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


INSERM, UMR 1153, ECSTRA team, Saint-Louis Hospital, Paris

ISCB 2016, August 21-25th, Birmingham
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)

\[ y_{ik} = x_{ik} \beta + z_{ik} b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \]

\[ \hat{\beta} \text{ and associated variability } \text{var}(\hat{\beta}) \]
MI methods 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

Some MI methods to impute multilevel data

- FCS-2lnorm (van Buuren, 2010): continuous / sporadic
- **FCS-1step** (Jolani et al., 2015; Resche-Rigon et al., 2013): mixed / systematic
- **FCS-2step** (Resche-Rigon and White, 2016): mixed / systematic and sporadic
- JM-Pan (Schafer, 1997): continuous / systematic and sporadic
- **JM-jomo** (Quartagno and Carpenter, 2016): mixed / systematic and sporadic

However, only **FCS-1step, FCS-2step** and **JM-jomo** handle systematically missing values and mixed data
Continuous variables

Heteroscedastic mixed-effects model as imputation model

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

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 \( \theta_m \)
   - draw \( b_k^m | y_{ik}^{obs}, \theta_m \)
   - draw \( y_{ik}^{miss} | \theta_m, b_k^m \)

Specific issues

1. how to generate \( \Sigma_k \) without \( y_{ik} \)? (systematic)
2. how to draw \( b_k^m \) without \( y_{ik} \) (systematic) or given \( y_{ik} \) (sporadic)?
FCS-1step (Jolani et al., 2015)

Conditional imputation models

\[ y_{ik} = x_{ik}\beta + z_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \]

For each incomplete variable

1. generate \( \theta_m = (\beta_m, \Psi_m, \sigma^2_m) \) \( 1 \leq m \leq M \)
   - estimate \( \theta \) and \( var(\hat{\theta}) \) by REML
   - draw \( \theta_m \) from an appropriate distribution with expectation \( \hat{\theta} \), and variance \( \hat{var}(\hat{\theta}) \)

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-1step (Jolani et al., 2015)

Conditional imputation models

\[ y_{ik} = x_{ik}\beta + z_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \]

For each incomplete variable

1. generate \( \theta_m = (\beta_m, \Psi_m, \sigma^2_m) \quad 1 \leq m \leq M \)
   - estimate \( \theta \) and \( \text{var} (\hat{\theta}) \) by REML
   - draw \( \theta_m \) from an appropriate distribution with expectation \( \hat{\theta} \), and variance \( \hat{\text{var}} (\hat{\theta}) \)

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

\[ y_{ik} = x_{ik}\beta_k + \varepsilon_{ik} \quad \beta_k = \beta + b_k \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma_k^2) \]

→ the same imputation model, with heteroscedastic assumption

1. generate \( \theta_m = (\beta_m, \Psi_m, (\sigma_1^2, \ldots, \sigma_K^2)_m) \)
   - estimate \( \theta \) and \( \text{var} \left( \hat{\theta} \right) \) by a two-step estimator:
     - step a fit \( y_{ik} = x_{ik}\beta_k + \varepsilon_{ik} \) to each cluster
     - step b combine the \( \hat{\beta}_k \) and \( \hat{\sigma}_k^2 \) by multivariate meta-analysis (by REML or MM)
   - draw \( \theta_m \) from an appropriate distribution with expectation \( \hat{\theta} \), and variance \( \hat{\text{var}} \left( \hat{\theta} \right) \)

2. impute in each cluster \( k \)
   - 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
**JM-jomo** (Quartagno and Carpenter, 2016)

\[
\mathbf{y}_{ik} = \mathbf{x}_{ik} \beta + \mathbf{z}_{ik} b_k + \mathbf{e}_{ik} \\
\mathbf{b}_k \sim \mathcal{N}(0, \Psi) \\
\mathbf{e}_{ik} \sim \mathcal{N}(0, \Sigma_k)
\]

1. **Bayesian formulation to generate** $\theta_m = (\beta_m, \Psi_m, \Sigma_m)_{1 \leq m \leq M}$
   - **prior:** $\Sigma_k^{-1} \sim W(\nu_1, \Lambda_1), \quad \Psi^{-1} \sim W(\nu_2, \Lambda_2), \quad \beta \propto 1$
   - **posterior:** unknown $\rightarrow$ **Gibbs sampler**

   \[
   \beta^{(\ell+1)} \sim \mathcal{P} \left( \beta | \mathbf{X}^{obs}, \mathbf{X}^{miss(\ell)}, \Sigma^{(\ell)}, b^{(\ell)} \right) \\
   b_k^{(\ell+1)} \sim \mathcal{P} \left( b_k | \mathbf{X}^{obs}, \mathbf{X}^{miss(\ell)}, \beta^{(\ell+1)}, \Psi^{(\ell)}, \Sigma_k^{(\ell)} \right) \\
   \Psi^{-1(\ell+1)} \sim \mathcal{P} \left( \Psi^{-1} | \mathbf{X}^{obs}, \mathbf{X}^{miss(\ell)}, b^{(\ell+1)} \right) \\
   \Sigma_k^{-1(\ell+1)} \sim \mathcal{P} \left( \Sigma_k^{-1} | \mathbf{X}^{obs}, \mathbf{X}^{miss(\ell)}, b_k^{(\ell+1)} \right) \\
   \mathbf{X}_k^{miss(\ell+1)} \sim \mathcal{P} \left( \mathbf{X}_k^{miss} | \mathbf{X}^{obs}, \beta^{(\ell+1)}, \Psi^{(\ell+1)}, \Sigma^{(\ell+1)}, b_k^{(\ell+1)} \right)
   \]

2. **Imputation** (given by step 1)
Binary variables

- **FCS-1step** (Jolani et al., 2015)
  - fit a logistic model with mixed effect to all clusters
  → sporadically missing values not handled

- **FCS-2step** (Resche-Rigon and White, 2016)
  - fit a logistic model with fixed effect to each cluster
  - combine estimates using a meta-analysis
  → large clusters are required

- **JM-jomo** (Quartagno and Carpenter, 2016)
  - draw latent normal variables
  - derive categories
  → more time consuming
Simulation design

- **Data generation**: 500 incomplete data sets are independently simulated
  \((n = 11685, \quad K = 28, \quad 18 \leq n_k \leq 1834)\)

  - \(y_{ik} = \beta^0 + \beta^1 x_{ik}^{(1)} + \beta^2 x_{ik}^{(2)} + b_k^0 + b_k^1 x_{ik}^{(1)} + \varepsilon_{ik}\)
    
    \(\text{with } \beta = (.72, .11, .03), \quad \Psi = \begin{bmatrix} .0077 \\ .0015 \\ .0015 \\ .0004 \end{bmatrix}, \quad \sigma = .15\)

  - \(x_{ik}^{(1)} : \mathcal{N}(\mu + \mu_k, .36)\)

  - \(x_{ik}^{(2)} : \text{logit}\left(\mathcal{P}\left(x_{ik}^{(2)} = 1\right)\right) = \nu + \nu_k\)

  - add missing values on \(x^{(1)}, x^{(2)}\) varying \(\pi_{syst}\) and \(\pi_{spor}\)

- **Analysis**: \(\beta\) and \(\text{var}\left(\hat{\beta}\right)\) estimated by applying MI methods using \(M = 5\) imputed arrays

- **Criteria**: bias, rmse, variance estimate, coverage
Results: $\pi_{syst} = .1$, $\pi_{spor} = .375$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$ $\beta_1$</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$ $\beta_2$</th>
<th>95% Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>93.8</td>
<td>94.2</td>
</tr>
<tr>
<td>CC</td>
<td>0.0063</td>
<td>0.0052</td>
<td>90.4</td>
<td>93.4</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>0.0086</td>
<td>0.0044</td>
<td>71.8</td>
<td>57.0</td>
</tr>
<tr>
<td>FCS-1step</td>
<td>0.0050</td>
<td>0.0044</td>
<td>91.6</td>
<td>95.4</td>
</tr>
<tr>
<td>FCS-2step-mm</td>
<td>0.0055</td>
<td>0.0047</td>
<td>94.2</td>
<td>95.5</td>
</tr>
<tr>
<td>FCS-2step-re</td>
<td>0.0056</td>
<td>0.0045</td>
<td>93.2</td>
<td>95.9</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0059</td>
<td>0.0053</td>
<td>95.8</td>
<td>96.0</td>
</tr>
</tbody>
</table>

$\hat{\beta}$ is the estimated coefficient, and $\text{var} (\hat{\beta})$ is the variance of the estimated coefficient.
Results: $\pi_{syst} = 0.1$, $\pi_{spor} = 0.375$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} \left( \hat{\beta} \right)}$</th>
<th>$\text{var} \left( \hat{\beta} \right)$</th>
<th>95% Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>0.0048</td>
<td>0.0030</td>
</tr>
<tr>
<td>CC</td>
<td>0.0063</td>
<td>0.0052</td>
<td>0.0066</td>
<td>0.0053</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>0.0086</td>
<td>0.0044</td>
<td>0.0086</td>
<td>0.0068</td>
</tr>
<tr>
<td>FCS-1step</td>
<td>0.0050</td>
<td>0.0044</td>
<td>0.0055</td>
<td>0.0043</td>
</tr>
<tr>
<td>FCS-2step-mm</td>
<td>0.0055</td>
<td>0.0047</td>
<td>0.0055</td>
<td>0.0042</td>
</tr>
<tr>
<td>FCS-2step-re</td>
<td>0.0056</td>
<td>0.0045</td>
<td>0.0056</td>
<td>0.0042</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0059</td>
<td>0.0053</td>
<td>0.0054</td>
<td>0.0043</td>
</tr>
</tbody>
</table>
Results: $\pi_{\text{syst}} = .1$, $\pi_{\text{spor}} = .375$
Results: $\pi_{syst} = 0.1$, $\pi_{spor} = 0.375$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$ $\beta_1$</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$ $\beta_2$</th>
<th>95% Cover $\beta_1$</th>
<th>95% Cover $\beta_2$</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>0.0048</td>
<td>0.0030</td>
<td>93.8</td>
</tr>
<tr>
<td>CC</td>
<td>0.0063</td>
<td>0.0052</td>
<td>0.0066</td>
<td>0.0053</td>
<td>90.4</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>0.0086</td>
<td>0.0044</td>
<td>0.0086</td>
<td>0.0068</td>
<td>71.8</td>
</tr>
<tr>
<td>FCS-1step</td>
<td>0.0050</td>
<td>0.0044</td>
<td>0.0055</td>
<td>0.0043</td>
<td>91.6</td>
</tr>
<tr>
<td>FCS-2step-mm</td>
<td>0.0055</td>
<td>0.0047</td>
<td>0.0055</td>
<td>0.0042</td>
<td>94.2</td>
</tr>
<tr>
<td>FCS-2step-re</td>
<td>0.0056</td>
<td>0.0045</td>
<td>0.0056</td>
<td>0.0042</td>
<td>93.2</td>
</tr>
<tr>
<td>JM-jomo</td>
<td><strong>0.0059</strong></td>
<td><strong>0.0053</strong></td>
<td><strong>0.0054</strong></td>
<td><strong>0.0043</strong></td>
<td>95.8</td>
</tr>
</tbody>
</table>
Results: $\pi_{\text{syst}} = .1$, $\pi_{\text{spor}} = .375$
Results: $\pi_{syst} = .25, \pi_{spor} = .25$

![Boxplot representing the differences between estimated and true values of $\beta_1$ and $\beta_2$ for various methods.]

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>95% Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>0.0048</td>
<td>0.0030</td>
</tr>
<tr>
<td>CC</td>
<td>0.0070</td>
<td>0.0053</td>
<td>0.0071</td>
<td>0.0053</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>0.0105</td>
<td>0.0038</td>
<td>0.0106</td>
<td>0.0058</td>
</tr>
<tr>
<td>FCS-1step</td>
<td>0.0049</td>
<td>0.0046</td>
<td>0.0056</td>
<td>0.0043</td>
</tr>
<tr>
<td>FCS-2step-mm</td>
<td>0.0059</td>
<td>0.0054</td>
<td>0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>FCS-2step-re</td>
<td>0.0059</td>
<td>0.0049</td>
<td>0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0066</td>
<td>0.0069</td>
<td>0.0057</td>
<td>0.0050</td>
</tr>
</tbody>
</table>
Results: $\pi_{\text{syst}} = .25$, $\pi_{\text{spor}} = .25$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>95% Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>0.0048</td>
<td>0.0030</td>
</tr>
<tr>
<td>CC</td>
<td>0.0070</td>
<td>0.0053</td>
<td>0.0071</td>
<td>0.0053</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>0.0105</td>
<td>0.0038</td>
<td>0.0106</td>
<td>0.0058</td>
</tr>
<tr>
<td>FCS-1step</td>
<td>0.0049</td>
<td>0.0046</td>
<td>0.0056</td>
<td>0.0043</td>
</tr>
<tr>
<td>FCS-2step-mm</td>
<td>0.0059</td>
<td>0.0054</td>
<td>0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>FCS-2step-re</td>
<td>0.0059</td>
<td>0.0049</td>
<td>0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0066</td>
<td>0.0069</td>
<td>0.0057</td>
<td>0.0050</td>
</tr>
</tbody>
</table>
An overview of MI methods for multilevel mixed data

- standard methods are irrelevant
- FCS-1step, FSC-2step and JM-jomo all appear to perform well
- inference performances are quite similar
- FCS-2step is quicker to perform

Perspectives

- a larger simulation study (MAR mechanism, number of clusters, size of clusters,...)
- a precise guidance
References I


