EP16: Missing Values in Clinical Research: Multiple Imputation

2. Imputation Step

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

✉ n.erler@erasmusmc.nl

Erasmus MC
University Medical Center Rotterdam

Univariate Missing Data

How can we actually get imputed values?

Univariate Missing Data

How can we actually get imputed values?

For now: assume only one continuous variable has missing values (**univariate missing data**).

X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	✓	✓	✓
✓	NA	✓	✓
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Univariate Missing Data

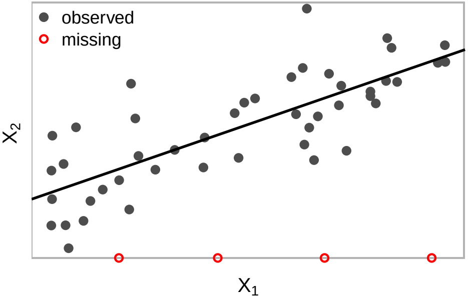
How can we actually get imputed values?

For now: assume only one continuous variable has missing values (**univariate missing data**).

X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	✓	✓	✓
✓	NA	✓	✓
⋮	⋮	⋮	⋮

Idea: Predict values

Model: $x_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i3} + \beta_3 x_{i4} + \varepsilon_i$



Univariate Missing Data

How can we actually get imputed values?

For now: assume only one continuous variable has missing values (**univariate missing data**).

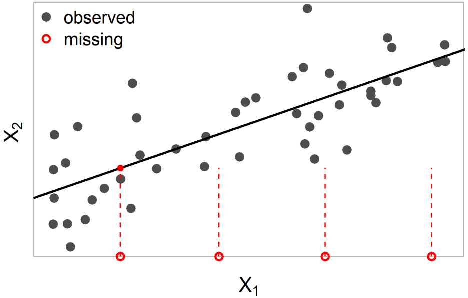
X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	✓	✓	✓
✓	NA	✓	✓
⋮	⋮	⋮	⋮

Idea: Predict values

Model: $x_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i3} + \beta_3 x_{i4} + \varepsilon_i$

Imputed/predicted value:

$\hat{x}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i3} + \hat{\beta}_3 x_{i4}$



Univariate Missing Data

Problem:

- We can obtain **only one imputed value** per missing value (but we wanted a whole distribution).
- The predicted values do not take into account the added **uncertainty** due to the missing values.

Univariate Missing Data

Problem:

- We can obtain **only one imputed value** per missing value (but we wanted a whole distribution).
- The predicted values do not take into account the added **uncertainty** due to the missing values.

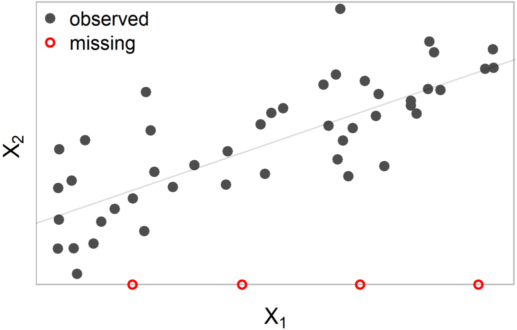
➡ We need to take into account **two sources of uncertainty**:

- The **parameters** are estimated with **uncertainty** (represented by the standard error).
- There is **random variation / prediction error** (variation of the residuals).

Univariate Missing Data

Taking into account uncertainty about the parameters:
We assume that β **has a distribution**, and we can sample realizations of β from that distribution.

When plugging the different realizations of β into the predictive model, we obtain **slightly different regression lines**.

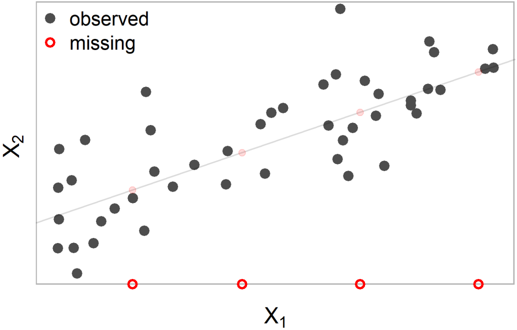


Univariate Missing Data

Taking into account uncertainty about the parameters:
We assume that β **has a distribution**, and we can sample realizations of β from that distribution.

When plugging the different realizations of β into the predictive model, we obtain **slightly different regression lines**.

With each set of coefficients, we also get slightly **different predicted values**.

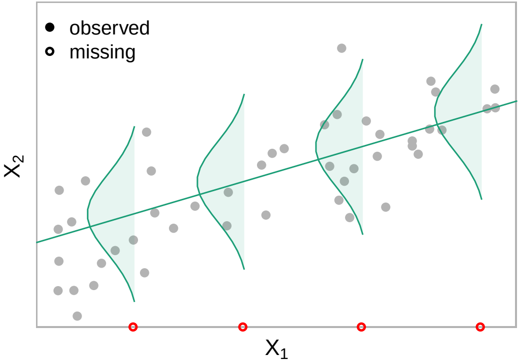


Univariate Missing Data

Taking into account the prediction error:
The model does not fit the data perfectly: observations are scattered around the regression lines.

We assume that the **data have a distribution**, where

- the **mean** for each value is given by the **predictive model**, and
- the **variance** is determined by the variance of the residuals ε.



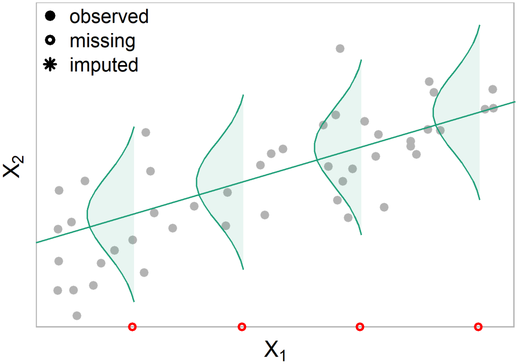
Univariate Missing Data

Taking into account the prediction error:
The model does not fit the data perfectly: observations are scattered around the regression lines.

We assume that the **data have a distribution**, where

- the **mean** for each value is given by the **predictive model**, and
- the **variance** is determined by the variance of the residuals ε.

➡ sample imputed values from this distribution.



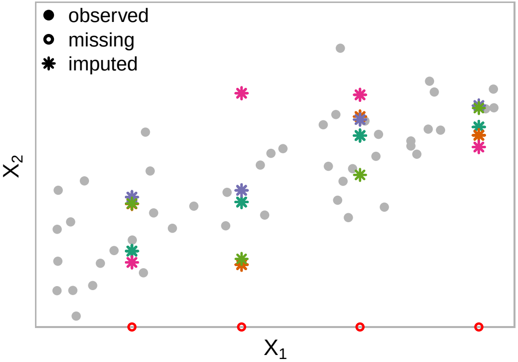
Univariate Missing Data

Taking into account the prediction error:
The model does not fit the data perfectly: observations are scattered around the regression lines.

We assume that the **data have a distribution**, where

- the **mean** for each value is given by the **predictive model**, and
- the **variance** is determined by the variance of the residuals ε.

➡ sample imputed values from this distribution.



In the end, we obtain one imputed dataset for each colour.

Multivariate Missing Data

Multivariate missing data:
What if we have **missing values in more than one variable**?

Multivariate Missing Data

Multivariate missing data:
What if we have **missing values in more than one variable**?

In case of **monotone missing values** we can use the technique for univariate missing data in a chain:
impute x_4 given x_1
impute x_3 given x_1 and x_4
impute x_2 given x_1, x_4 and x_3

X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	NA	NA	✓
✓	NA	NA	NA
⋮	⋮	⋮	⋮

Multivariate Missing Data

Multivariate missing data:
What if we have **missing values in more than one variable**?

In case of **monotone missing values** we can use the technique for univariate missing data in a chain:
impute x_4 given x_1
impute x_3 given x_1 and x_4
impute x_2 given x_1, x_4 and x_3

X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	NA	NA	✓
✓	NA	NA	NA
⋮	⋮	⋮	⋮

When we have **non-monotone missing data** there is no sequence without conditioning on unobserved values.

X_1	X_2	X_3	X_4
✓	NA	✓	✓
✓	✓	NA	NA
✓	NA	✓	NA
⋮	⋮	⋮	⋮

Multivariate Missing Data

There are **two popular approaches** for the imputation step in **multivariate non-monotone** missing data:

Fully Conditional Specification

- Multiple Imputation using Chained Equations (**MICE**)
- sometimes also: sequential regression
- implemented in SPSS, R, Stata, SAS, …
- our focus here

Multivariate Missing Data

There are **two popular approaches** for the imputation step in **multivariate non-monotone** missing data:

Fully Conditional Specification

- Multiple Imputation using Chained Equations (**MICE**)
- sometimes also: sequential regression
- implemented in SPSS, R, Stata, SAS, …
- our focus here

Joint Model Imputation

(more details later)

MICE / FCS

MICE (**M**ultiple **I**mputation using **C**hained **E**quations) or
FCS (multiple imputation using **F**ully **C**onditional **S**pecification)

extends univariable imputation to the setting with multivariate non-monotone missingness:

MICE / FCS

- imputes **multivariate** missing data on a **variable-by-variable** basis,
- using the technique for univariate missing data.

MICE / FCS

MICE (**M**ultiple **I**mputation using **C**hained **E**quations) or
FCS (multiple imputation using **F**ully **C**onditional **S**pecification)

extends univariable imputation to the setting with multivariate non-monotone missingness:

MICE / FCS

- imputes **multivariate** missing data on a **variable-by-variable** basis,
- using the technique for univariate missing data.

Moreover, MICE/FCS is

- an **iterative** procedure, specifically
- a **Markov Chain Monte Carlo (MCMC)** method,
- uses the idea of the **Gibbs sampler**

MICE / FCS: Sidenote

Markov Chain Monte Carlo

- a technique to **draw samples from a complex probability distribution**
- works via creating a chain of random variables (a Markov chain)
➡ The distribution that each element in the chain is sampled from depends on the value of the previous element.
- When certain conditions are met, the chain eventually stabilizes
- samples of the chain are then a sample from the complex distribution of interest

MICE / FCS: Sidenote

Gibbs sampling

- a MCMC method to obtain a **sample from a multivariate distribution**
- the multivariate distribution is split into a set of univariate full conditional distributions
- a sample from the multivariate distribution can be obtained by repeatedly drawing from each of the univariate distributions

MICE / FCS: Notation

- X: $n \times p$ data matrix with n rows and p variables $x_1, \ldots, x_p$
- R: $n \times p$ missing indicator matrix containing 0 (missing) or 1 (observed)

$$\begin{array}{cc} & \begin{array}{cccc} X_{-2} & X_2 & X_{-2} & \end{array} \\ \mathbf{X} = & \begin{vmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{vmatrix} \end{array}$$

$$\mathbf{R} = \begin{vmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,p} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n,1} & R_{n,2} & \cdots & R_{n,p} \end{vmatrix}$$

MICE / FCS: Notation

- X: $n \times p$ data matrix with n rows and p variables $x_1, \ldots, x_p$
- R: $n \times p$ missing indicator matrix containing 0 (missing) or 1 (observed)

$$\begin{array}{c c c c c}
 & X_{-2} & X_2 & \multicolumn{2}{c}{X_{-2}} \\
\mathbf{X} = & \multicolumn{4}{c}{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}}
\end{array}
\qquad
\mathbf{R} = \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,p} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n,1} & R_{n,2} & \cdots & R_{n,p} \end{bmatrix}$$

For example:

$$\mathbf{X} = \begin{array}{cccc}
X_1 & X_2 & X_3 & X_4 \\
\hline
\checkmark & \text{NA} & \checkmark & \checkmark \\
\checkmark & \checkmark & \text{NA} & \text{NA} \\
\checkmark & \text{NA} & \checkmark & \text{NA}
\end{array}
\qquad
\Rightarrow \mathbf{R} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, p$:                                              ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \ldots, p$:                                         ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t$ in $1, \ldots, T$:                       ▷ loop through iterations
    for $j$ in $1, \ldots, p$:                   ▷ loop through variables



    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, p$:                                        ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t$ in $1, \dots, T$:                          ▷ loop through iterations
    for $j$ in $1, \dots, p$:                      ▷ loop through variables
        Define currently complete data except $X_j$
        $\dot{X}_{-j}^t = \left(\dot{X}_1^t, \dots, \dot{X}_{j-1}^t, \dot{X}_{j+1}^{t-1}, \dots, \dot{X}_p^{t-1}\right)$.



    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, p$:                                         ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t$ in $1, \dots, T$:                                         ▷ loop through iterations
    for $j$ in $1, \dots, p$:                                     ▷ loop through variables
        Define currently complete data except $X_j$
        $\dot{X}_{-j}^t = \left(\dot{X}_1^t, \dots, \dot{X}_{j-1}^t, \dot{X}_{j+1}^{t-1}, \dots, \dot{X}_p^{t-1}\right)$.
        Draw parameters $\dot{\theta}_j^t \sim p(\theta_j^t \mid X_j^{obs}, \dot{X}_{-j}^t, R)$.

    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, p$:                                                ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t$ in $1, \dots, T$:                                ▷ loop through iterations
    for $j$ in $1, \dots, p$:                             ▷ loop through variables
        Define currently complete data except $X_j$
        $\dot{X}_{-j}^t = \left(\dot{X}_1^t, \dots, \dot{X}_{j-1}^t, \dot{X}_{j+1}^{t-1}, \dots, \dot{X}_p^{t-1}\right)$.
        Draw parameters $\dot{\theta}_j^t \sim p(\theta_j^t \mid X_j^{obs}, \dot{X}_{-j}^t, R)$.
        Draw imputations $\dot{X}_j^t \sim p(X_j^{mis} \mid \dot{X}_{-j}^t, R, \dot{\theta}_j^t)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \ldots, 4$:                                        ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 1$:                                      ▷ loop through iterations
    for $j = 1$:                                   ▷ loop through variables
        Define currently complete data except $X_1$
        $\dot{X}_{-1}^{1} = \left(\dot{X}_2^0, \dot{X}_3^0, \dot{X}_4^0\right)$.
        Draw parameters $\dot{\theta}_1^1 \sim p(\theta_1^1 \mid X_1^{obs}, \dot{X}_{-1}^1, R)$.
        Draw imputations $\dot{X}_1^1 \sim p(X_1^{mis} \mid \dot{X}_{-1}^1, R, \dot{\theta}_1^1)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \ldots, 4$:                                                    ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 1$:                                          ▷ loop through iterations
    for $j = 2$:                                       ▷ loop through variables
        Define currently complete data except $X_2$
        $\dot{X}_{-2}^1 = \left(\dot{X}_1^1, \dot{X}_3^0, \dot{X}_4^0\right)$.
        Draw parameters $\dot{\theta}_2^1 \sim p(\theta_2^1 \mid X_2^{obs}, \dot{X}_{-2}^1, R)$.
        Draw imputations $\dot{X}_2^1 \sim p(X_2^{mis} \mid \dot{X}_{-2}^1, R, \dot{\theta}_2^1)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \ldots, 4$:                                        ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 1$:                                   ▷ loop through iterations
    for $j = 3$:                                ▷ loop through variables
        Define currently complete data except $X_3$
        $\dot{X}_{-3}^1 = \left(\dot{X}_1^1, \dot{X}_2^1, \dot{X}_4^0\right)$.
        Draw parameters $\dot{\theta}_3^1 \sim p(\theta_3^1 \mid X_3^{obs}, \dot{X}_{-3}^1, R)$.
        Draw imputations $\dot{X}_3^1 \sim p(X_3^{mis} \mid \dot{X}_{-3}^1, R, \dot{\theta}_3^1)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \ldots, 4$:                                          ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 1$:                                       ▷ loop through iterations
    for $j = 4$:                                   ▷ loop through variables
        Define currently complete data except $X_4$
        $\dot{X}_{-4}^1 = \left(\dot{X}_1^1, \dot{X}_2^1, \dot{X}_3^1\right)$.
        Draw parameters $\dot{\theta}_4^1 \sim p(\theta_4^1 \mid X_4^{obs}, \dot{X}_{-4}^1, R)$.
        Draw imputations $\dot{X}_4^1 \sim p(X_4^{mis} \mid \dot{X}_{-4}^1, R, \dot{\theta}_4^1)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, 4$:                                              ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 2$:                                      ▷ loop through iterations
    for $j = 1$:                                   ▷ loop through variables
        Define currently complete data except $X_1$
        $\dot{X}_{-1}^2 = \left(\dot{X}_2^1, \dot{X}_3^1, \dot{X}_4^1\right)$.
        Draw parameters $\dot{\theta}_1^2 \sim p(\theta_1^2 \mid X_1^{obs}, \dot{X}_{-1}^2, R)$.
        Draw imputations $\dot{X}_1^2 \sim p(X_1^{mis} \mid \dot{X}_{-1}^2, R, \dot{\theta}_1^2)$.
    end for
end for
```

The MICE Algorithm (Van Buuren, 2012)

```
for $j$ in $1, \dots, 4$:                                                    ▷ Setup
    Specify imputation model for variable $X_j$
    $p(X_j^{mis} \mid X_j^{obs}, X_{-j}, R)$
    Fill in starting imputations $\dot{X}_j^0$ by random draws from $X_j^{obs}$.
end for

for $t = 2$:                                              ▷ loop through iterations
    for $j = 2$:                                          ▷ loop through variables
        Define currently complete data except $X_2$
        $\dot{X}^{2}_{-2} = \left(\dot{X}^2_1, \dot{X}^1_3, \dot{X}^1_4\right)$.
        Draw parameters $\dot{\theta}^2_2 \sim p(\theta^2_2 \mid X^{obs}_2, \dot{X}^2_{-2}, R)$.
        Draw imputations $\dot{X}^2_2 \sim p(X^{mis}_2 \mid \dot{X}^2_{-2}, R, \dot{\theta}^2_2)$.
    end for
end for
```

The MICE Algorithm

The imputed values from the **last iteration**,

$$\left(\dot{X}_1^T, \ldots, \dot{X}_p^T\right),$$

are then used to replace the missing values in the original data.

One run through the algorithm ➡ one imputed dataset.

The MICE Algorithm

The imputed values from the **last iteration**,

$$\left(\dot{X}_1^T, \ldots, \dot{X}_p^T\right),$$

are then used to replace the missing values in the original data.

One run through the algorithm ➡ one imputed dataset.

➡ To obtain m imputed datasets: **repeat m times**

Iterations & Convergence

- The **sequence of imputations** for one missing value (from starting value to final iteration) is called a **chain**.
- Each run through the MICE algorithm produces one chain per missing value.

Why iterations?

Iterations & Convergence

- The **sequence of imputations** for one missing value (from starting value to final iteration) is called a **chain**.
- Each run through the MICE algorithm produces one chain per missing value.

Why iterations?

- **Imputed values** in one variable **depend on** the imputed values of the **other variables** (Gibbs sampling).
- If the starting values (random draws) are far from the actual distribution, imputed values from the first few iterations are not draws from the distribution of interest.

Iterations & Convergence

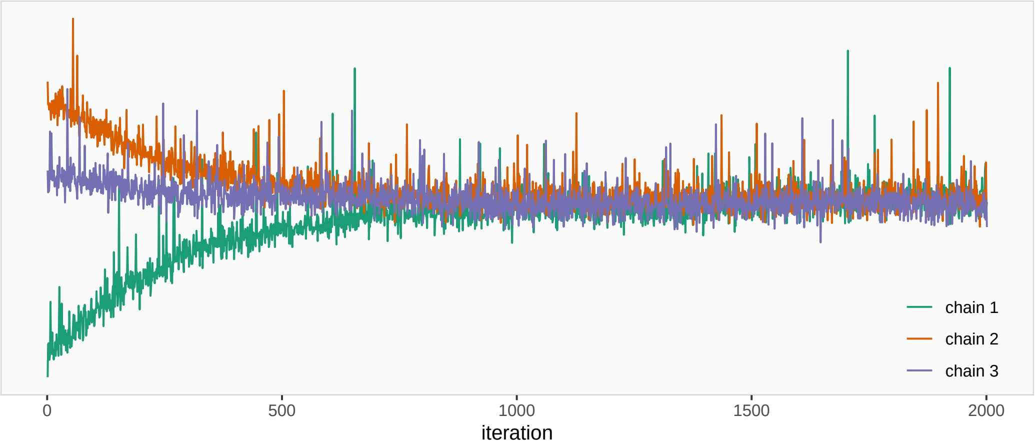
How many iterations?
Until **convergence**
= when the sampling distribution does not change any more
(Note: the imputed value will still vary between iterations.)

Iterations & Convergence

How many iterations?
Until **convergence**
= when the sampling distribution does not change any more
(Note: the imputed value will still vary between iterations.)

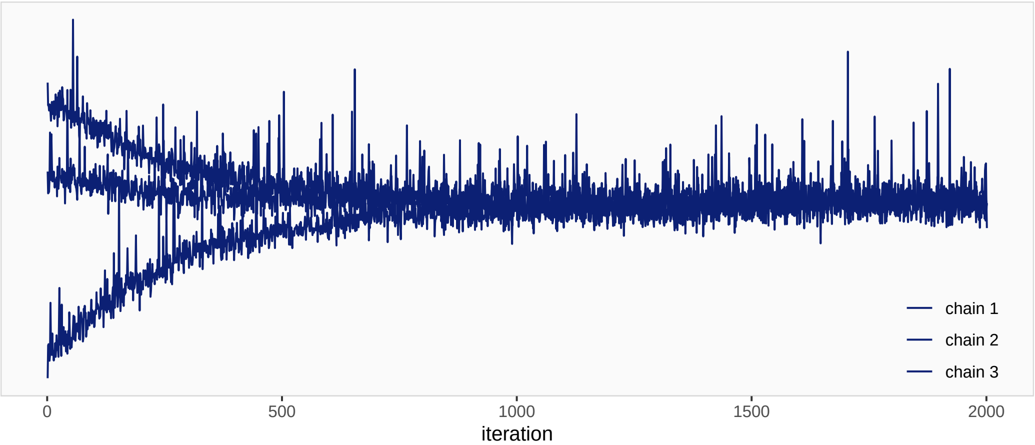
How to evaluate convergence?
The **traceplot** (x-axis: iteration number, y-axis: imputed value) should show a horizontal band

Checking Convergence



Each chain is the sequence of imputed values (from starting value to final imputed value) for the same missing value.

Checking Convergence



Each chain is the sequence of imputed values (from starting value to final imputed value) for the same missing value.

Checking Convergence

In imputation we have

- several **variables** with missing values (e.g., p)
- several missing **values** in each of these variables
- m **chains** for each missing value

➡ possibly a large number of MCMC chains

To check all chains separately could be very time consuming in large datasets.

Checking Convergence

In imputation we have

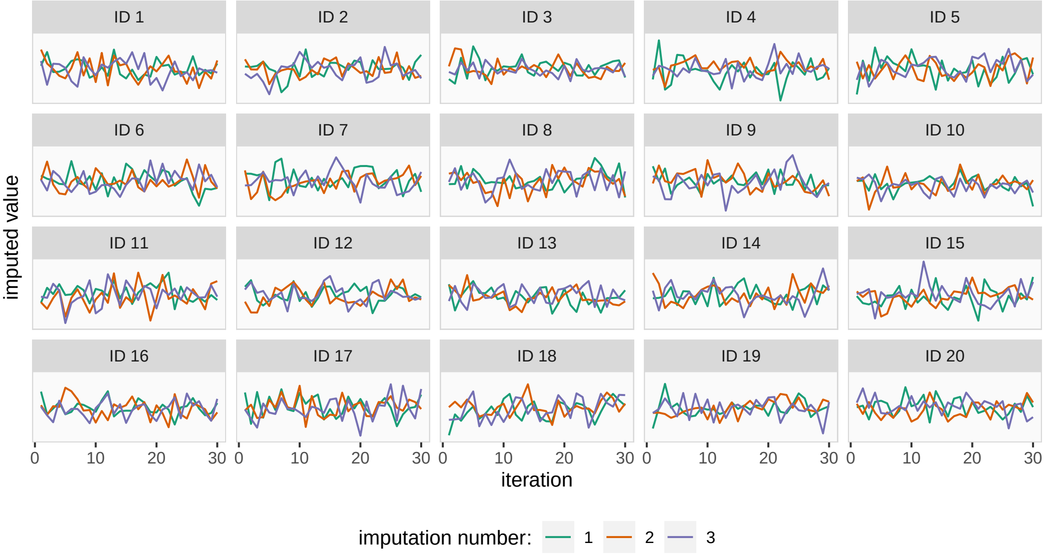
- several **variables** with missing values (e.g., p)
- several missing **values** in each of these variables
- m **chains** for each missing value

➡ possibly a large number of MCMC chains

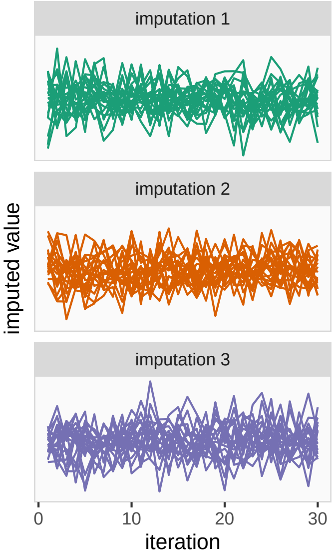
To check all chains separately could be very time consuming in large datasets.

Alternative: Calculate and plot a summary (e.g., the mean) of the imputed values over all subjects, separately per chain and variable
➡ only $m \times p$ chains to check

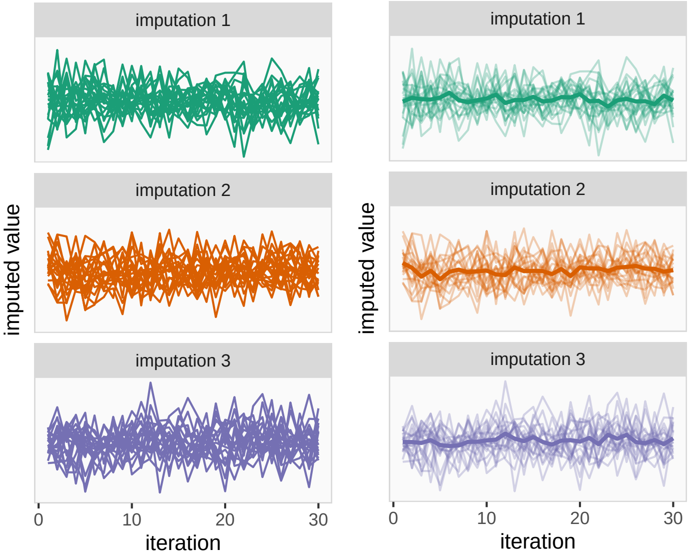
Checking Convergence



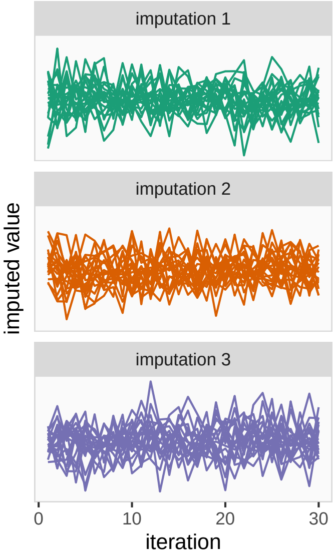
Checking Convergence

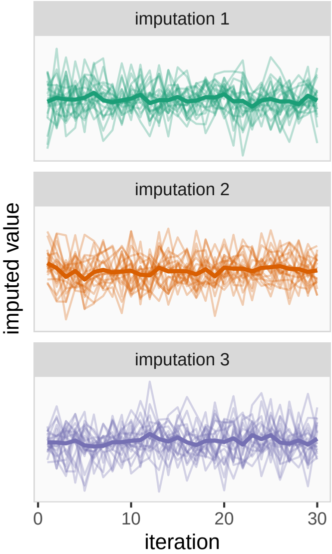


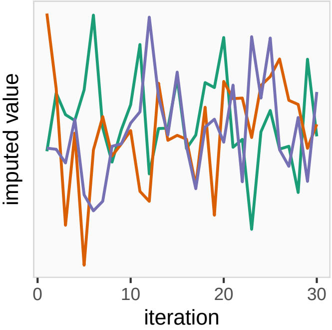
Checking Convergence



Checking Convergence







References

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Taylor & Francis. https://stefvanbuuren.name/fimd/