Comparison of multiple imputation methods for systematically and sporadically missing multilevel data

V. Audigier, I. White , S. Jolani, T. Debray, M. Quartagno, S. van Buuren, M. Resche-Rigon

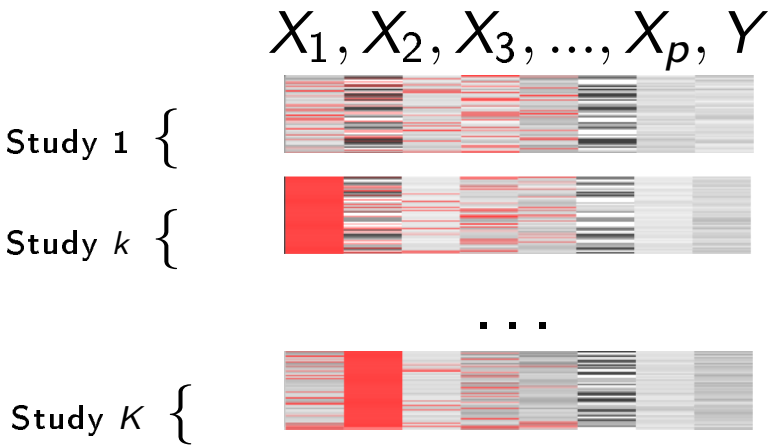
INSERM, UMR 1153, Equipe ECSTRA, Hôpital Saint-Louis, Paris

Journées 2016 du GDR, June 27, Lyon

Motivation

A regression framework

- small sample size
- bias due to the collecting process
- missing values (sporadically missing)



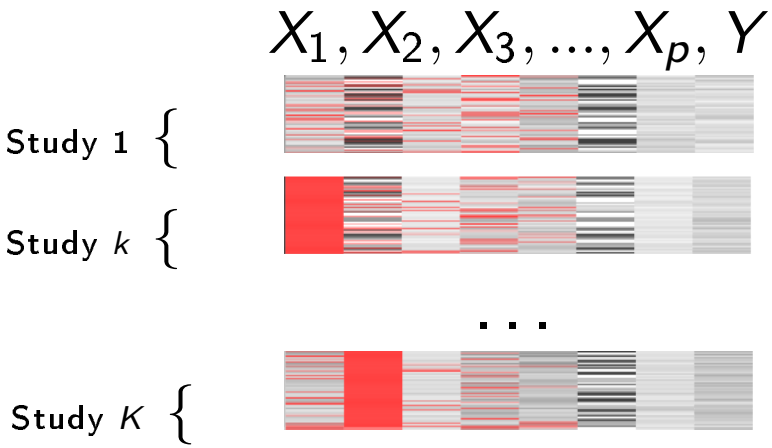
$\Rightarrow$ Aggregate several studies (IPD meta-analysis)

- multilevel structure
  $\mathbf{y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \quad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_k \sim \mathcal{N}(0, \sigma^2)$
- sporadically and systematically missing data

Motivation

A regression framework

- small sample size
- bias due to the collecting process
- missing values (sporadically missing)



$\Rightarrow$ Aggregate several studies (IPD meta-analysis)

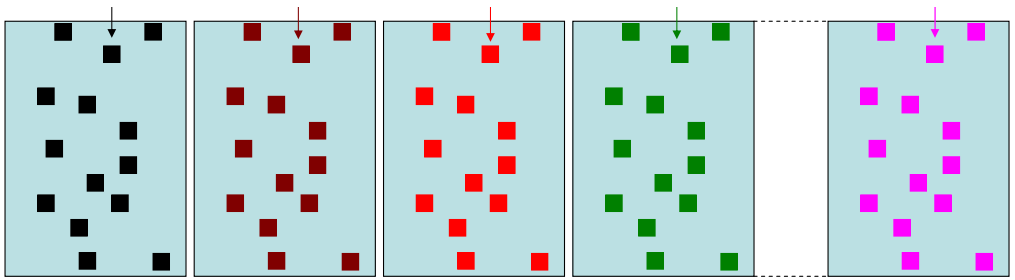
- multilevel structure
  $\mathbf{y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \quad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma^2\right)$
- sporadically and systematically missing data

$\hat{\beta}$ and associated variability T $\rightarrow$ a complex inference task

Multiple imputation (Rubin, 1987)

1. Provide a set of M parameters θ_m of an imputation model to generate M plausible imputed data sets

$$P\left(X^{miss}|X^{obs},\theta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(X^{miss}|X^{obs},\theta_M\right)$$



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

3. Combine the results: $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$

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

$\Rightarrow$ Aim: provide estimation of the parameters and of their variability

MI for mixed-effects model

The mixed-effects model as imputation model

$$\begin{aligned} \mathbf{Y}_k &= \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \\ b_k &\sim \mathcal{N}(0, \Psi) \\ \varepsilon_k &\sim \mathcal{N}(0, \Sigma_k) \end{aligned}$$

Multiple imputation under this model

1. generating M sets of parameters $\theta_m = (\beta^m, \Psi^m, \Sigma_k^m)$
2. imputing the data according each set θ_m
   - draw $b_k^m | \mathbf{Y}_k^{obs}, \theta_m$
   - draw $\mathbf{Y}_k^{miss} | \theta_m, b_k^m$

Specific issues

- systematically missing data: $\mathbf{Y}_k$ is not observed $\rightarrow \Sigma_k$?
- sporadically missing data: $P\left(b_k | \mathbf{Y}_k^{obs}, \theta_m\right)$ can be untractable

Quartagno and Carpenter (2015)

$$\mathbf{Y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \qquad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_k \sim \mathcal{N}(0, \Sigma_k)$$

1. Bayesian formulation to generate $\theta_m = (\beta^m, \Psi^m, \Sigma^m)_{1 \leq m \leq M}$
   - prior: $\Sigma_k^{-1} \sim W(\nu_1, \Lambda_1), \quad \Psi^{-1} \sim W(\nu_2, \Lambda_2), \quad \beta \propto 1$
   - draw in the posterior (Gibbs sampler)

$$b_k^{(\ell+1)} \sim P\left(b_k | \mathbf{Y}, \beta^{(\ell)}, \Psi^{(\ell)}, \Sigma_k^{(\ell)}\right) (\textit{Gaussian})$$
$$\Psi^{-1(\ell+1)} \sim P\left(\Psi^{-1} | \mathbf{Y}, \beta^{(\ell)}, \Sigma^{(\ell)}, b^{(\ell+1)}\right) (\textit{Wishart})$$
$$\Sigma_k^{-1(\ell+1)} \sim P\left(\Sigma_k^{-1} | \mathbf{Y}, \beta^{(\ell)}, \Psi^{(\ell+1)}, b_k^{(\ell+1)}\right) (\textit{Wishart})$$
$$\beta^{(\ell+1)} \sim P\left(\beta | \mathbf{Y}, \Psi^{(\ell+1)}, \Sigma^{(\ell+1)}, b^{(\ell+1)}\right) (\textit{Gaussian})$$

2. Imputation

Quartagno and Carpenter (2015)

$$\mathbf{Y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \qquad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_k \sim \mathcal{N}(0, \Sigma_k)$$

1. Bayesian formulation to generate $\theta_m = (\beta^m, \Psi^m, \Sigma^m)_{1 \leq m \leq M}$
   - prior: $\Sigma_k^{-1} \sim W(\nu_1, \Lambda_1), \quad \Psi^{-1} \sim W(\nu_2, \Lambda_2), \quad \beta \propto 1$
   - draw in the posterior (Gibbs sampler)

$$\mathbf{Y}^{miss\,(\ell)} \sim P\left(\mathbf{Y}^{miss} | \mathbf{Y}^{obs}, \beta^{(\ell-1)}, \Psi^{(\ell-1)}, \Sigma^{(\ell-1)}, b_k^{(\ell-1)}\right)$$

$$b_k^{(\ell+1)} \sim P\left(b_k | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Psi^{(\ell)}, \Sigma_k^{(\ell)}\right) (Gaussian)$$

$$\Psi^{-1\,(\ell+1)} \sim P\left(\Psi^{-1} | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Sigma^{(\ell)}, b^{(\ell+1)}\right) (Wishart)$$

$$\Sigma_k^{-1\,(\ell+1)} \sim P\left(\Sigma_k^{-1} | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Psi^{(\ell+1)}, b_k^{(\ell+1)}\right) (Wishart)$$

$$\beta^{(\ell+1)} \sim P\left(\beta | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \Psi^{(\ell+1)}, \Sigma^{(\ell+1)}, b^{(\ell+1)}\right) (Gaussian)$$

2. Imputation

Quartagno and Carpenter (2015)

$$\mathbf{Y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \qquad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_k \sim \mathcal{N}(0, \Sigma_k)$$

1. Bayesian formulation to generate $\theta_m = (\beta^m, \Psi^m, \Sigma^m)_{1 \leq m \leq M}$
   - prior: $\Sigma_k^{-1} \sim W(\nu_1, \Lambda_1), \quad \Psi^{-1} \sim W(\nu_2, \Lambda_2), \quad \beta \propto 1$
   - draw in the posterior (Gibbs sampler)

$$\mathbf{Y}^{miss\,(\ell)} \sim P\left(\mathbf{Y}^{miss} | \mathbf{Y}^{obs}, \beta^{(\ell-1)}, \Psi^{(\ell-1)}, \Sigma^{(\ell-1)}, b_k^{(\ell-1)}\right)$$

$$b_k^{(\ell+1)} \sim P\left(b_k | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Psi^{(\ell)}, \Sigma_k^{(\ell)}\right) (Gaussian)$$

$$\Psi^{-1\,(\ell+1)} \sim P\left(\Psi^{-1} | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Sigma^{(\ell)}, b^{(\ell+1)}\right) (Wishart)$$

$$\Sigma_k^{-1\,(\ell+1)} \sim P\left(\Sigma_k^{-1} | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \beta^{(\ell)}, \Psi^{(\ell+1)}, b_k^{(\ell+1)}\right) (Wishart)$$

$$\beta^{(\ell+1)} \sim P\left(\beta | \mathbf{Y}^{obs}, \mathbf{Y}^{miss\,(\ell)}, \Psi^{(\ell+1)}, \Sigma^{(\ell+1)}, b^{(\ell+1)}\right) (Gaussian)$$

2. Imputation

$$\mathbf{Y}_k^{miss} \sim P\left(\mathbf{Y}^{miss} | \mathbf{Y}_k^{obs}, \theta^m, b_k\right)$$

Jolani et al. (2015)

A fully conditional specification approach

$$\mathbf{y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \varepsilon_k \qquad logit\left(P\left(\mathbf{y}_k = 1\right)\right) = \mathbf{X}_k\beta + \mathbf{Z}_k b_k$$
$$b_k \sim \mathcal{N}\left(0, \Psi\right) \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma^2\right)$$

For each incomplete variable

1. generate $\theta_m = \left(\beta^m, \Psi^m, \sigma^{2\,m}\right)$
   - Non-informative prior for θ (Jeffrey)
   - Posterior: large sample approximation

$$\beta|\mathbf{X}, \mathbf{Z}, \mathbf{y}^{obs}, b, \Psi, \sigma^2 \sim \mathcal{N}\left(\widehat{\beta}^{ML}, \mathcal{I}^{-1}\left(\widehat{\beta}^{ML}\right)\right) \qquad \Psi|\mathbf{X}, \mathbf{Z}, \mathbf{y}^{obs}, b, \beta, \sigma^2 \sim \mathcal{W}\left(\widehat{\Psi}^{ML}, p\right)$$
$$b_k|\mathbf{X}, \mathbf{Z}, \mathbf{y}^{obs}, \Psi, \beta, \sigma^2 \sim \mathcal{N}\left(0, \Psi\right) \qquad \sigma^2|\mathbf{X}, \mathbf{Z}, \mathbf{y}^{obs}, b, \Psi, \beta \sim \widehat{\sigma^2}\left(I - p\right)/\chi^2\left(1\right)$$

2. imputation

$$\mathbf{y}_k^{miss} \sim P\left(\mathbf{y}^{miss}|\mathbf{X}, \mathbf{Z}, \beta, \Psi, \sigma^2, b_k\right)$$

Resche-Rigon and White (2016)

A fully conditional specification approach

$$\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \qquad logit\left(P\left(\mathbf{y}_k = 1\right)\right) = \mathbf{X}_k\beta_k$$
$$\beta_k = \beta + b_k \qquad b_k \sim \mathcal{N}\left(0, \Psi_b\right) \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

1.
   - generate ${\sigma_k^2}^m$
   - generate $\left(\beta^m, \Psi^m | {\sigma_k^2}^m\right)$
     - step 1: on each cluster $\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \; \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right) \rightarrow \widehat{\beta}_k$, $\widehat{var}\left(\widehat{\beta}_k | b_k, \sigma_k\right)$
     - step 2: $\widehat{\beta_k} = \beta + \varepsilon_\beta \;\; \varepsilon_\beta \sim \mathcal{N}\left(0, \widehat{var}\left(\widehat{\beta}_k | b_k, \sigma_k\right) + \Psi_b\right) \rightarrow \widehat{\beta}, \widehat{\Psi}_b$
     - $\beta_k \sim \mathcal{N}\left(\widehat{\beta}, \widehat{var}\left(\widehat{\beta}_k | b_k, \sigma_k\right) + \widehat{\Psi}_b\right)$
2. imputation

Resche-Rigon and White (2016)

A fully conditional specification approach

$$\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \qquad logit\left(P\left(\mathbf{y}_k = 1\right)\right) = \mathbf{X}_k\beta_k$$
$$\beta_k = \beta + b_k \qquad b_k \sim \mathcal{N}\left(0, \Psi_b\right) \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

1.
   - generate $\sigma_k^{2\,m}$ (prior: $log\sigma_k = log\sigma + c_k \quad c_k \sim \mathcal{N}\left(0, \Psi_c\right)$)
     - step 1: $\log\widehat{\sigma}_k$, $\widehat{var}\left(\log\widehat{\sigma}_k|c_k\right)$
     - step 2: $\widehat{log\sigma_k} = \log\sigma + \varepsilon_\sigma \quad \varepsilon_\sigma \sim \mathcal{N}\left(0, \widehat{var}\left(log\widehat{\sigma}_k|c_k\right) + \Psi_c\right)$ $\rightarrow \log\widehat{\sigma}, \widehat{\Psi}_c$
     - $\log\sigma_k \sim \mathcal{N}\left(\log\widehat{\sigma}, \widehat{var}\left(\log\widehat{\sigma}_k|c_k\right) + \widehat{\Psi}_c\right)$
   - generate $\left(\beta^m, \Psi^m|\sigma_k^{2\,m}\right)$
     - step 1: on each cluster $\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \; \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right) \rightarrow \widehat{\beta}_k$, $\widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right)$
     - step 2: $\widehat{\beta_k} = \beta + \varepsilon_\beta \quad \varepsilon_\beta \sim \mathcal{N}\left(0, \widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right) + \Psi_b\right) \rightarrow \widehat{\beta}, \widehat{\Psi}_b$
     - $\beta_k \sim \mathcal{N}\left(\widehat{\beta}, \widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right) + \widehat{\Psi}_b\right)$
2. imputation

Resche-Rigon and White (2016)

A fully conditional specification approach

$$\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \qquad logit\left(P\left(\mathbf{y}_k = 1\right)\right) = \mathbf{X}_k\beta_k$$
$$\beta_k = \beta + b_k \qquad b_k \sim \mathcal{N}\left(0, \Psi_b\right) \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

1.
   - generate ${\sigma_k^2}^m$ (prior: $log\sigma_k = log\sigma + c_k \quad c_k \sim \mathcal{N}\left(0, \Psi_c\right)$)
     - step 1: $\log\widehat{\sigma}_k$, $\widehat{var}\left(\log\widehat{\sigma}_k|c_k\right)$
     - step 2: $\widehat{log\sigma_k} = \log\sigma + \varepsilon_\sigma \quad \varepsilon_\sigma \sim \mathcal{N}\left(0, \widehat{var}\left(log\widehat{\sigma}_k|c_k\right) + \Psi_c\right)$ $\rightarrow \log\widehat{\sigma}, \widehat{\Psi}_c$
     - $\log\sigma_k \sim \mathcal{N}\left(\log\widehat{\sigma}, \widehat{var}\left(\log\widehat{\sigma}_k|c_k\right) + \widehat{\Psi}_c\right)$
   - generate $\left(\beta^m, \Psi^m|{\sigma_k^2}^m\right)$
     - step 1: on each cluster $\mathbf{y}_k = \mathbf{X}_k\beta_k + \varepsilon_k \; \varepsilon_k \sim \mathcal{N}\left(0, \sigma_k^2\right) \rightarrow \widehat{\beta}_k$, $\widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right)$
     - step 2: $\widehat{\beta_k} = \beta + \varepsilon_\beta \quad \varepsilon_\beta \sim \mathcal{N}\left(0, \widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right) + \Psi_b\right) \rightarrow \widehat{\beta}, \widehat{\Psi}_b$
     - $\beta_k \sim \mathcal{N}\left(\widehat{\beta}, \widehat{var}\left(\widehat{\beta}_k|b_k, \sigma_k\right) + \widehat{\Psi}_b\right)$
2. imputation $\quad {\mathbf{y}_k}^{miss} \sim P\left(\mathbf{y}^{miss}|\mathbf{X}, \beta_k, \Psi_b, \sigma_k^2\right)$

Simulation design

500 incomplete data sets ($I = 15000, p = 3, K = 30$)

- 2 continuous variables $\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(3)}$ from $\mathcal{N}(b_k, \mathbb{I}_2)$
- 1 binary variable $\mathbf{x}_k^{(2)}$: $logit\left(P\left(\mathbf{x}_k^{(2)} = 1\right)\right) = b_k$
- $\mathbf{y}_k = \beta^1 \mathbf{x}_k^{(1)} + \beta^2 \mathbf{x}_k^{(2)} + b_k^0 + b_k^1 \mathbf{x}_k^{(1)} + b_k^2 \mathbf{x}_k^{(2)} + \varepsilon_k$
  with $\beta = (.5, 1)$, $\Psi = \begin{bmatrix} 0.250 & 0.075 & 0.075 \\ 0.075 & 0.250 & 0.075 \\ 0.075 & 0.075 & 0.250 \end{bmatrix}$, $\sigma^2 = 1$
- add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ $\pi_{syst} = .2$, $\pi_{spor} = .2$ (MCAR)

Apply 3 MI methods using $M = 5$ imputed arrays

Criteria for β: bias, rmse, variance estimate, coverage

Results: random intercept ($b^1 = 0$, $b^2 = 0$)

Method	β	True	Bias	$\widehat{T}$	95% Cover	RMSE
Full	coef_X1	0.500	0.001	0.000	0.956	0.017
	coef_X2	1.000	-0.000	0.001	0.948	0.039
CC	coef_X1	0.500	0.002	0.001	0.948	0.028
	coef_X2	1.000	0.002	0.004	0.938	0.062
FCS-2steps	coef_X1	0.500	-0.006	0.001	0.962	0.023
	coef_X2	1.000	-0.032	0.003	0.926	0.060
FCS-glm	coef_X1	0.500	-0.001	0.001	0.968	0.022
	coef_X2	1.000	-0.023	0.003	0.942	0.055
JM	coef_X1	0.500	-0.017	0.001	0.962	0.031
	coef_X2	1.000	-0.016	0.004	0.968	0.055

Table: Inference results for MI methods

FCS-2step	FCS-glm	JM
0.39	11.62	0.06

Table: Average time required to multiply impute one data set (in min)

Results: random effects

Method	β	True	Bias	$\widehat{T}$	95% Cover	RMSE
Full	coef_X1	0.500	0.000	0.009	0.940	0.092
	coef_X2	1.000	0.001	0.010	0.940	0.101
CC	coef_X1	0.500	-0.002	0.014	0.938	0.121
	coef_X2	1.000	0.005	0.017	0.938	0.128
FCS-2steps	coef_X1	0.500	-0.029	0.009	0.900	0.110
	coef_X2	1.000	-0.057	0.012	0.916	0.124
FCS-glm	coef_X1	0.500	-0.038	0.008	0.870	0.110
	coef_X2	1.000	-0.041	0.010	0.908	0.116
JM	coef_X1	0.500	-0.057	0.008	0.888	0.114
	coef_X2	1.000	-0.071	0.013	0.928	0.130

Table: Inference results for MI methods

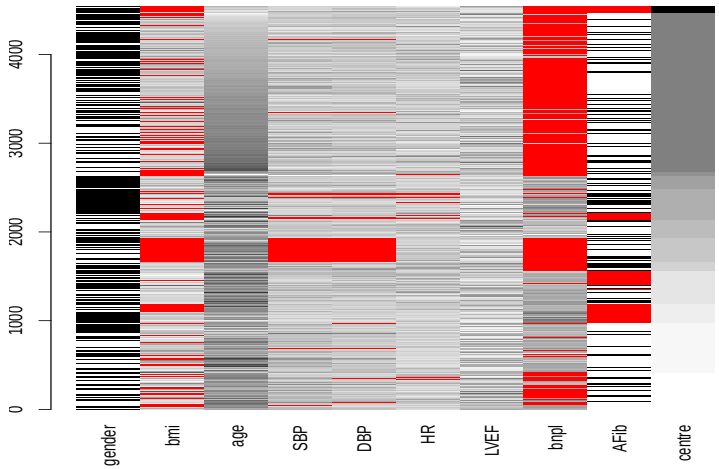
FCS-2step	FCS-glm	JM
0.59	10.92	0.06

Table: Average time required to multiply impute one data set (in min)

Appplication to GREAT data

Risk factors associated with acute heart failure

- y=left ventricular ejection fraction, $\mathbf{X}$ = BNP, AFib,...
- $I = 4546$, $p = 9$, $K = 12$
- MI using $M = 20$ imputed arrays



	CC		FCS-2steps		FCS-glm		JM	
	$\widehat{\beta}$	$\sqrt{\widehat{T}}$	$\widehat{\beta}$	$\sqrt{\widehat{T}}$	$\widehat{\beta}$	$\sqrt{\widehat{T}}$	$\widehat{\beta}$	$\sqrt{\widehat{T}}$
(Intercept)	53.060	5.640	56.990	5.710	60.160	3.680	61.310	4.650
gender	-6.910	0.870	-5.970	0.580	-5.560	0.460	-6.130	0.600
bmi	0.170	0.070	0.110	0.070	0.110	0.050	0.140	0.060
age	0.170	0.030	0.120	0.020	0.130	0.020	0.110	0.020
SBP	0.180	0.020	0.150	0.020	0.150	0.010	0.150	0.010
DBP	-0.190	0.030	-0.170	0.030	-0.170	0.020	-0.180	0.020
HR	-0.060	0.020	-0.050	0.010	-0.050	0.010	-0.050	0.010
bnpl	-10.200	0.900	-9.290	1.510	-10.420	0.650	-8.860	0.940
AFib1	3.280	0.840	3.130	0.660	2.810	0.540	3.050	0.570

FCS-2step	FCS-glm	JM
5.90	495.4	2.10

Conclusion

An overview of MI methods with multilevel structure

- inference performances quite similar
- JM and FCS-2step quick to perform
- limits with complex analysis model

Perspectives

- larger simulation study (missing data mechanism)
- strategies for complex analysis model
- R package

References I

D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.

M. Quartagno and J.R. ; Carpenter. Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Stat Med*, 2015. doi: 10.1002/sim.6837.

S. Jolani, T. P. A. Debray, H. Koffijberg, S. van Buuren, and K. G. M. Moons. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine*, 34(11):1841–1863, 2015.

M. Resche-Rigon and I. White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *smmr*, 2016. in revision.