Introduction

Nowadays: Availability of imputation methods in standard software facilitates automated imputation of incomplete data. For example in R:

```
### Multiple Imputation (MI) using chained equations
mice
### Joint Model MI using a multi-
### variate normal model (MVN)
Jomo
### Bayesian joint model sequential-
### factorization imputation
JointAI
```

Impute missing values by draws from the (posterior) predictive distribution of an incomplete variable, conditional on (all) other variables.

- The predictive distributions need to fit the data well!

However:
- imputation models are specified automatically by the software
- in practice often no effort is made to check the validity of the postulated models

Robustness of MI - Some Findings from Literature

Normal imputation model for non-normal data
- MICE & MVN robust for inference about the mean
- more flexible distributions necessary when interest is in quantiles
- MICE: non-/semi-parametric methods often better

Comparison between approaches
- MVN & MICE similarly robust
- misspecified MICE better than compl. case analysis
- doubly robust IPW may be even better than MICE

Sequental Factorization Imputation

Fully Bayesian approach allowing simultaneous analysis and imputation:
- factorize joint distribution as sequence of conditional distributions,
- one of which one is the analysis model of interest:

\[
p(y | x) = \prod_i p(y_i | x) = \prod_i p(y_i | x_i, \ldots, x_{i-1}, \theta)
\]

\[
\pi(\theta) = \prod_i \pi(\theta_i)
\]

Notation:
- \( X = (X_{x_{obs}}, x_{mis}) \) design matrix of completely observed and incomplete covariates
- \( x_{mis} = (x_1, \ldots, x_i) \)
- \( \theta = (\theta_1, \ldots, \theta_i) \)

- Draw imputations from the posterior Predictive Distribution (PPD) (e.g., for a covariate \( x_i \)):

\[
p(x_i | y, x, x_i, x_{mis}, \theta) \propto p(y | x, x_i, \theta) \prod_i p(x_i | x_i, x_{mis}, \theta)
\]

- \( \prod_i p(x_i | x_i, \theta) \pi(\theta) \prod_i \pi(\theta_i) \)

- \( \pi(\theta) \)

<table>
<thead>
<tr>
<th>cond. distr. of ( x_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi(\theta) )</td>
</tr>
</tbody>
</table>

- PPD specified indirectly => direct evaluation of its fit not possible

Our Research Question:

How robust is sequential factorization imputation to misspecification of conditional distributions?

Investigating Robustness by Simulation

- Analysis model:
  - linear regression with 4 covariates
  - \( y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \sigma^2) \)
  - \( x_1 \sim N(0, 1) \) or \( x_2 \sim \text{Gamma}(5, 10) \)
  - \( x_3 \sim \text{Bin}(0.5) \)
  - \( x_4 \sim \text{Bin}(\exp(x_1 + 0.5 x_2 + 0.5 x_3)) \)

- \( \beta = (0.1, 0.5) \) or \( \beta = (0.25, 0.5, 0.75, 1.0) \)

- sequence: bias in all \( \hat{\beta} \), but worse for MICE; coverage \( \geq 0.65 \)
- interaction: MICE more severely biased in all \( \beta \), coverage \( \geq 0.5 \), JointAI \( \geq 0.75 \)

Conclusions

- Misspecification of the cond. distributions translates to misspecified imputation models.
- In most of our scenarios: JointAI performed (slightly) better than MICE.
- Fit of the cond. distributions needs to be validated to obtain unbiased results.
- More flexible models are needed to assure appropriate performance in practice, where imputation is often used in a “black-box” manner.