micemd: a smart multiple imputation R package for missing multilevel data

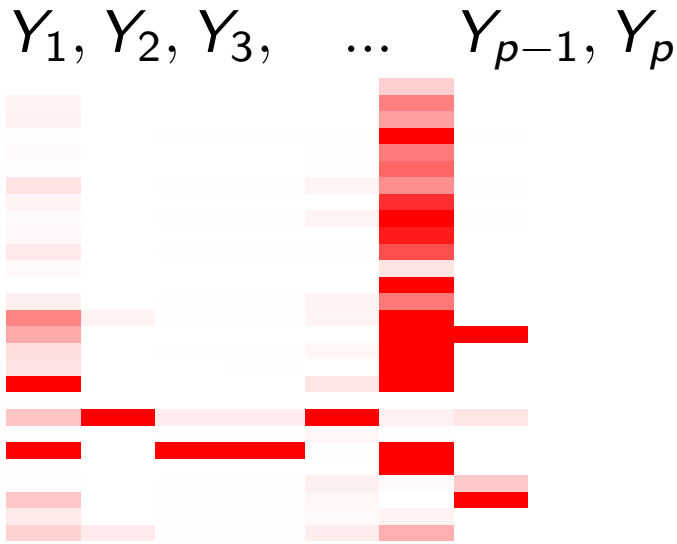
V. Audigier, M. Resche-Rigon

CEDRIC, MSDMA team, CNAM, Paris

UseR, 2019 July 11th, Toulouse

Motivation: GREAT data (GREAT Network, 2013)

- Risk factors associated with short-term mortality in acute heart failure
- 28 observational cohorts, 11685 patients, 2 **binary** and 8 **continuous** variables (patient characteristics and potential risk factors)
- sporadically and systematically **missing data**



Aim: explain the relationship between biomarkers (BNP, AFIB,...) and the left ventricular ejection fraction (LVEF)

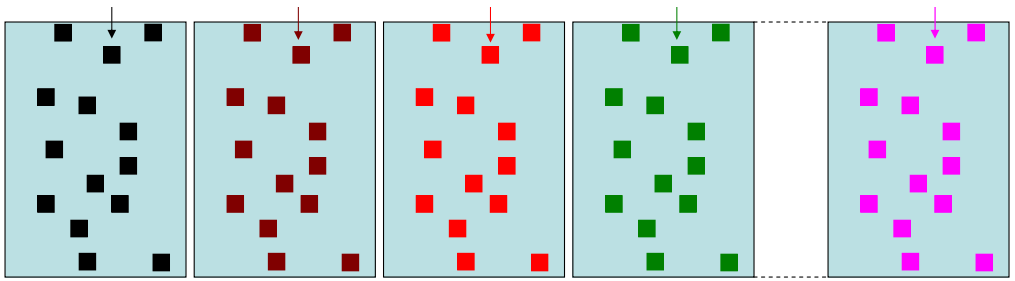
$$y_{ik}^{LVEF} = \beta^0 + \beta^1 \mathbf{y}_{ik}^{BNP} + \beta^2 \mathbf{y}_{ik}^{AFIB} + b_k^0 + b_k^1 \mathbf{y}_{ik}^{BNP} + \varepsilon_{ik}$$
$$b_k \sim \mathcal{N}(0, \Psi) \, \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

$\hat{\beta}$ **and associated variability** $var\left(\widehat{\beta}\right)$

Multiple imputation (Rubin, 1987)

1. Generate a set of M parameters $(\theta_m)_{1 \leq m \leq M}$ of an imputation model to generate M plausible imputed data sets

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



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

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

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

$\Rightarrow$ **Provide estimation of the parameters and of their variability**

Imputation model for multilevel data

Two standard ways to perform MI

- **Fully conditional specification** (FCS, MICE): a conditional imputation model for each variable
- **Joint modelling** (JM): a joint imputation model for all variables

The imputation model (joint or conditional) needs to

- account for the **heterogeneity** between clusters
- account for the **types** of variables (continuous and binary)
- be identifiable with **sporadically and systematically** missing values

MI for multilevel data

Method (type - name)	deal with missing values:				Coded in R
	Spor.?	Syst.?	continuous ?	binary?	
JM-pan	yes	yes	yes	no	yes
JM-REALCOM	yes	yes	yes	yes	no
JM-jomo	yes	yes	yes	yes	yes
JM-Mplus	yes	yes	yes	yes	no
JM-RCME	yes	yes	yes	no	no
FCS-pan	yes	yes	yes	no	yes
FCS-2lnorm	yes	no	yes	no	yes
FCS-GLM	yes	yes	yes	yes	yes
FCS-2stage	yes	yes	yes	yes	yes

FCS-2stage (Resche-Rigon and White, 2016)

Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\left(\beta + b_k\right) + \varepsilon_{ik} \qquad b_k \sim \mathcal{N}\left(0, \Psi\right) \qquad \varepsilon_{ik} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$
   - estimate θ and var $\left(\widehat{\theta}\right)$ with **a two-stage estimator**

     stage 1 fit $y_{ik} = \mathbf{z}_{ik}\beta_k + \varepsilon_{ik}$ to each cluster

     stage 2 combine the $\widehat{\beta}_k$ (and $\widehat{\sigma}_k^2$) with a random intercept model:

     $$\widehat{\beta}_k = \left(\beta + b_k\right) + \varepsilon'_k \quad b_k \sim \mathcal{N}\left(0, \Psi\right) \; \varepsilon'_k \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

   - draw θ_m from the **asymptotic** distribution of the estimator with expectation $\widehat{\theta}$ and variance $\widehat{\text{var}}\left(\widehat{\theta}\right)$
2. impute in each cluster k with **systematically missing data**
   - draw b_k **from their marginal distribution**
   - impute data according to the imputation model

FCS-2stage (Resche-Rigon and White, 2016)

Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\left(\beta + b_k\right) + \varepsilon_{ik} \qquad b_k \sim \mathcal{N}\left(0, \Psi\right) \qquad \varepsilon_{ik} \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{k}}\right)$$

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma^2_1, \ldots, \sigma^2_K\right)_m\right)$
   - estimate θ and var $\left(\widehat{\theta}\right)$ with **a two-stage estimator**

     stage 1 fit $y_{ik} = \mathbf{z}_{ik}\beta_k + \varepsilon_{ik}$ to each cluster

     stage 2 combine the $\widehat{\beta}_k$ (and $\widehat{\sigma}^2_k$) with a random intercept model:

     $$\widehat{\beta}_k = \left(\beta + b_k\right) + \varepsilon'_k \quad b_k \sim \mathcal{N}\left(0, \Psi\right) \; \varepsilon'_k \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

   - draw θ_m from the **asymptotic** distribution of the estimator with expectation $\widehat{\theta}$ and variance $\widehat{\text{var}}\left(\widehat{\theta}\right)$
2. impute in each cluster k with **sporadically missing data**
   - draw b_k **conditionally to** $\widehat{\beta_k}$ from stage 2
   - impute data according to the imputation model

Differences between MI methods (Audigier et al., 2018)

	Prior	heteroscedasticity assumption	link binary
FCS-2stage		yes	logit
FCS-GLM	Jeffrey	no	logit
JM-jomo	conjugate	yes	probit

- conjugate prior distributions are known to very informative in GLMM
- heteroscedastic assumption is more flexible

Simulation design: data generation

500 incomplete data sets are independently simulated

- a multilevel structure
- 4 variables (1 binary, 3 continuous)
- sporadically and systematically missing data
- parameters are tuned to mimic GREAT data

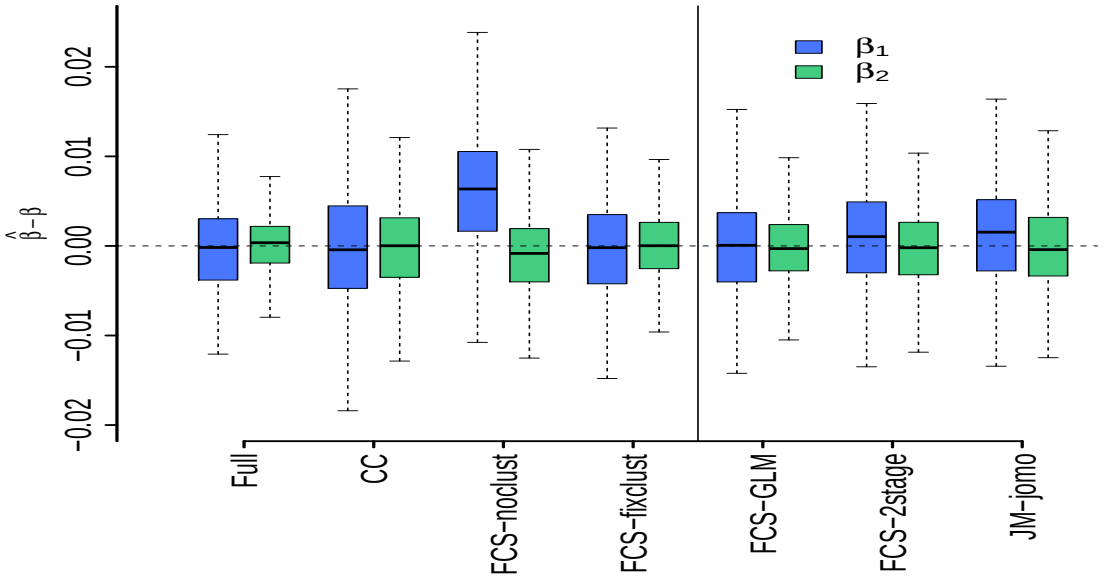
More precisely,

- $\mathbf{x}_{ik}^{(1)} : \mathcal{N}(2.9 + \mu_k, .36)$
- $\mathbf{x}_{ik}^{(2)} : logit\left(P\left(\mathbf{x}_{ik}^{(2)} = 1\right)\right) = 4.2 + \nu_k \qquad (\mu_k, \nu_k, \xi_k) \sim \mathcal{N}\left(0, \begin{bmatrix} .12 & .001 & .001 \\ .001 & .12 & .001 \\ .001 & .001 & .12 \end{bmatrix}\right)$
- $\mathbf{x}_{ik}^{(3)} : \mathcal{N}(2.9 + \xi_k, .36)$
- $y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$
  with $\beta = (.72, -.11, .03)$, $\Psi = \begin{bmatrix} .0077 & .0015 \\ .0015 & .0004 \end{bmatrix}$, $\sigma = .15$
- add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ with $\pi_{syst} = .25$ and $\pi_{spor} = .25$

Simulation design

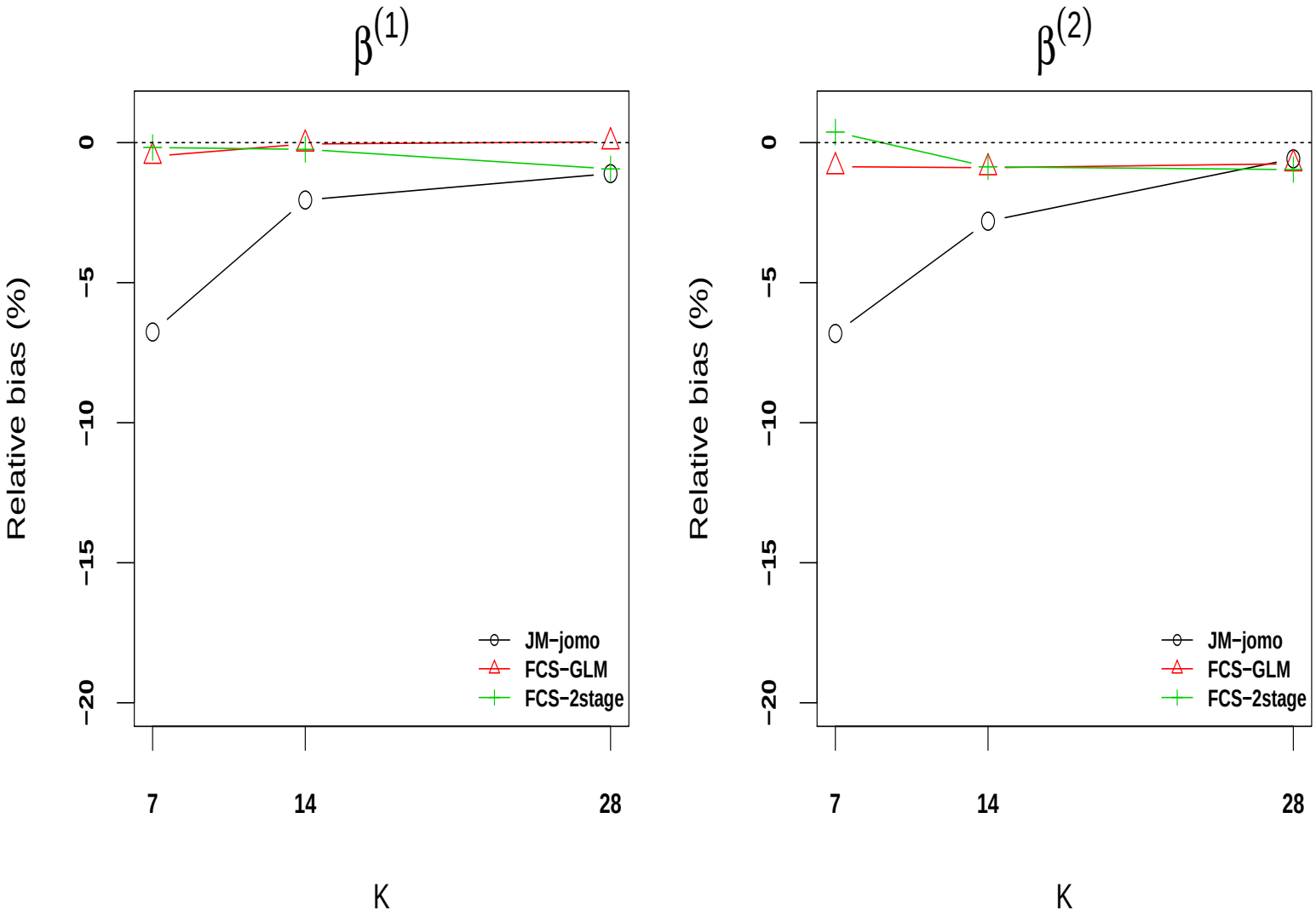
- **Methods**
  - JM-jomo, FCS-GLM, FCS-2stage
  - Full, CC, FCS-fixclust, FCS-noclust
  - $M = 5$ imputed arrays
- **Estimands**: β and $var\left(\widehat{\beta}\right)$
- **Criteria**: bias, rmse, variance estimate, coverage

Results: base-case

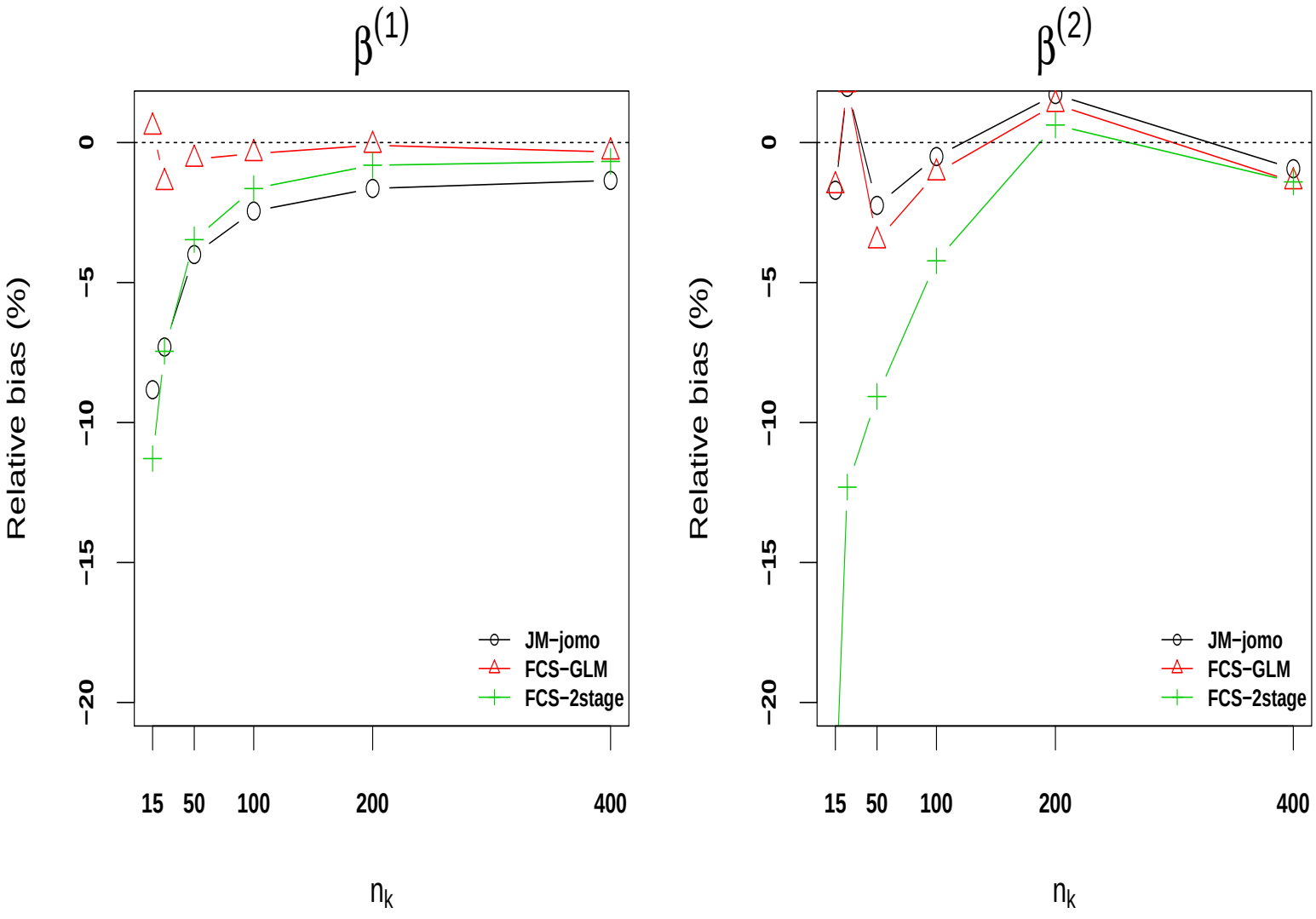


Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0070	0.0053	0.0071	0.0053	92.2	94.4	
FCS-noclust	0.0041	0.0043	0.0067	0.0045	58.2	92.0	0.9
FCS-fixclust	0.0043	0.0043	0.0058	0.0042	87.0	94.6	1.1
FCS-GLM	0.0047	0.0046	0.0057	0.0043	89.7	95.8	103.3
FCS-2stage	0.0059	0.0049	0.0058	0.0044	95.0	96.2	0.9
JM-jomo	0.0066	0.0069	0.0056	0.0049	98.4	97.6	7.8

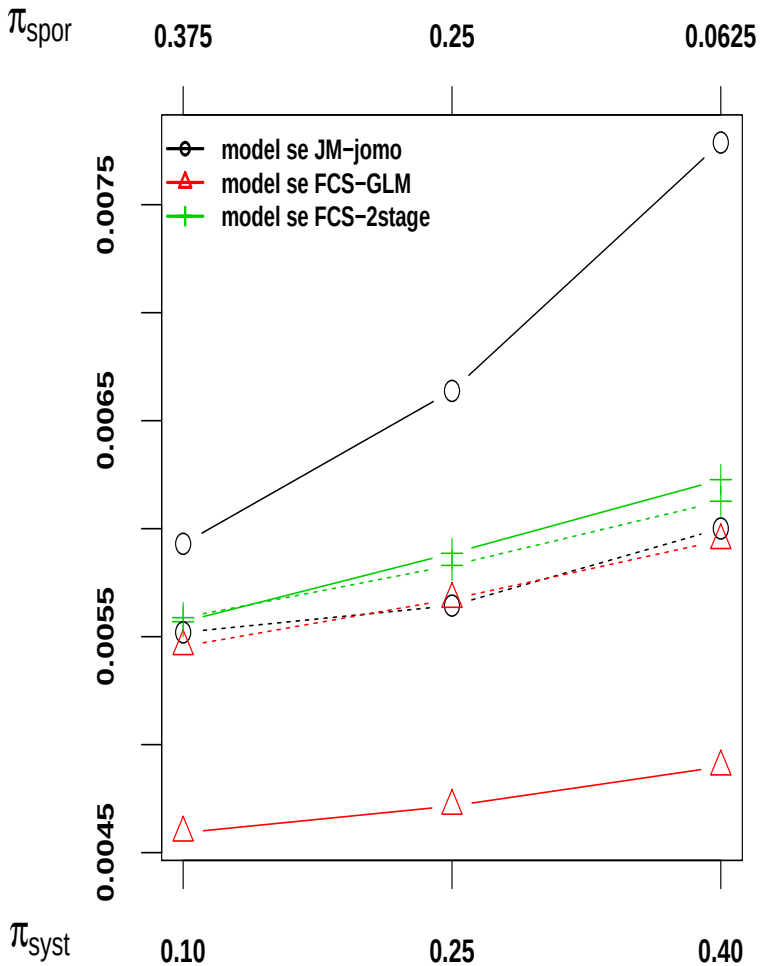
Influence of the number of clusters

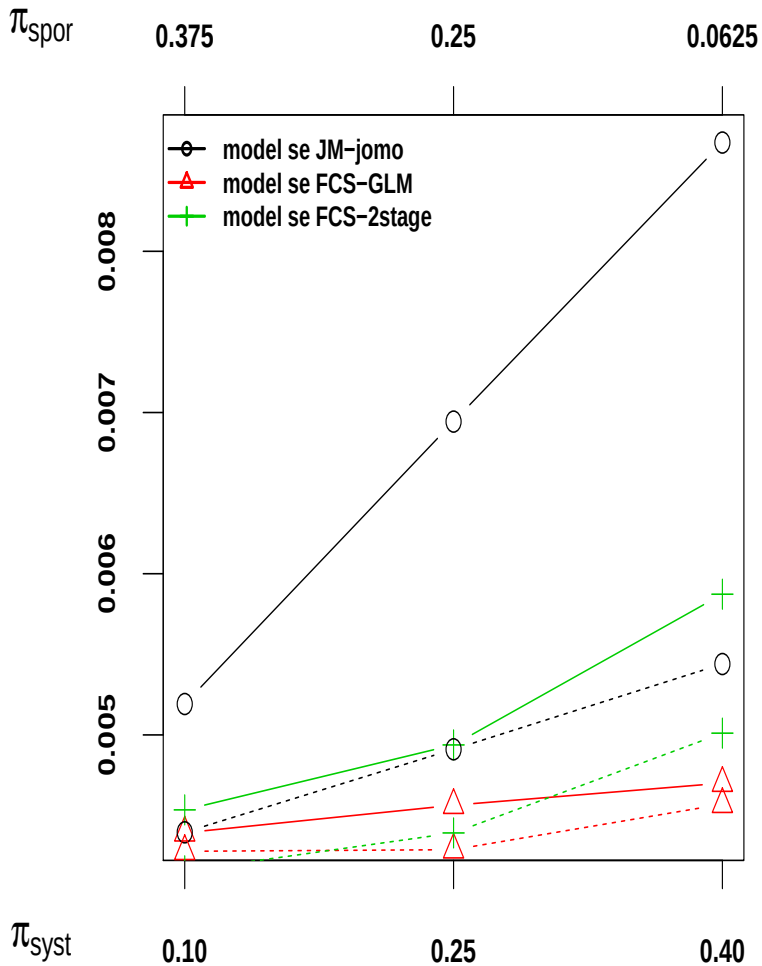


Influence of the cluster size



Influence of the proportion of systematically missing values





Conclusion

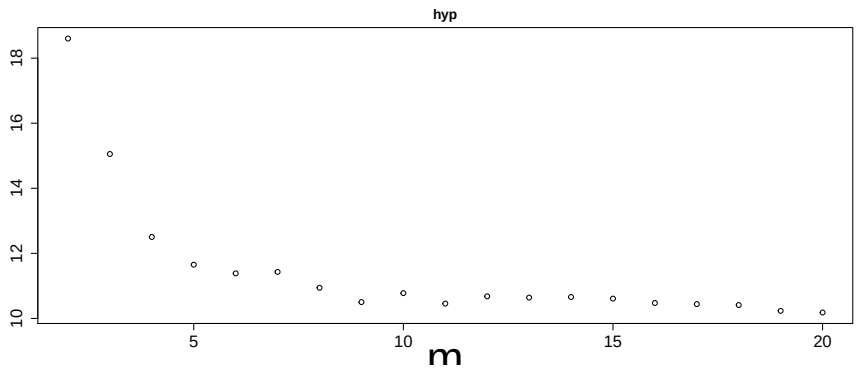
micemd is an add-on for the mice R package [1] which performs multiple imputation by chained equations.

imputation models for

- multilevel data
- sporadically and systematically missing values
- continuous, binary or count variables

tools to facilitate its use

- automatic choice of imputation models
- choice of the number of imputed tables
- overimputation for model checking
- parallel calculation



[1] van Buuren and Groothuis-Oudshoorn (2011)

References I

Global Research on Acute conditions Team (GREAT) Network. Managing Acute Heart Failure in the ED - Case Studies from the Acute Heart Failure Academy, 2013. http://www.greatnetwork.org.

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

Matthieu Resche-Rigon and Ian White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 2016. http://dx.doi.org/10.1177/0962280216666564.

V. Audigier, I. R. White, S. Jolani, T. P. A. Debray, M. Quartagno, J. Carpenter, S. van Buuren, and M. Resche-Rigon. Multiple imputation for multilevel data with continuous and binary variables. *Statist. Sci.*, 33(2):160–183, 05 2018. doi: 10.1214/18-STS646. URL https://doi.org/10.1214/18-STS646.

Stef van Buuren and Karin Groothuis-Oudshoorn. *mice*: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

V. Audigier and M. Resche-Rigon. *micemd: Multiple Imputation by Chained Equations with Multilevel Data*, 2019. R package version 1.6.0.

M. Quartagno and J. Carpenter. *jomo: A package for Multilevel Joint Modelling Multiple Imputation*, 2019. URL http://CRAN.R-project.org/package=jomo. R package version 2.6-8.