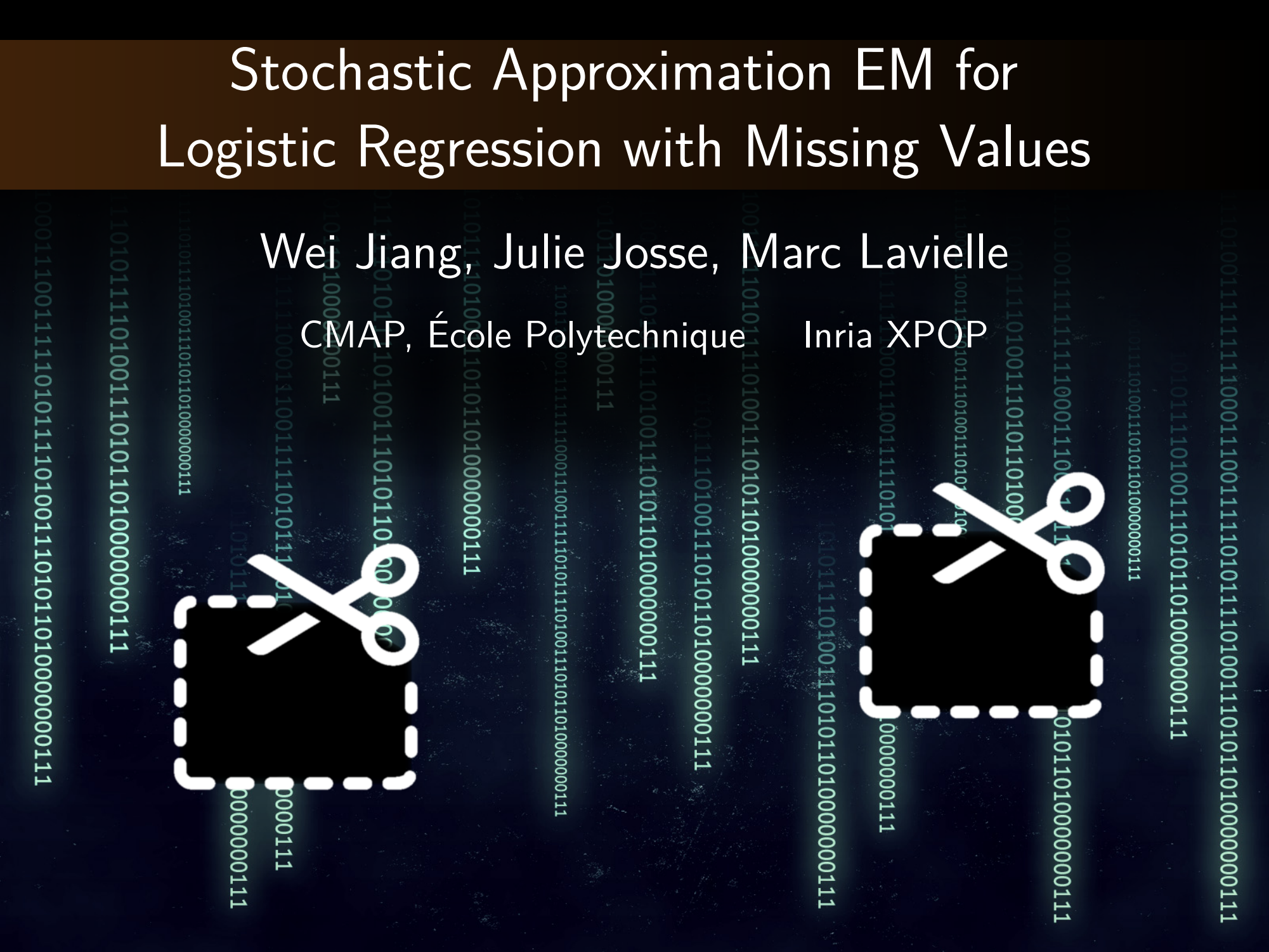
Stochastic Approximation EM for Logistic Regression with Missing Values

Wei Jiang, Julie Josse, Marc Lavielle

CMAP, École Polytechnique Inria XPOP



1 Background

public health polytraumatized patients major trauma

modeling with missing data

1 Background: Public health for trauma patients

Traumabase data:
7495 trauma patients + 244 measurements (quantitative & categorical).

1. Develop the models to help the emergency doctors to take decisions

Type of Accident	Age	Sex	Blood pressure	Lactate	Hemorrhagic shock
Falling	50	M	140	**NM**	Yes
Fire	28	F	**NR**	4.8	No
Knife	30	M	120	1.2	No
Traffic accident	23	M	110	3.6	No
Knife	33	M	106	**NM**	No
Traffic accident	58	F	150	**NM**	Yes

$$\text{Measurements of patients} \xrightarrow{\text{Predict}} \text{Hemorrhagic shock}$$

$$X \text{ mixed} \xrightarrow{\textbf{Logistic regression}} Y = 0, 1$$

Challenge: Modeling with **missing data**
(**NA** = Not Applicable, **NM** = Not Made, **NR** = Not Recorded)

1 State of the art to handle missing values

- Complete case $\Rightarrow$ loss of information $\times$
- Imputation (**mice**, **missMDA**, **missForest**)
- **Algorithm Expectation-Maximization** to get the maximum likelihood estimators + other algorithms to get the variances
  $\Rightarrow$ **Natural model selection procedure !**
  $\Rightarrow$ Difficult to establish?
  $\Rightarrow$ Not many implementations, even for simple models.

2 Logistic regression with missing data

SAEM algorithm to estimate the parameters

Estimate variance

Model selection procedure

2 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

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

Covariables

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

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

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

Assumption: Missing data are **Missing at Random**.
Decomposition: $x = (x_{\text{obs}}, x_{\text{mis}})$.
Aim: $\arg\max \mathcal{LL}(\theta; x_{\text{obs}}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\text{mis}}$.

EM:

- **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:** Update the estimation of θ:
  $\theta_k = \arg\max_\theta Q_k(\theta)$.

Unfeasible computation of expectation!

2 Stochastic Approximation EM

(book, Lavielle 2014)

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

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

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

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

  where (γ_k) is a decreasing sequence of positive numbers.
- **Maximization:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

2 Metropolis-Hastings algorithm

Target distribution

$$f_i(x_{i,\text{mis}}) = \mathrm{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta)$$
$$\propto \mathrm{p}(y_i|x_i; \beta)\, \mathrm{p}(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma).$$

Proposal distribution

$$g_i(x_{i,\text{mis}}) = \mathrm{p}(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma) \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$\mu_i = \mu_{\text{mis}} + \Sigma_{\text{mis,obs}}\Sigma_{\text{obs,obs}}^{-1}(x_{i,\text{obs}} - \mu_{\text{obs}}),$$
$$\Sigma_i = \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}}\Sigma_{\text{obs,obs}}^{-1}\Sigma_{\text{obs,mis}},$$

Metropolis:

- $z_{im}^{(k)} \sim g_i(\boldsymbol{x}_{i,mis})$, $u \sim \mathcal{U}[0,1]$
- $r = \dfrac{f_i(z_{im}^{(k)})/g_i(z_{im}^{(k)})}{f_i(z_{i,m-1}^{(k)})/g_i(z_{i,m-1}^{(k)})}$
- If $u < r$, accept $z_{im}^{(k)}$

Variance estimation with Louis formula: $\mathcal{I}_{\text{obs}} = \mathcal{I}_{\text{comp}} - \mathcal{I}_{\text{mis}}$

Model selection criterion:

$$\text{AIC}(\mathcal{M}) = -2\mathcal{LL}(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y) + 2d(\mathcal{M}),$$
$$\text{BIC}(\mathcal{M}) = -2\mathcal{LL}(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y) + \log(n)d(\mathcal{M}),$$

Observed likelihood:

$$\begin{aligned}
\mathrm{p}(y_i, x_{i,\text{obs}}; \theta) &= \int \mathrm{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta)\mathrm{p}(x_{i,\text{mis}}; \theta) dx_{i,\text{mis}} \\
&= \int \mathrm{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta)\frac{\mathrm{p}(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})} g_i(x_{i,\text{mis}}) dx_{i,\text{mis}} \\
&= \mathbb{E}_{g_i}\left(\mathrm{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta)\frac{\mathrm{p}(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})}\right).
\end{aligned}$$

Sample from g_i (proposal distribution) $\Rightarrow$ Empirical mean.

2 Implementation in R

Package: **misaem**

```
library(devtools); install_github("wjiang94/misaem")
library(misaem)
list.saem = miss.saem(X.obs, y, maxruns = 500, tol_em = 1e-07)
```

Arguments:

- X.obs: Design matrix with missingness $n \times p$;
- y: Response vector $n \times 1$.

Return a list with components:

- beta: Estimated β;
- var: Estimated variance for estimated parameters;
- logl: Observed log-likelihood.

3 Simulation study

SAEM behavior

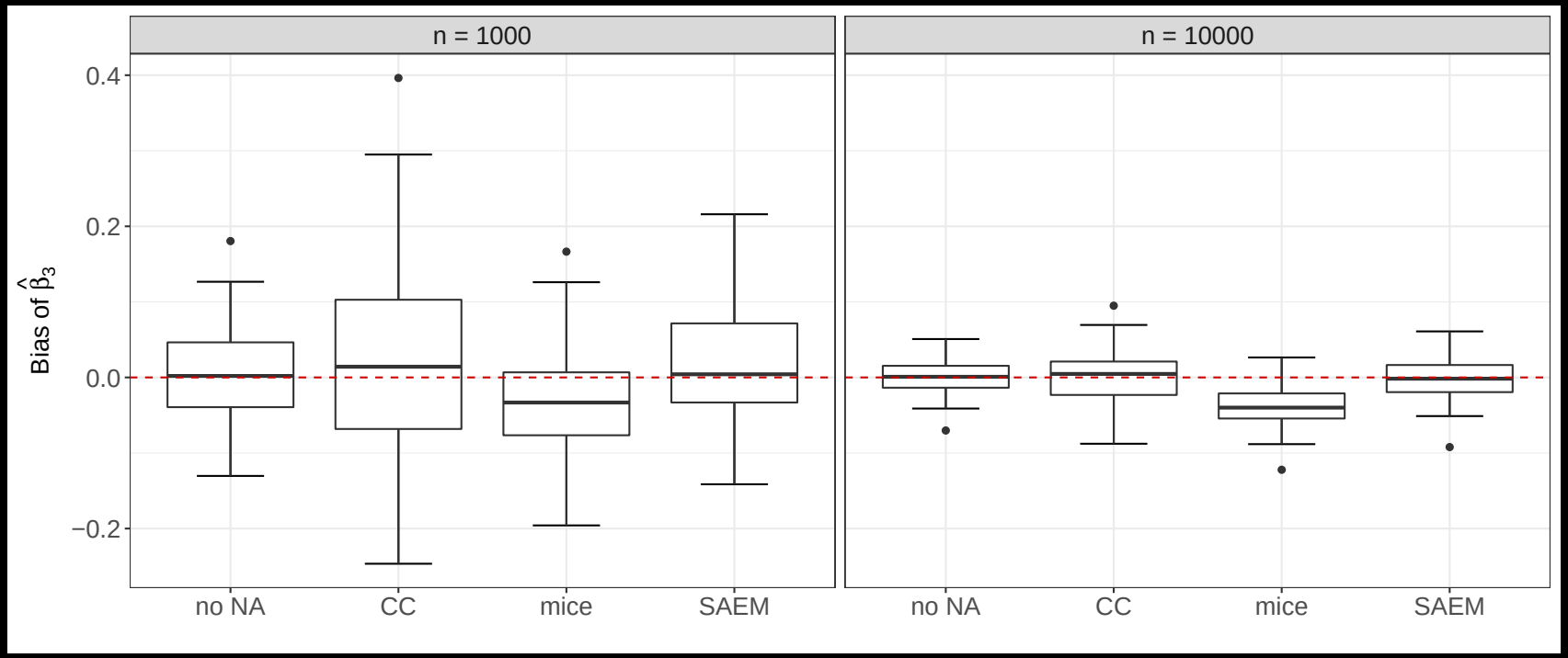
model selection effects

bias of estimation

Comparison with other methods

3 Comparison with competitors: bias

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



no NA: classical estimation on original dataset without NA;
CC: complete case analysis method;
mice: multiple imputation implemented by package *mice*.

3 Comparison with competitors: confidence interval

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

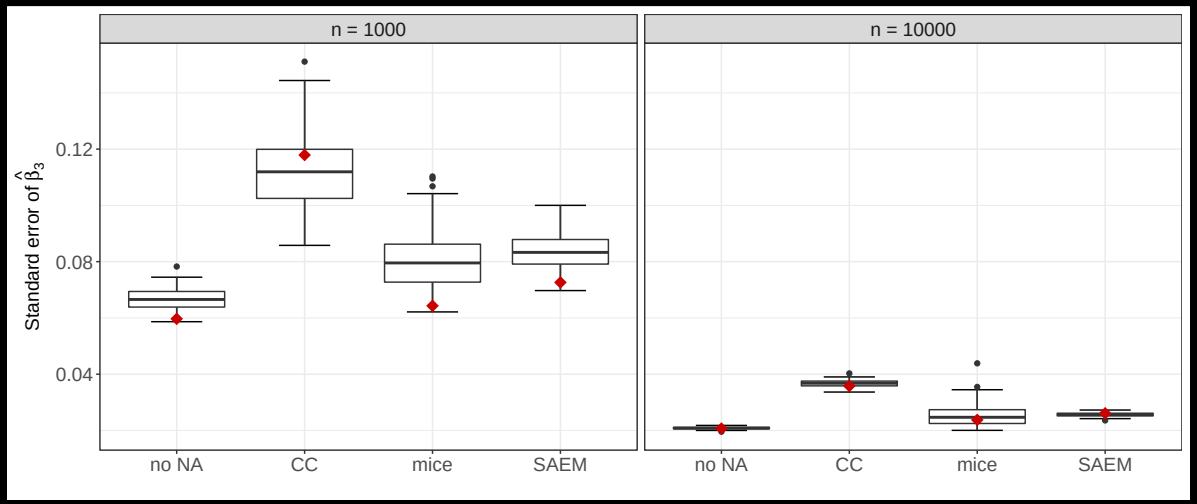


Figure: Length of confidence interval.

3 Comparison with competitors : execution time

Table: Comparison of execution time (in seconds) with $n = 1000$ calculated over 1000 simulations.

Time	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

Extract of code (MCMC):

```
for (m in (1:nmcmc)){
 xina.c <- mi + rnorm(njna)%*%chol(Oi)
 if (y[i]==1)
  alpha <- (1+exp(-sum(xina*betana))/cobs)/(1+exp(-sum(xina.c*betana))/cobs)
 else
  alpha <- (1+exp(sum(xina*betana))*cobs)/(1+exp(sum(xina.c*betana))*cobs)
 if (runif(1) < alpha){xina <- xina.c}
}
```

3 Model selection results

Table: Percentage of times that different criterion selects the correct true model (C), overfit (O), and underfit (U).

	Non-Correlated			Correlated		
Criterion	C	O	U	C	O	U
BIC_{obs}	**92**	3	5	**94**	2	4
BIC_{orig}	96	2	2	93	0	7
BIC_{cc}	79	1	20	91	0	9

Extract of codes:

```
for (j in 1:(nrow(subsets)-1)){
  variables = subsets[j,]
  pos_var=which(variables==1)
  nb.x = sum(variables)
  nb.para = (nb.x + 1) + p + p*p
  list.saem.subset=miss.saem(X.obs,y,pos_var,ll_obs_cal=TRUE)
  BIC[nb,j] = -2*list.saem.subset$ll+ nb.para * log(n)
}
```

4 Application on Traumabase

Exploration of dataset
Cross validation

Predictive performance

Risk of severe hemorrhage for Traumabase

4 Exploration of dataset

Data preprocessing ⇒ **6384 patients**.
Clinical experience ⇒ **14 influential quantitative measurements**
Missingness: 0 - **60%**.

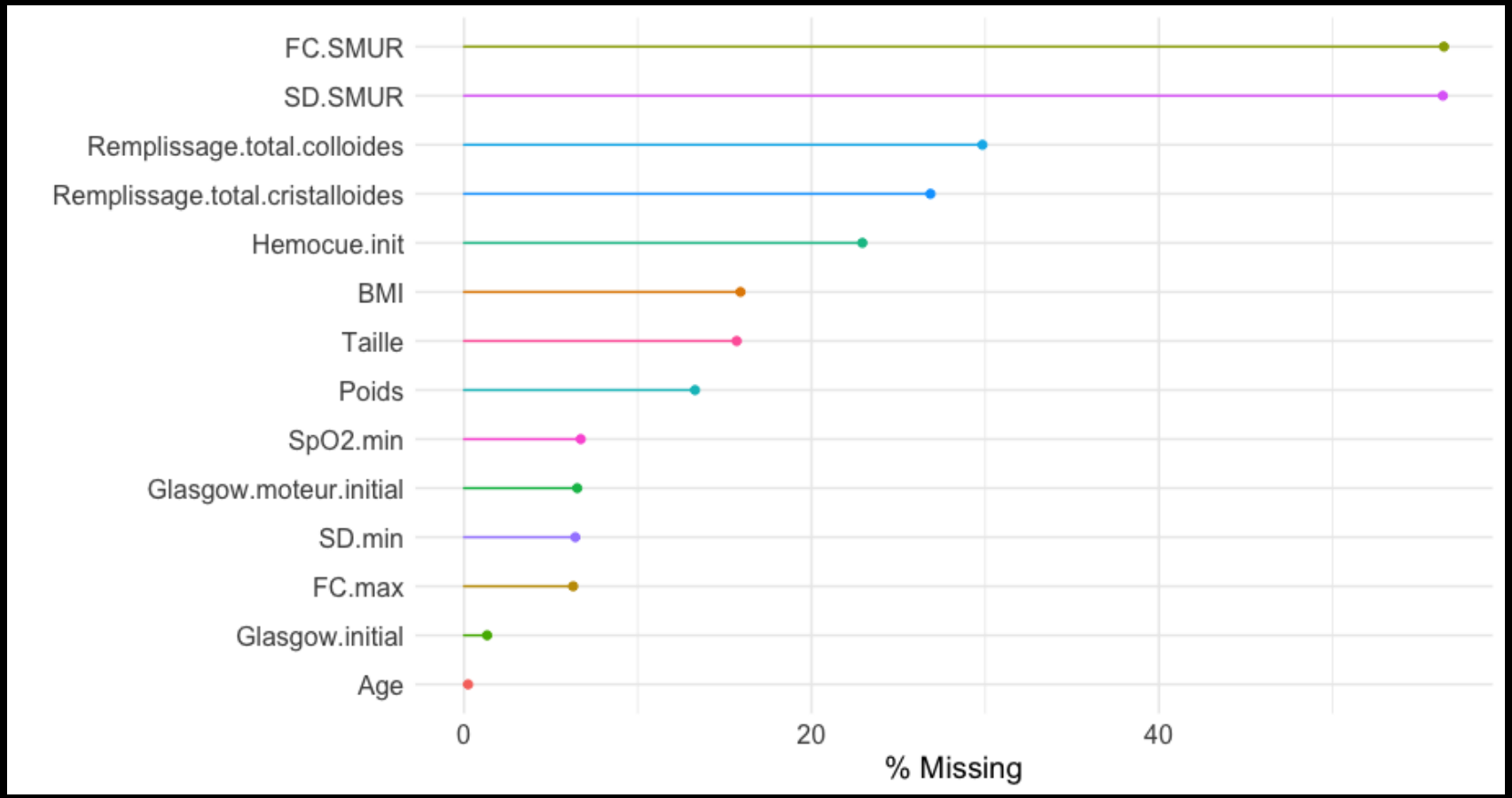
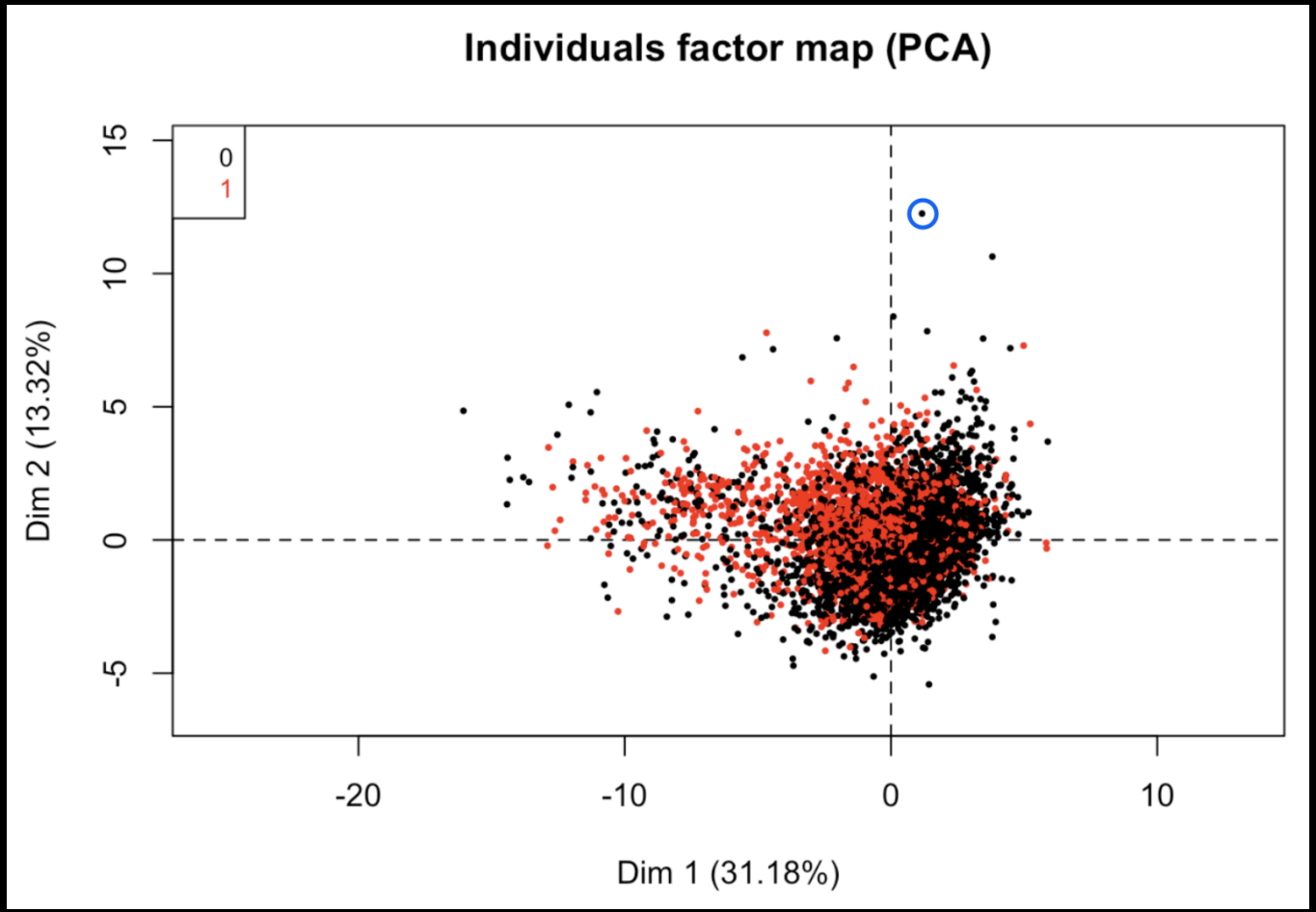


Figure: Percentage of missingness in each variable.

4 Exploration of dataset

Two observations resulting in a very small value of **observed log-likelihood**:
For the 3302 nd patient, the calculation of BMI is wrong.
For 1144 th patient, the values of Weight (200 kg) and Height (100 cm) have a large possibility to be wrong.



Random split : training set (80%) + test set (20%)

Variables	Estimate (se)
(*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)

- The more one bleed, the lower the HemoCu is, and the more the blood will be transfused. Then the more likely one will end up in a hemorrhagic shock.

4 Predictive performance : comparison

Random split : training set (80%) + test set (20%) (repeated 15 times)

Table: Comparison of the median of the predictive performances (values are multiplied by 100) of different methods dealing with missing data.

Metrics	SAEM	impPCA	impMean	*mice*
AUC	**86.70**	86.67	86.62	86.62
Accuracy	**83.23**	81.96	82.74	**83.54**
Sensitivity	**78.26**	77.59	76.86	76.29
Specificity	83.70	82.21	83.15	**84.58**
Precision	30.56	30.68	30.97	**32.88**

5 Conclusion

- SAEM for logistic regression with missingness leads to unbiased estimation and a more reasonable coverage of confidence interval;
- Model selection by criterion BIC with missing data can be well performed;
- R package *misaem*: `github.com/wjiang94/misaem`, `logReg`
- arXiv:1805.04602

1. Deal with both quantitative and categorical data;
2. Deal with both MAR and MNAR missing values.
3. Causal inference and propensity score analysis.