rightarrow$ loss of information

Missing values problematic

List-wise deletion (default `lm` function in R)

⇒ loss of information



⇒ less than 10% remained

Single imputation

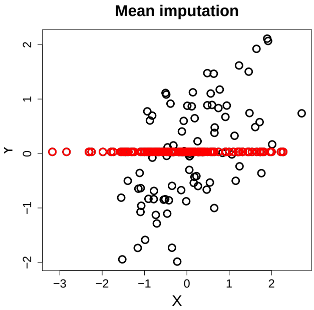
- $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$ *i.i.d.*
- 70% missing entries on y randomly

Date completion by the mean of observed values in y
$\Rightarrow$ Estimate parameters:

Single imputation

- $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$ *i.i.d.*
- 70% missing entries on y randomly

Date completion by the mean of observed values in y
$\Rightarrow$ Estimate parameters:



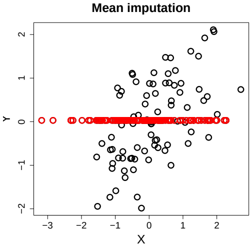
$$\mu_y = 0 \quad \sigma_y = 1 \quad \rho = 0.6$$

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

$\Rightarrow$ Biased estimates

Imputation methods

- Mean imputation

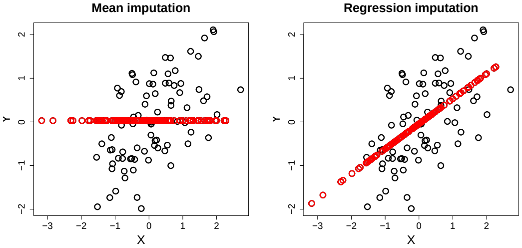


$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

0.01
0.5
0.30

Imputation methods

- Mean imputation
- Impute by regression: impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  $\Rightarrow$ variance underestimated and correlation overestimated.

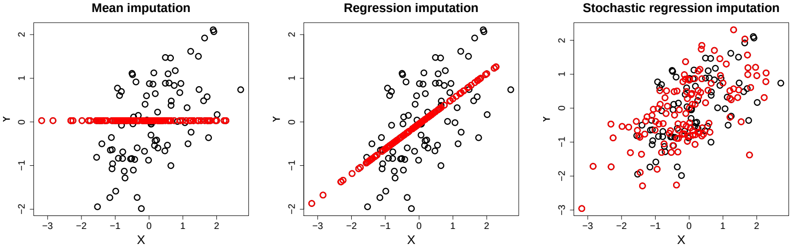


$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

0.01	0.01
0.5	0.72
0.30	0.78

Imputation methods

- Mean imputation
- Impute by regression: impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  $\Rightarrow$ variance underestimated and correlation overestimated.
- Impute by stochastic regression: impute $\hat{y}_i \sim \mathcal{N}\left(x_i\hat{\beta}, \hat{\sigma}^2\right)$
  $\Rightarrow$ preserve distribution



$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Missing pattern and mechanism

Q: How about real dataset?

Dealing with missing values depends on:

- the pattern of missing values
- the mechanism leading to missing values

R packages: `VIM`, `naniar` (Matthias Templ, Nick Tierney)
`FactoMineR` (YouTube)

Ozone data set

112 daily records of meteorological variables (wind speed, temperature, rainfall, etc.) and ozone concentration recorded in Rennes

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	NA	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	17	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0622	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0642	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

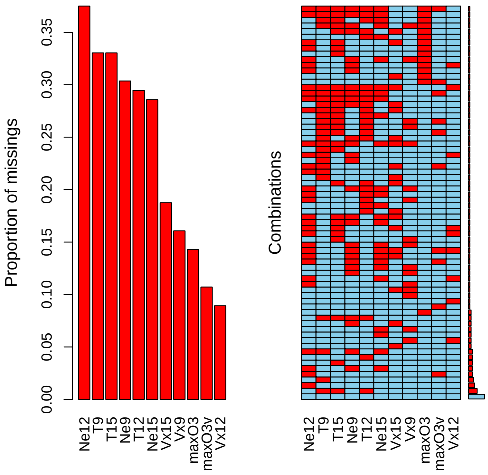
http://www.airbreizh.asso.fr/

Aim: complete ozone

Count missing values

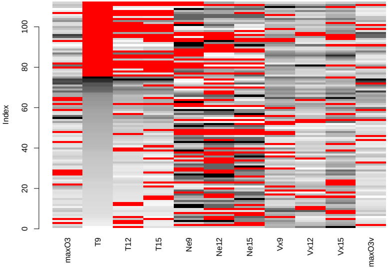
```
> library(missMDA)
> WindDirection <- ozo[,12]
> don <- ozo[,1:11]
> library(VIM)
> res <- summary(aggr(don, sortVar = TRUE))$combinations
> res[rev(order(res[, 2])),]

Variables sorted by
number of missings:                    Combinations Count    Percent
Variable        Count         0:0:0:0:0:0:0:0:0:0:0    13 11.6071429
    Ne12 0.37500000           0:1:1:1:0:0:0:0:0:0:0     7  6.2500000
      T9 0.33035714           0:0:0:0:0:1:0:0:0:0:0     5  4.4642857
     T15 0.33035714           0:1:0:0:0:0:0:0:0:0:0     4  3.5714286
     Ne9 0.30357143           0:1:0:0:1:1:1:0:0:0:0     3  2.6785714
     T12 0.29464286           0:0:1:0:0:0:0:0:0:0:0     3  2.6785714
    Ne15 0.28571429           0:0:0:1:0:0:0:0:0:0:0     3  2.6785714
    Vx15 0.18750000           0:0:0:0:1:1:1:0:0:0:0     3  2.6785714
     Vx9 0.16071429           0:0:0:0:0:1:0:0:0:0:1     3  2.6785714
   maxO3 0.14285714           0:1:1:1:1:0:0:0:0:0:0     2  1.7857143
  maxO3v 0.10714286           0:0:0:0:1:0:0:0:0:1:0     2  1.7857143
    Vx12 0.08928571           0:0:0:0:0:0:1:1:0:0:0     2  1.7857143
                              0:0:0:0:0:0:1:0:0:0:0     2  1.7857143
                              .....................     .  ...
```

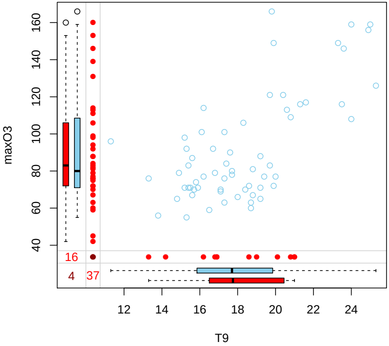
Pattern visualization




```
 #library(VIM)
> aggr(don, sortVar = TRUE)
```

Visualization





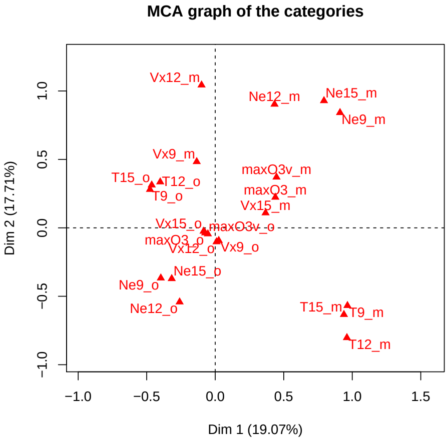

```
# library(VIM)
> matrixplot(don, sortby = 2)
> marginplot(don[ ,c("T9", "maxO3")])
```

Visualization with Multiple Correspondence Analysis

⇒ Create the missingness matrix

```
> mis.ind <- matrix("o", nrow = nrow(don), ncol = ncol(don))
> mis.ind[is.na(don)] = "m"
> dimnames(mis.ind) = dimnames(don)
> mis.ind

         maxO3 T9  T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
20010601 "o"   "o" "o" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010602 "o"   "m" "m" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010603 "o"   "o" "o" "o" "o" "m"  "m"  "o" "m"  "o"  "o"
20010604 "o"   "o" "o" "m" "o" "o"  "o"  "m" "o"  "o"  "o"
20010605 "o"   "m" "o" "o" "m" "m"  "m"  "o" "o"  "o"  "o"
20010606 "o"   "o" "o" "o" "o" "m"  "o"  "o" "o"  "o"  "o"
20010607 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
20010610 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
```

Visualization with Multiple Correspondence Analysis

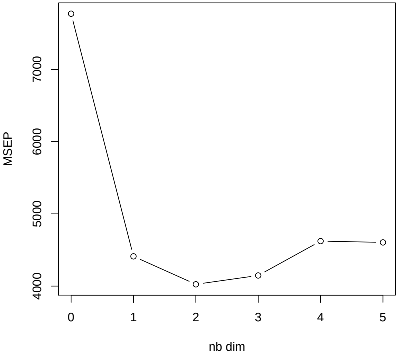



```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA, invis = "ind", title = "MCA graph of the categories")
```

Imputation with Principal Component Analysis in practice

⇒ Step 1: Estimation of the number of dimensions
(Cross Validation)

```
> library(missMDA)
> nb <- estim_ncpPCA(don, method.cv = "Kfold")
> nb$ncp    #2
> plot(0:5, nb$criterion, xlab = "nb dim", ylab ="MSEP")
```



Imputation with PCA in practice

⇒ Step 2: Imputation of the missing values

```
> res.comp <- imputePCA(don, ncp = 2)
> res.comp$completeObs[1:3, ]
     maxO3    T9   T12   T15 Ne9 Ne12 Ne15   Vx9  Vx12  Vx15 maxO3v
0601    87 15.60 18.50 20.47   4 4.00 8.00  0.69 -1.71 -0.69     84
0602    82 18.51 20.88 21.81   5 5.00 7.00 -4.33 -4.00 -3.00     87
0603    92 15.30 17.60 19.50   2 3.98 3.81  2.95  1.97  0.52     82
```

Complete ozone

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
20010601	87.000	15.600	18.500	20.471	4.000	4.000	8.000	0.695	-1.710	-0.695	84.000
20010602	82.000	18.505	20.870	21.799	5.000	5.000	7.000	-4.330	-4.000	-3.000	87.000
20010603	92.000	15.300	17.600	19.500	2.000	3.984	3.812	2.954	1.951	0.521	82.000
20010604	114.000	16.200	19.700	24.693	1.000	1.000	0.000	2.044	0.347	-0.174	92.000
20010605	94.000	18.968	20.500	20.400	5.294	5.272	5.056	-0.500	-2.954	-4.330	114.000
20010606	80.000	17.700	19.800	18.300	6.000	7.020	7.000	-5.638	-5.000	-6.000	94.000
20010607	79.000	16.800	15.600	14.900	7.000	8.000	6.556	-4.330	-1.879	-3.759	80.000
20010610	79.000	14.900	17.500	18.900	5.000	5.000	5.016	0.000	-1.042	-1.389	99.000
20010611	101.000	16.100	19.600	21.400	2.000	4.691	4.000	-0.766	-1.026	-2.298	79.000
20010612	106.000	18.300	22.494	22.900	5.000	4.627	4.495	1.286	-2.298	-3.939	101.000
20010613	101.000	17.300	19.300	20.200	7.000	7.000	3.000	-1.500	-1.500	-0.868	106.000
.....											
20010915	69.000	17.100	17.700	17.500	6.000	7.000	8.000	-5.196	-2.736	-1.042	71.000
20010916	71.000	15.400	18.091	16.600	4.000	5.000	5.000	-3.830	0.000	1.389	69.000
20010917	60.000	15.283	18.565	19.556	4.000	5.000	4.000	0.000	3.214	0.000	71.000
20010918	42.000	14.091	14.300	14.900	8.000	7.000	7.000	-2.500	-3.214	-2.500	60.000
20010919	65.000	14.800	16.425	15.900	7.000	7.982	7.000	-4.341	-6.062	-5.196	42.000
20010920	71.000	15.500	18.000	17.400	7.000	7.000	6.000	-3.939	-3.064	0.000	65.000
20010924	76.000	13.300	17.700	17.700	5.631	5.883	5.453	-0.940	-0.766	-0.500	65.139
20010925	75.573	13.300	18.434	17.800	3.000	5.000	5.001	0.000	-1.000	-1.286	76.000
20010927	77.000	16.200	20.800	20.499	5.368	5.495	5.177	-0.695	-2.000	-1.473	71.000
20010928	99.000	18.074	22.169	23.651	3.531	3.610	3.561	1.500	0.868	0.868	93.135
20010929	83.000	19.855	22.663	23.847	5.374	5.000	3.000	-4.000	-3.759	-4.000	99.000
20010930	70.000	15.700	18.600	20.700	7.000	6.405	7.000	-2.584	-1.042	-4.000	83.000

```
> library(missMDA)
> res.comp <- imputePCA(ozo[, 1:11])
> res.comp$comp
```

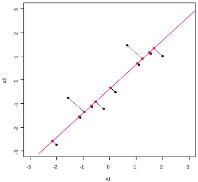
PCA reconstruction

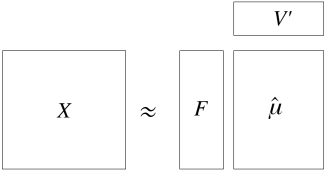
X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq S \right\}$$

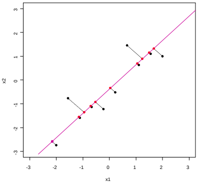
PCA reconstruction

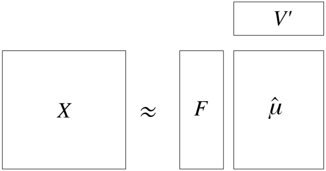
X

-2.00	-2.74
NA	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	NA
0.22	-0.52
0.67	1.46
NA	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq S \right\}$$

$\Rightarrow$ PCA with missing values: weighted least squares

$$\mathrm{argmin}_\mu \left\{ \|W_{n\times p} * (X - \mu)\|_2^2 : \mathrm{rank}\,(\mu) \leq S \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise

References on more imputation methods

- PCA or MCA
  R package: `missMDA`
- *k*-nearest neighbor
  R packages: `VIM`, `yaImpute`, `impute`
- random forest
  R package: `missForest`
- chained equation (conditional distribution)
  R packages: `mice`

⇒ R-miss-tastic (Josse et al.): Methods and references for managing missing data
⇒ Flexible imputation of missing data (Stef van Buuren)

Recommended methods

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates

Recommended methods

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates

Eample: Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$, point estimates with EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

Recommended methods

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates

Eample: Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$, point estimates with EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

$\Rightarrow$ Natural model selection procedure!
$\Rightarrow$ One specific algorithm for each statistical method.
$\Rightarrow$ Not many implementations even for simple models.

Recommended methods

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates

Eample: Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$, point estimates with EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

$\Rightarrow$ Natural model selection procedure!
$\Rightarrow$ One specific algorithm for each statistical method.
$\Rightarrow$ Not many implementations even for simple models.

Specialized focus on: Logistic regression with missing covariates,

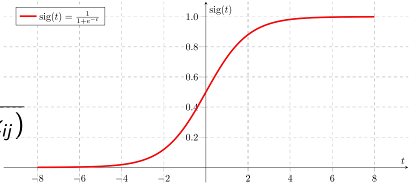
joint work with Julie Josse, Marc Lavielle and TraumaBase Group

Logistic regression model

$x = (x_{ij})$ a $n \times p$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model

$$\mathbb{P}\left(y_i = 1 | x_i; \beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

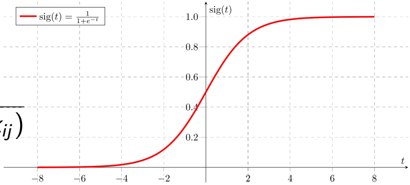


Logistic regression model

$x = (x_{ij})$ a $n \times p$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model

$$\mathbb{P}\left(y_i = 1 | x_i; \beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$



Covariables

$$x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$$

Log-likelihood for complete-data with the set of parameters $\theta = (\mu, \Sigma, \beta)$

$$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^{n} \Big(\log(\mathrm{p}(y_i | x_i; \beta)) + \log(\mathrm{p}(x_i; \mu, \Sigma)) \Big).$$

Missing data mechanism

Decomposition: $x = (x_{\mathrm{obs}}, x_{\mathrm{mis}})$.

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

Missing data mechanism

Decomposition: $x = (x_{\text{obs}}, x_{\text{mis}})$.

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

- Missing completely at random (MCAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R)$$

- Missing at random (MAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R|x_{\text{obs}})$$

- Missing not at random (MNAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R|x_{\text{mis}})$$

Example: age and income.

Missing data mechanism

Decomposition: $x = (x_{\text{obs}}, x_{\text{mis}})$.

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

- Missing completely at random (MCAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R)$$

- Missing at random (MAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R|x_{\text{obs}})$$

- Missing not at random (MNAR):

$$\mathrm{p}(R|x_{\text{obs}}, x_{\text{mis}}) = \mathrm{p}(R|x_{\text{mis}})$$

Example: age and income.

Assumption: Missing data are **Missing at Random**

$\Rightarrow$ Ignore missing mechanism when doing inferences.

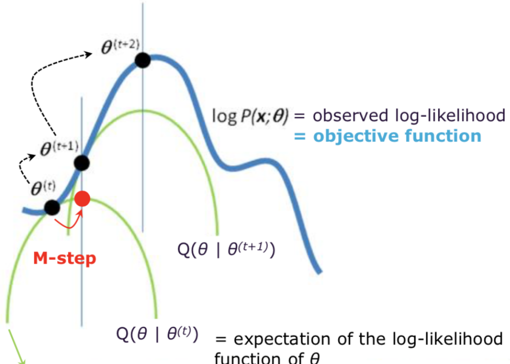
EM algorithm with missing data

Aim: argmax $\mathcal{LL}(\theta; x_{\text{obs}}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\text{mis}}$.

- **E-step:** Evaluate the quantity

$$Q_k(\theta) = \mathbb{E}[\mathcal{LL}(\theta; x, y) | x_{\text{obs}}, y; \theta_{k-1}]$$
$$= \int \mathcal{LL}(\theta; x, y) \text{p}(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1}) dx_{\text{mis}}.$$

- **M-step:** $\theta_k = \text{argmax}_\theta \, Q_k(\theta)$.



EM algorithm with missing data

Aim: $\text{argmax}\,\mathcal{LL}(\theta; x_{\text{obs}}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\text{mis}}$.

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\mathcal{LL}(\theta; x, y) | x_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \mathcal{LL}(\theta; x, y) \mathrm{p}(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1}) dx_{\text{mis}}. \end{aligned}$$

- **M-step:** $\theta_k = \text{argmax}_\theta\, Q_k(\theta)$.

Unfeasible computation of expectation!

MCEM (Wei & Tanner 1990): Generate a large set of samples of missing data from $\mathrm{p}(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1})$ and replaces the expectation by an empirical mean.

Require a huge number of samples to converge!

Stochastic Approximation EM

(book, Lavielle 2014) Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $x_{i,\mathrm{mis}}^{(k)}$ from

$$\mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}, y_i; \theta_{k-1}).$$

- **Stochastic approximation:** Update the function Q

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\mathrm{obs}}, x_{\mathrm{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$

  where (γ_k) is a decreasing sequence of positive numbers.

- **Maximization:** $\theta_k = \mathrm{argmax}_\theta \, Q_k(\theta)$.

Stochastic Approximation EM

(book, Lavielle 2014) Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $x_{i,\mathrm{mis}}^{(k)}$ from

$$\mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}, y_i; \theta_{k-1}).$$

- **Stochastic approximation:** Update the function Q

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\mathrm{obs}}, x_{\mathrm{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$

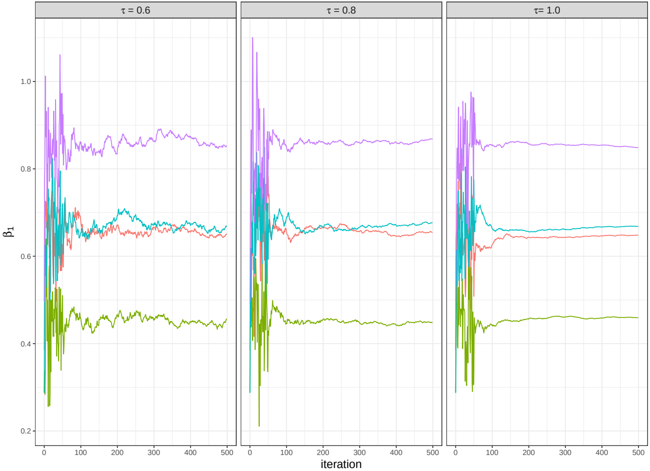
  where (γ_k) is a decreasing sequence of positive numbers.

- **Maximization:** $\theta_k = \mathrm{argmax}_\theta \, Q_k(\theta)$.

Convergence: (Allassonniere et al. 2010)
The choice of the sequence (γ_k) is important for ensuring the almost sure convergence of SAEM to a MLE.

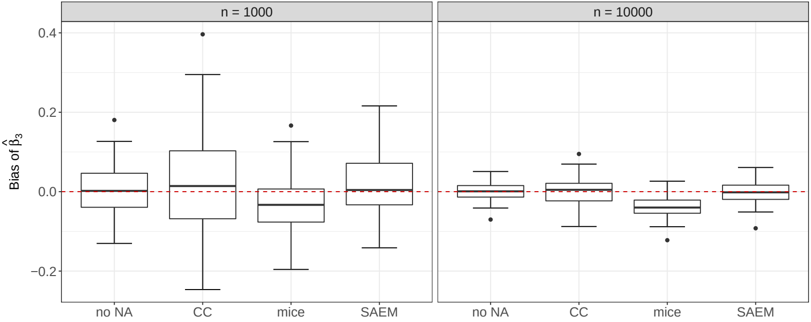
Simulation study: SAEM behavior

Step size : $\gamma_k = (k - k_1)^{-\tau}$
$k_1 = 50$ and $\tau = (0.6, 0.8, 1.0)$.
$N = 1000$, $p = 5$, percentage of missingness= 10%
4 repetitions of simulations and 500 iterations:



Comparison with competitors: estimates

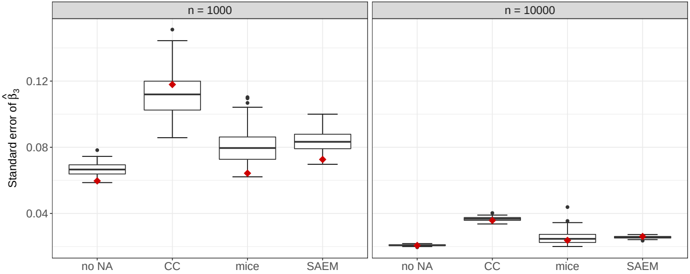
x: $p = 5$, $n = 1000$ / $n = 10\,000 \Rightarrow y \in \{0, 1\}$
percentage of missingness = 10%.
Repeat 1000 times for each setting.



Comparison with competitors : coverage

Table: Coverage (%) for $n = 10\,000$, calculated over 1000 simulations.

parameter	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	**86.5**	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	**89.6**	94.7



Comparison with competitors : execution time

Table: Comparison of execution time between no NA, MCEM, mice, and SAEM with $n = 1000$ calculated over 1000 simulations.

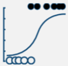
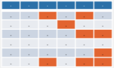
Execution time (seconds)	no NA	MCEM	mice	SAEM
min	2.87×10^{-3}	492	0.64	9.96
mean	4.65×10^{-3}	773	0.70	13.50
max	43.50×10^{-3}	1077	0.76	16.79

Application on TraumaBase

Age
Weight
Height
BMI
Glasgow
Systolic BP
Diastolic BP
Heart Rate
Hb Hemocue
SpO_2
Volume Expander
Pulse Pressure

- 14 continuous variables
- Gaussian distribution assumption

Logistic regression with missing values

 + 

Hemorrhagic shock

$P(y = 1 \mid X; \hat{\beta})$?

Exploration of dataset

Data preprocessing ⇒ **6384 patients** in the dataset.
Clinical experience ⇒ **14 influential quantitative measurements**
The percentage of missingness of some variables varies form 0 to **60%**, which indicates the importance of analysis of missing data.

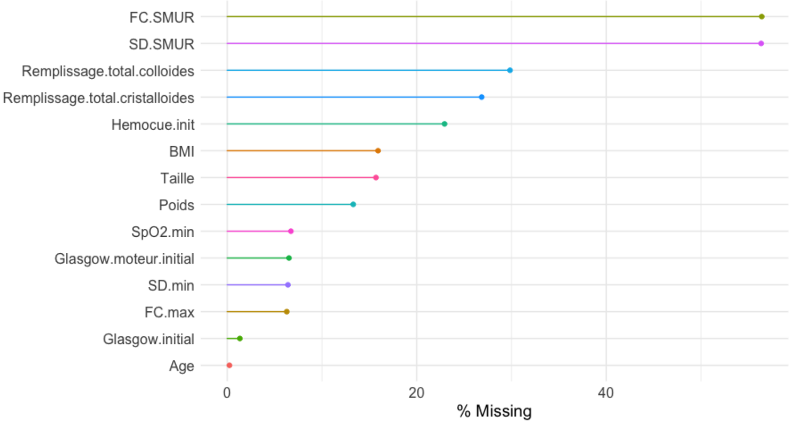


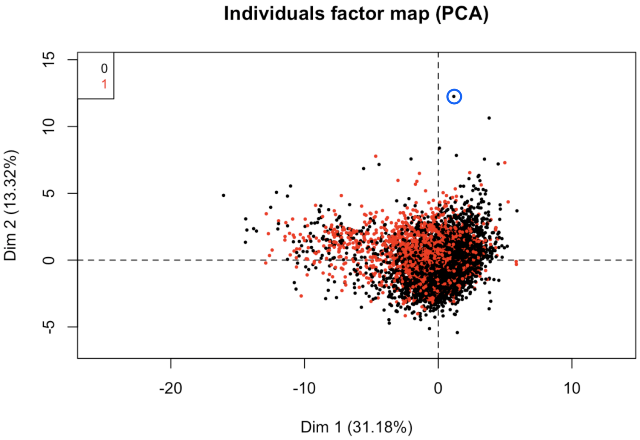
Figure: Percentage of missing information in each variable.

Exploration of dataset

Based on **penalized observed log-likelihood**

⇒ Observations resulting in a very small value of the log-likelihood.

⇒ wrong records



Estimation and interpretation

Estimation and model selection:

Variable	**Effect**	Estimate (std error)
Intercept		-0.52 (0.59)
Age	+	0.011 (0.0033)
Glasgow.moteur	-	-0.16 (0.036)
FC.max	+	0.026 (0.0025)
Hemocue.init	-	-0.23 (0.031)
RT.cristalloides	+	0.00090 (0.00010)
RT.colloides	+	0.0019 (0.00021)
SD.min	-	-0.025 (0.0050)
SD.SMUR	-	-0.021 (0.0056)

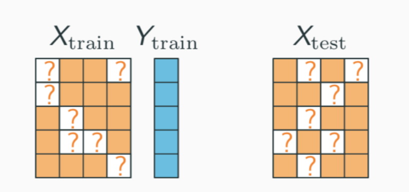
Estimation and interpretation

Estimation and model selection:

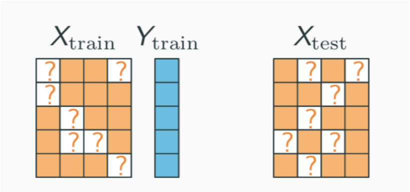
Variable	Effect	Estimate (std error)
Intercept		-0.52 (0.59)
Age	+	0.011 (0.0033)
Glasgow.moteur	-	-0.16 (0.036)
FC.max	+	0.026 (0.0025)
Hemocue.init	-	-0.23 (0.031)
RT.cristalloides	+	0.00090 (0.00010)
RT.colloides	+	0.0019 (0.00021)
SD.min	-	-0.025 (0.0050)
SD.SMUR	-	-0.021 (0.0056)

- Older people tend to have a larger possibility to suffer from hemorrhagic shock.
- A low Glasgow score means one makes no motor response, often in the case of hemorrhagic shock.

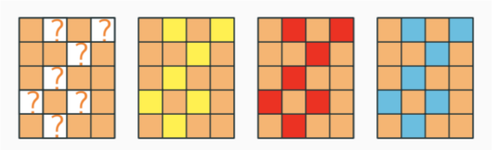
Missing values in test set



Missing values in test set



$$x_{\text{mis}}^{(1)},\, x_{\text{mis}}^{(2)},\, \cdots\, x_{\text{mis}}^{(M)} \sim \mathrm{p}\left(x_{\text{mis}} \mid x_{\text{obs}}\right) \Downarrow$$

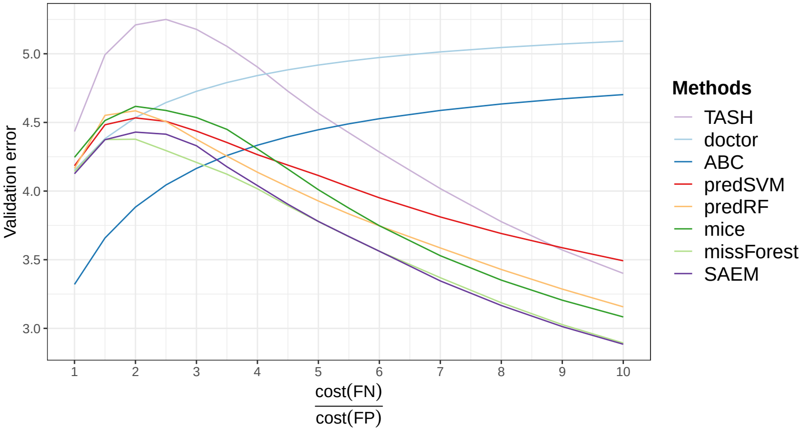


$$p_m(y) = \mathrm{p}\left(y | x_{\text{obs}}, x_{\text{mis}}^{(m)}\right): \quad p_1 \quad p_2 \quad \cdots \quad p_M$$

$$\hat{y} = \arg\max_y \mathrm{p}\left(y | x_{\text{obs}}\right) = \arg\max_y \sum_{m=1}^{M} p_m(y)$$

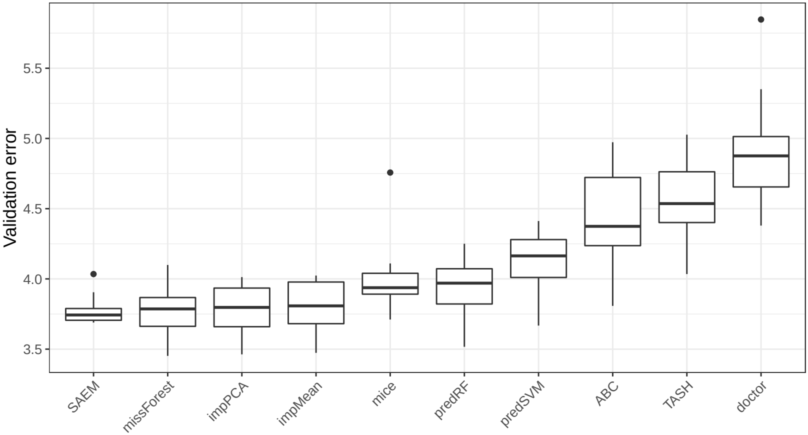
Predictive performance

Random split : training set (70%) + test set (30%) (repeated 15 times)



Predictive performance

Random split : training set (70%) + test set (30%) (repeated 15 times)



Conclusion

- SAEM for logistic regression with missingness can be computationally efficient
- Unbiased estimation and a more reasonable coverage of confidence interval than competitors
- Application in TraumaBase: good perdictive performance
- R package **misaem** & arXiv:1805.04602

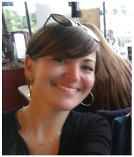
Ongoing work:

- Deal with logistic regression with missing and heterogeneous data, based on general location model
- High-dimensional model selection with missing values (joint work with Gosia Bogdan, arXiv:1909.06631, R package **ABSLOPE**)

Thank you for listening!

Dziękuję

Julie Josse



Marc Lavielle

Tobias Gauss

Sophie Hamada

Some references

Schafer (1997),

Joseph L. Schafer
Analysis of incomplete data

Little & Rubin (1987, 2002)

Roderick Little Donald Rubin
Statistical analysis with missing values

Suggested reading: chap 25 of Gelman & Hill (2006)

Andrew Gelman Jennifer L. Hill
Data Analysis Using Regression and Multilevel/Hierarchical Models

R-miss-tastic: https://rmisstastic.netlify.com/
A resource website on missing values - Methods for managing missing data