Multiple imputation for multilevel data with continuous and binary variables

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

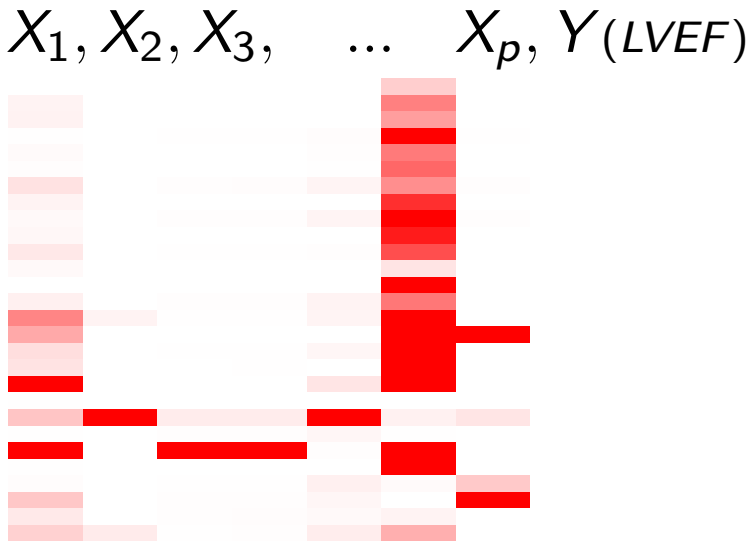
CNAM, MSDMA team, Paris

Séminaire de Statistique et Probabilités Appliquées du LJK, December 14th, Grenoble

Outline

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)

$$y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}\left(0, \sigma^2\right)$$

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

Methods to handle missing values

Missing data are often assumed to be missing at random (MAR)

Ad-hoc methods

- **Complete-case analysis**
  - → generally leads to biased estimates
  - → increases standard errors

- **Single imputation**
  - → leads to unbiased estimates
  - → standard errors are downwardly biased

Methods to handle missing values

Relevant methods

- **Likelihood approaches**
  - Frequentist framework: EM algorithm
  - Bayesian framework: Data Augmentation

  → leads to unbiased estimates
  → not always feasible
  → specific to the analysis model

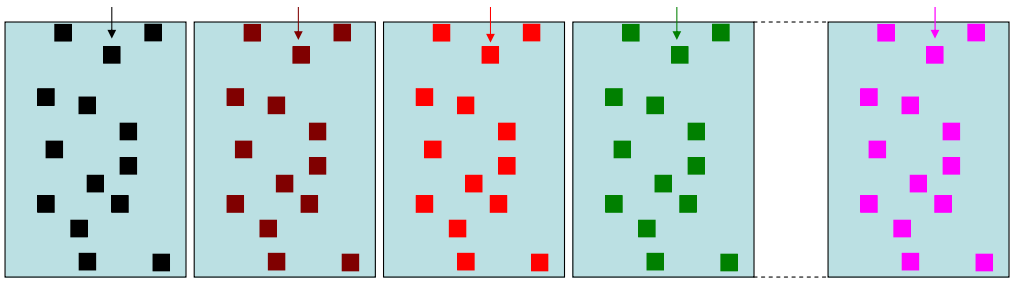
- **Multiple imputation**

  → leads to unbiased estimates
  → can be used for several analysis models

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

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



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

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

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

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

MI 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 be in line with the data

- need to account for the **heterogeneity** between clusters
- need to account for the **types** of data (continuous and binary)

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

* using variant reported in this work

Outline

Challenges: the continuous univariate case

Heteroscedastic mixed-effects model as imputation model

$$\begin{aligned} y_{ik} &= \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \\ b_k &\sim \mathcal{N}(0, \Psi) \\ \varepsilon_{ik} &\sim \mathcal{N}(0, \sigma_k) \end{aligned}$$

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 $\mathbf{y}_{ik}^{miss} | \theta_m, b_k^m$

Specific issues

1. how to generate σ_k without $\mathbf{y}_{ik}$? (systematic)
2. how to draw b_k^m without $\mathbf{y}_{ik}$ (systematic) or given $\mathbf{y}_{ik}$ (sporadic)?

FCS-GLM (Jolani, 2017)

Conditional imputation models

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

For each incomplete variable

1. generate $\theta_m = (\beta_m, \Psi_m, {\sigma^2}_m) \qquad 1 \leq m \leq M$
   - prior: non-informative (Jeffreys)
   - posterior distribution

$$\sigma^2|Y, b \sim \text{Inv-}\Gamma\left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma^2}}{2}\right) \quad \beta|Y, b, \sigma^2 \sim \mathcal{N}\left(\widehat{\beta}, \widehat{\text{var}}\left(\widehat{\beta}\right)\right) \quad \Psi^{-1}|Y, b \sim \mathcal{W}\left(K, \widehat{bb^\top}\right)$$

$\rightarrow$ **requires REML estimate**

2. impute in each cluster k with **systematically missing data**
   - draw $b_k \sim \mathcal{N}(0, \Psi_m)$
   - impute data according to the imputation model

FCS-GLM (Jolani, 2017)

Conditional imputation models

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

For each incomplete variable

1. generate $\theta_m = (\beta_m, \Psi_m, {\sigma^2}_m) \qquad 1 \leq m \leq M$
   - prior: non-informative (Jeffreys)
   - posterior distribution

$$\sigma^2|Y, b \sim \text{Inv-}\Gamma\left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma^2}}{2}\right) \quad \beta|Y, b, \sigma^2 \sim \mathcal{N}\left(\widehat{\beta}, \widehat{\text{var}}\left(\widehat{\beta}\right)\right) \quad \Psi^{-1}|Y, b \sim \mathcal{W}\left(K, \widehat{bb^\top}\right)$$

$\rightarrow$ **requires REML estimate**

2. impute in each cluster k with **sporadically missing data**
   - draw $b_k \sim \mathcal{N}\left(\mu_{b_k|y_k}, \Psi_{b_k|y_k}\right)$
   - impute data according to the imputation model

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

Conditional imputation models

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

→ **the same imputation model, with heteroscedastic assumption**

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**

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

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

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

   - 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

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

→ **the same imputation model, with heteroscedastic assumption**

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**

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

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

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

   - 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 step 2
   - impute data according to the imputation model

JM-jomo (Quartagno and Carpenter, 2015)

$$y_{ik} = x_{ik}\beta + z_{ik}b_k + \varepsilon_{ik} \qquad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_{ik} \sim \mathcal{N}(0, \Sigma_k)$$

1. Bayesian formulation to generate $\theta_m = (\beta_m, \Psi_m, (\Sigma_1, \Sigma_2, ..., \Sigma_K)_m)_{1 \leq m \leq M}$
   - (conjugate) prior: $\beta \propto 1, \quad \Psi^{-1} \sim W(\nu_1, \Lambda_1), \quad \Sigma_k^{-1} | \nu_2, \Lambda_2 \sim W(\nu_2, \Lambda_2)$
   - posterior: unknown explicitly **but...**
     - conditional posterior distributions are known without missing values $\rightarrow$ **Gibbs sampler**
     - with missing values, alternate imputation and draw from the posterior (**Data augmentation**)
2. Imputation (given by step 1)

Binary variables

- **FCS-GLM** (Jolani et al., 2015)
  - fit a logistic model with mixed effect to all clusters

    → sporadically missing values not properly handled
- **FCS-2stage** (Resche-Rigon and White, 2016)
  - fit a logistic model with fixed effect to each cluster
  - combine estimates using a random intercept model

    → large clusters are required
- **JM-jomo** (Quartagno and Carpenter, 2015)
  - probit link: outcomes are latent normal variables, variance for errors are fixed to 1
  - draw latent normal variables
  - derive categories

    → potentially more time consuming

Differences between MI methods

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

- conjugate prior distributions are known to very informative in GLMM
- heteroscedastic assumption is more flexible
- logit link does not allow imputation of sporadically binary variables for FCS-GLM

Outline

Simulation design

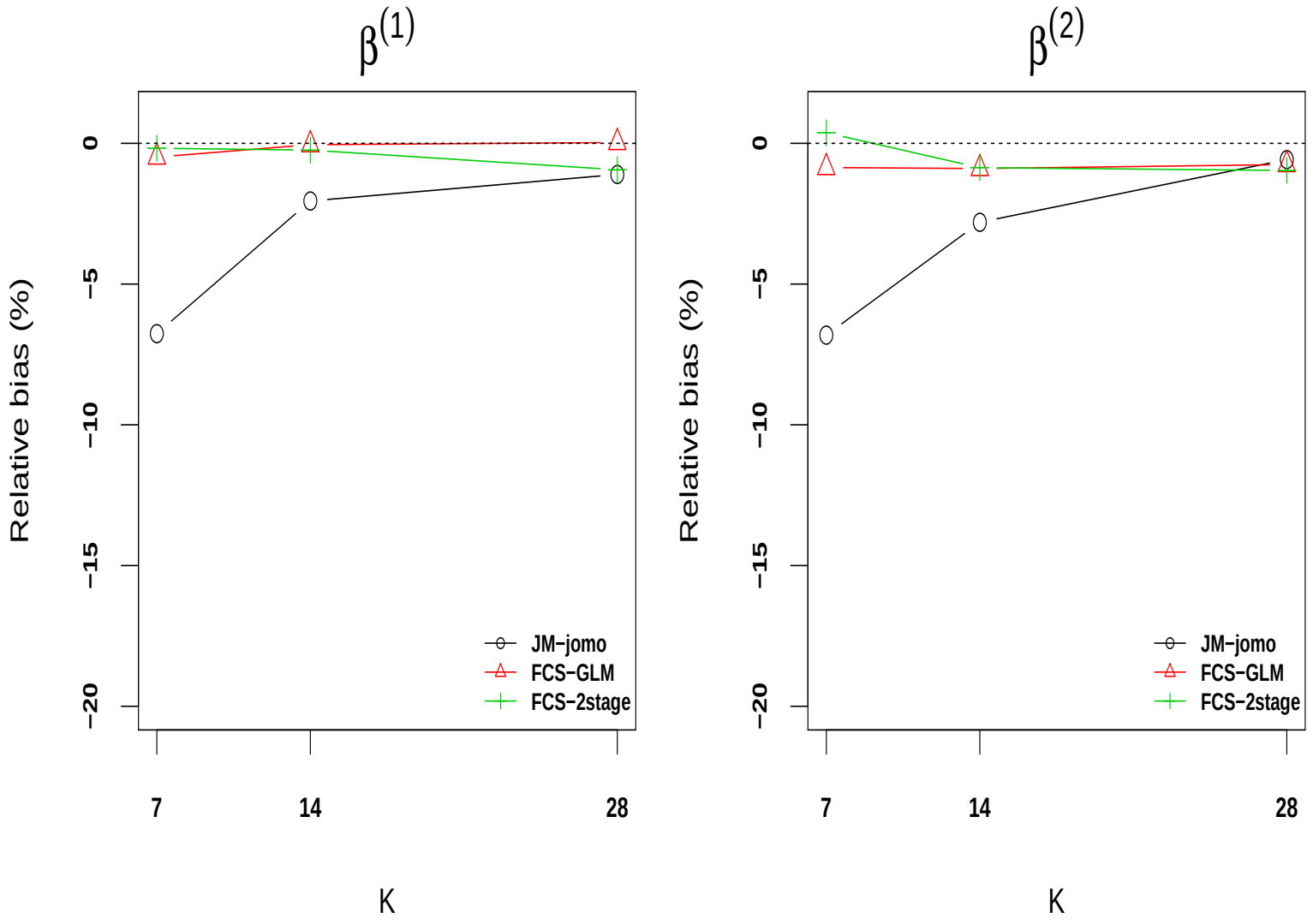
- **Data generation**: 500 incomplete data sets are independently simulated
  ($n = 11685, K = 28, 18 \leq n_k \leq 1834$)
  - $y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$
    with $\beta = (.72, -.11, .03)$, $\Psi = \begin{bmatrix} .0077 & .0015 \\ .0015 & .0004 \end{bmatrix}$, $\sigma = .15$
  - $(\mu_k, \nu_k, \xi_k) \sim \mathcal{N}\left(0, \begin{bmatrix} .12 & .001 & .001 \\ .001 & .12 & .001 \\ .001 & .001 & .12 \end{bmatrix}\right)$
  - $\mathbf{x}_{ik}^{(1)} : \mathcal{N}(2.9 + \mu_k, .36)$
  - $\mathbf{x}_{ik}^{(2)} : logit\left(P\left(\mathbf{x}_{ik}^{(2)} = 1\right)\right) = 4.2 + \nu_k$
  - $\mathbf{x}_{ik}^{(3)}$: $\mathcal{N}(2.9 + \xi_k, .36)$
  - add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ with $\pi_{syst} = .25$ and $\pi_{spor} = .25$

Simulation design

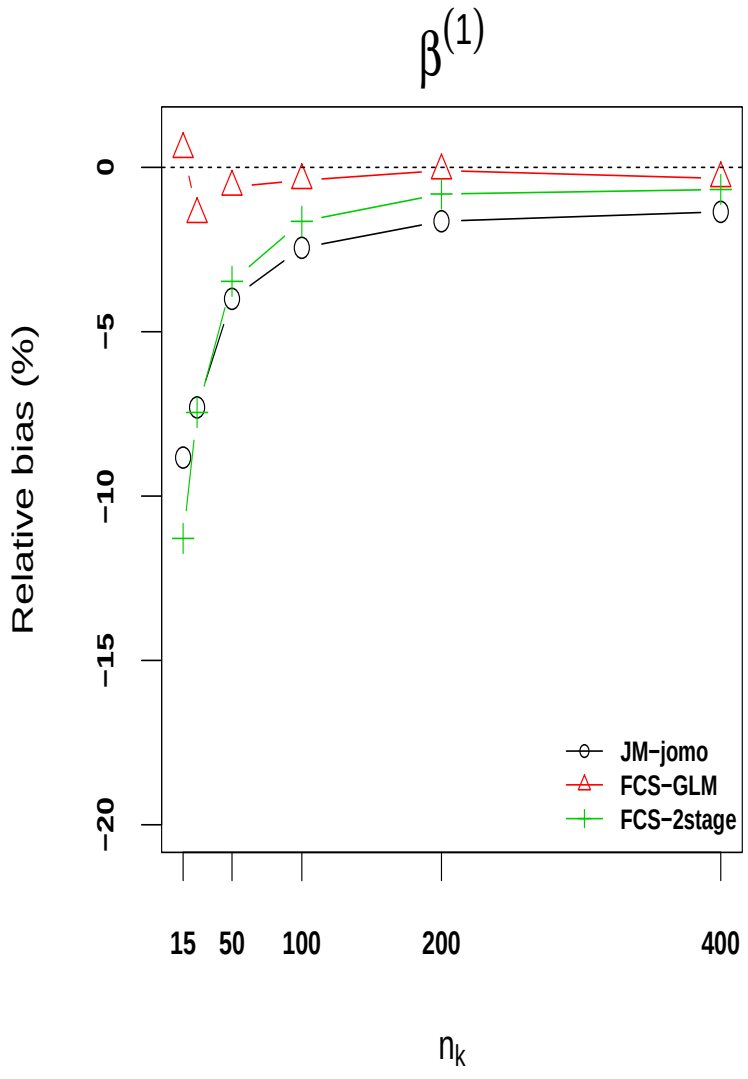
- **Methods**
  - JM-jomo, FCS-GLM, FCS-2stage
  - Full, CC, FCS-fix, FCS-noclust
  - $M = 5$ imputed arrays
- **Estimands**: β and $var\left(\widehat{\beta}\right)$
- **Criteria**: bias, rmse, variance estimate, coverage

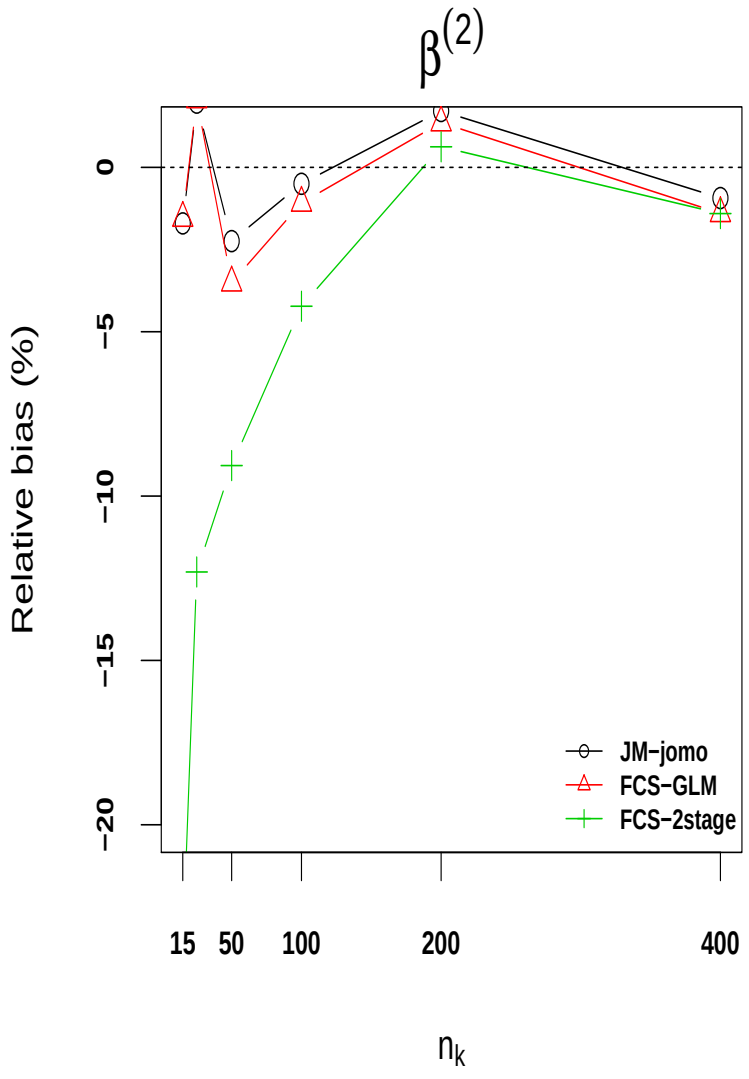
	Full	CC	FCS-noclust	FCS-fixclust	JM-jomo	FCS-GLM	FCS-2stage
$\beta^{(1)}$ est	-0.1101	-0.1104	-0.1039	-0.1102	-0.1088	-0.1100	-0.1090
$\beta^{(1)}$ rbias (%)	0.1	0.3	-5.6	0.2	-1.1	0.0	-0.9
$\beta^{(1)}$ model se	0.0047	0.0070	0.0041	0.0043	0.0066	0.0047	0.0059
$\beta^{(1)}$ emp se	0.0048	0.0071	0.0067	0.0058	0.0056	0.0057	0.0058
$\beta^{(1)}$ 95% cover	93.8	92.2	58.2	87.0	98.4	89.7	95.0
$\beta^{(1)}$ rmse	0.0048	0.0071	0.0091	0.0058	0.0058	0.0057	0.0059
$\beta^{(2)}$ est	0.0301	0.0299	0.0290	0.0300	0.0298	0.0298	0.0297
$\beta^{(2)}$ rbias (%)	0.2	-0.4	-3.4	0.0	-0.6	-0.8	-1.0
$\beta^{(2)}$ model se	0.0029	0.0053	0.0043	0.0043	0.0069	0.0046	0.0049
$\beta^{(2)}$ emp se	0.0030	0.0053	0.0045	0.0042	0.0049	0.0043	0.0044
$\beta^{(2)}$ 95% cover	94.2	94.4	92.0	94.6	97.6	95.8	96.2
$\beta^{(2)}$ rmse	0.0030	0.0053	0.0046	0.0042	0.0049	0.0043	0.0044
time (min)			0.9	1.1	7.8	103.3	0.9

Influence of the number of clusters

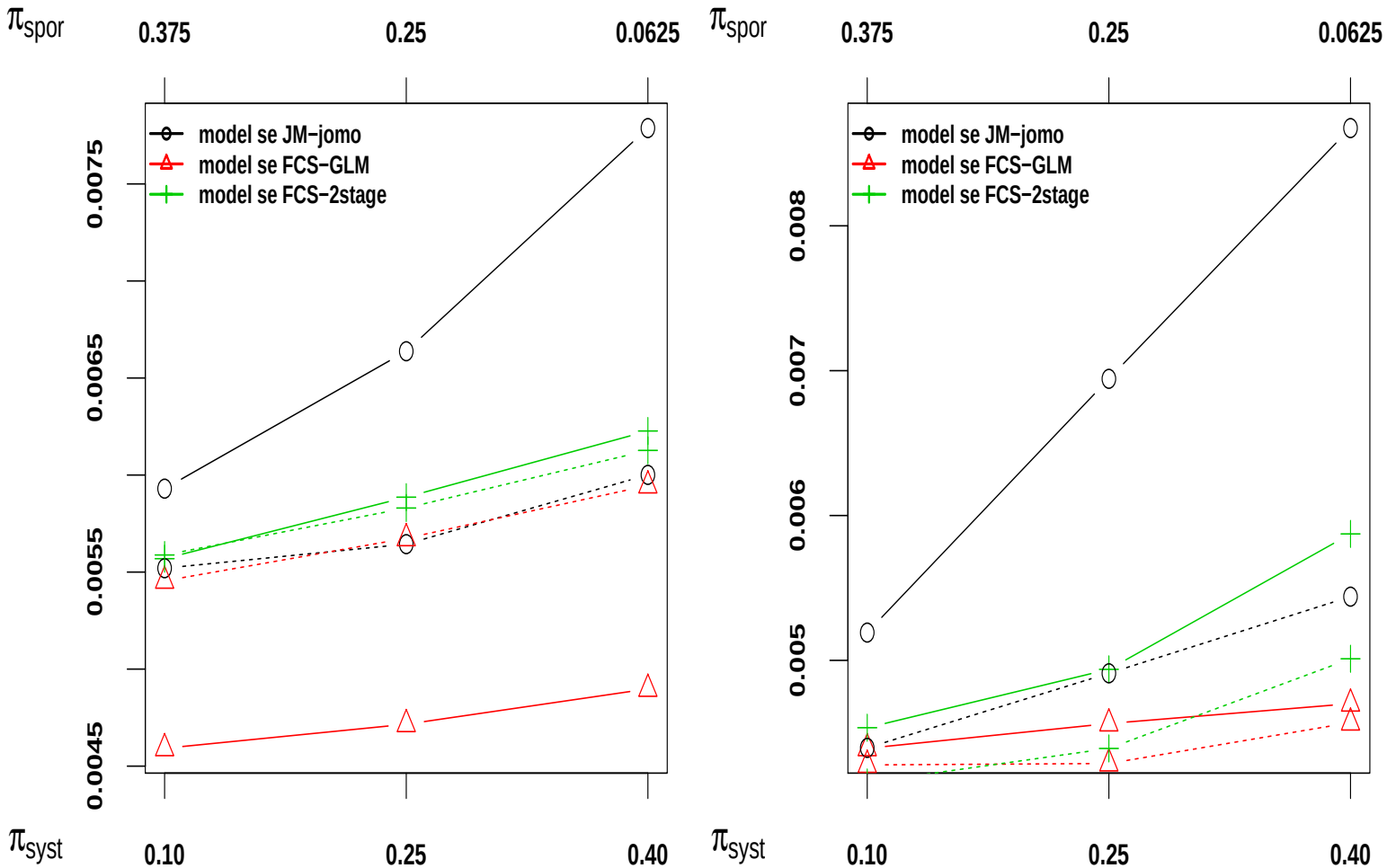


Influence of the cluster size



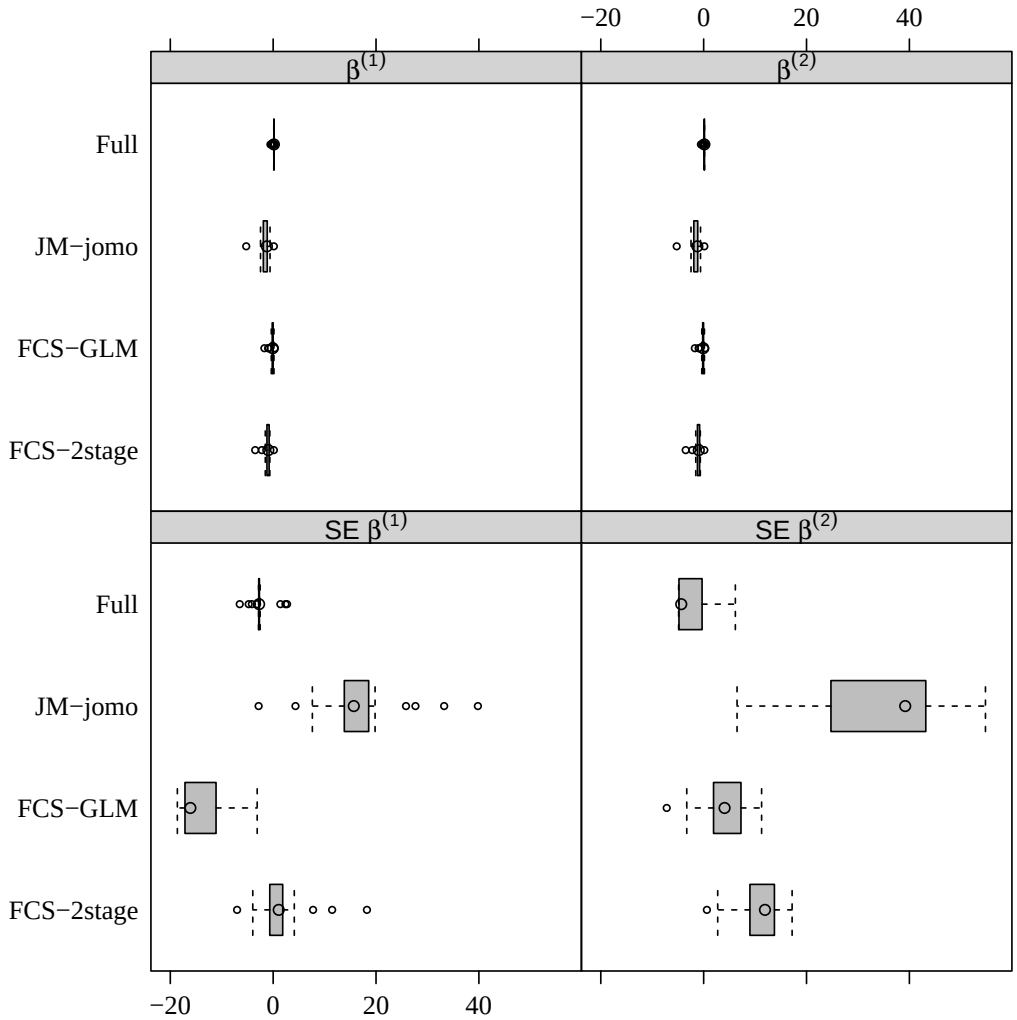


Influence of the proportion of systematically missing values



Other configurations

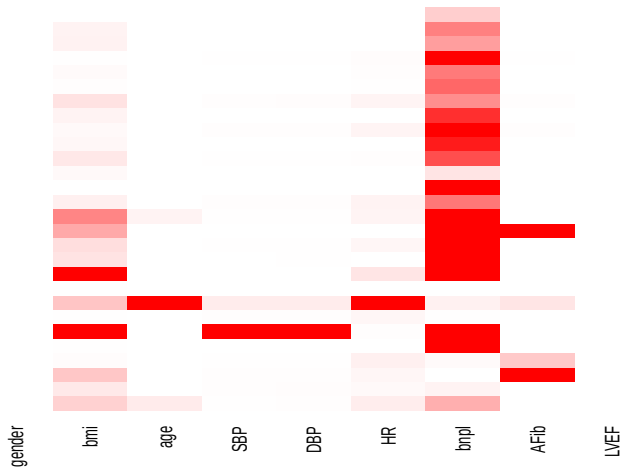
- 19 configurations
  - MAR mechanism
  - binary outcome
  - probit model for covariates
  - ...
- similar behaviours to base-case configuration



Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- $y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$
- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays



		CC	JM-jomo	FCS-GLM	FCS-2stage
β_{BNP}	est	-0.1132	-0.0891	-0.1002	-0.1009
	model se	0.0108	0.0078	0.0163	0.0112
β_{AFIB}	est	0.0268	0.0216	0.0218	0.0273
	model se	0.0071	0.0046	0.0066	0.0045
time (min)			94.0	30819.5	31.8

Conclusion

An overview of MI methods for multilevel data

- FCS-GLM, FCS-2stage and JM-jomo all appear to perform well
- Outperform had-hoc methods

FCS-2stage	• provides a quick way to obtain first results • for large clusters
FCS-GLM	• tends to underestimate the variance of the estimator because of the homoscedastic assumption • recommended with small clusters • time consuming (with binary variables)
JM-jomo	• tends to overestimate the variance of the fixed coefficients because of unsuitable prior distributions • recommended for large clusters when the proportion of binary variables is high

Methods are implemented in R (micemd, jomo)

Limits and perspectives

Limits

- congeniality (analysis model vs imputation model)
- choice of random and fixed effects in the imputation models
- binary variables are challenging

Perspectives

- imputation using machine learning
- large number of variables
- ordinal and nominal variables
- MNAR mechanism

References I

Great Network. Managing acute heart failure in the ed - case studies from the acute heart failure academy, 2013. http://www.greatnetwork.org.

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

S. Jolani. Hierarchical imputation of systematically and sporadically missing data: An approximate bayesian approach using chained equations. *Biometrical Journal*, 2017. doi: 10.1002/bimj.201600220.

Matthieu Resche-Rigon and Ian White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 2016.

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.

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. *ArXiv e-prints*, 2017.