# Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

September 13, 2023

# Multivariate missingness

## Context

- $\hookrightarrow$ For the regression imputation and stochastic regression imputation approaches, we have assumed that only one variable was subject to missingness.
- $\hookrightarrow$ Often we have missing values in more than one variable.
- $\hookrightarrow$ What should we do in such a case?

# Multivariate missingness

## Monotone pattern

- In the case of a monotone missing pattern, we can use the technique used for univariate missing data (e.g., stochastic regression imputation) in a chain.
- Let us consider a specific example to understand how we should proceed.

| $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ |
|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ |
| ✓ | ✓ | ✓ | NA |
| ✓ | ✓ | NA | NA |
| ✓ | NA | NA | NA |
| ✓ | NA | NA | NA |

- Impute $Y_2$ given $Y_1$.
- Impute $Y_3$ given $Y_1$ and $Y_2$.
- Impute $Y_4$ given $Y_1$, $Y_2$, and $Y_3$.

# Multivariate missingness

## Non-monotone pattern

- ↪ There are two popular approaches for imputation in multivariate non-monotone missing data:
  - ↪ Joint model imputation.
  - ↪ Fully conditional specification.

# Multivariate missingness

## Joint model imputation

- ↪ Joint model imputation fits a multivariate model to all the variables that have missingness, thus generalising what we have seen before.

- ↪ For instance, if we have $p$ variables, say $(Y_1, \ldots, Y_p)$ and if $Y_1$, $Y_2$, and $Y_3$ are subject to missingness, joint model imputation requires to specify a model for $f(Y_1, Y_2, Y_3 \mid Y^*)$, where $Y^* = (Y_4, \ldots, Y_p)$.

- ↪ The main drawback of this approach is that it is not always trivial to set up a reasonable multivariate regression model.

- ↪ As a consequence, in practice, an off-the-shelf model is typically used, most commonly the multivariate normal or $t$ distributions for continuous variables and a multinomial distribution for discrete variables.

- ↪ The positive point is that software exists to fit such models automatically (e.g., the `norm`, `cat`, `mix`, `jomo`, and `jointAI` packages).

# Multivariate missingness

## Fully conditional specification

- Fully conditional specification (FCS) (or multiple imputation by chained equations (MICE)) imputes multivariate missing data on a variable-by-variable basis.
- As before, suppose we have partially observed variables $(Y_1, Y_2, Y_3)$ and some fully observed variables $Y^* = (Y_4, \ldots, Y_p)$.
- Under the fully conditional specification approach, we specify regression models for

$$
\begin{aligned}
&f(Y_1 \mid Y_2, Y_3, Y^*),\\
&f(Y_2 \mid Y_1, Y_3, Y^*),\\
&f(Y_3 \mid Y_1, Y_2, Y^*).
\end{aligned}
$$

- If, for instance $Y_1$ is continuous, we might choose a linear regression for the first model.
- If, for instance $Y_2$ is binary, we might choose a logistic regression for the second model.

# Multivariate missingness

## Fully conditional specification

↪ Fully conditional specification would consist of the following steps:

1. Initially impute missing values in $Y_1$, $Y_2$, and $Y_3$ by randomly sampling from the observed values.

2. Impute missing values in $Y_1$ using the model $f(Y_1 \mid Y_2, Y_3, Y^*)$ (using observed $Y_1$ values and observed and imputed values of $Y_2$ and $Y_3$ and fully observed variables $Y^*$).

3. Impute missing values in $Y_2$ using the model $f(Y_2 \mid Y_1, Y_3, Y^*)$ (using observed $Y_2$ values and observed and imputed values of $Y_1$ and $Y_3$ and fully observed variables $Y^*$).

4. Impute missing values in $Y_3$ using the model $f(Y_3 \mid Y_1, Y_2, Y^*)$ (using observed $Y_3$ values and observed and imputed values of $Y_1$ and $Y_2$ and fully observed variables $Y^*$).

5. Iterate between the three steps above until approximate convergence.

↪ The imputed values from the last iteration are then used to replace the missing values in the original dataset.

# Multivariate missingness

## Fully conditional specification

- ↪ The major advantage of FCS/MICE (over the joint model imputation) is that the sequence of univariate regression models is easier to understand, thus allowing one to fit a reasonable model at each step.
- ↪ A theoretical issue with FCS/MICE is that there is no guarantee is that the algorithm draws imputations from a well defined joint/multivariate model.
- ↪ Recent work (e.g., Hughes et al. 2014) has identified certain conditions when it does and the key condition is that the conditional models are compatible.
- ↪ By compatible the authors mean that there exist multivariate distributions whose conditionals are those specified in FCS/MICE.