Clustering with missing data: pooling multiple imputation results with consensus clustering

Lilith Faucheux, Matthieu Resche-Rigon, Emmanuel Curis, Vassili Soumelis, Sylvie Chevret
PhD student at ECSTRRA team, UMR1153, Paris, France
Classification of breast tumors according to their immune Tumor MicroEnvironment

Analysis of soluble proteins produced and identified by immune cells

Issue: Missing data and left-censored data

N = 420 patients
Objective

Evaluate performance of clustering procedures using multiple imputation, on datasets with missing and left-censored data
Objective

Evaluate performance of clustering procedures using multiple imputation, on datasets with missing and left-censored data

Outline

• Simulation framework
• Clustering procedures using MI
• Performances on simulations
Simulation framework
Experimental data

Assays pipeline

Standard curve fitting

Fluorescence (FI) \rightarrow Parameter estimation

\[ FI = F(\text{concentration}) \]

Sample concentration estimation

\[ \hat{F} \rightarrow \text{Fluorescence (FI)} \rightarrow \text{Concentration estimation} \]

\[ \text{Concentration} = \hat{F}^{-1}(FI) \]
Experimental data: Incomplete data

- **Missing data**
- **Left-censored data**
  - Estimated concentration is lower than the lowest standard

- Lack of fluorescence acquisition
Data simulation

Simulation of samples exact concentration

\[ FL = F(\text{concentration}) \times \varepsilon, \quad \varepsilon \sim \text{Log}N(0, 0.1) \]

Exact concentration

\[ F \]

Fluorescence (FI)

Sample concentration estimation

\[ \hat{\text{Concentration}} = \hat{F}^{-1}(FI) \]

Assays pipeline

Curve fitting

\[ FI = F(\text{concentration}) \]

Parameter estimation

\[ \hat{F} \]

Fluorescence (FI)

Simulated data

C1

C2

C7

C?
Data simulation

Simulation of samples exact concentration

\[ FI = F(\text{concentration}) \times \varepsilon, \quad \varepsilon \sim \text{Log} \mathcal{N}(0, 0.1) \]

Exact concentration

\[ F \]

Fluorescence (FI)

3 variables (X1, X2, X3)
500 observations

Assays pipeline

Curve fitting

\[ FI = F(\text{concentration}) \]

C1, C2, ..., C7

Fluorescence (FI)

Parameter estimation

Sample concentration estimation

\[ \text{Concentration} = \hat{F}^{-1}(FI) \]

\[ \hat{F} \]

Fluorescence (FI)

Concentration estimation
Data simulation

**Simulation of samples exact concentration**

\[ FI = F(\text{concentration}) \times \varepsilon, \quad \varepsilon \sim \text{LogN}(0, 0.1) \]

Exact concentration

3 variables (X1, X2, X3)
500 observations
3 groups

\[ \Sigma = I(\sigma^2) \]

**Curve fitting**

\[ FI = F(\text{concentration}) \]

Parameter estimation

**Sample concentration estimation**

\[ \text{Concentration} = \hat{F}^{-1}(FI) \]

Fluorescence (Fl)

Concentration estimation
Data simulation

Simulation of samples exact concentration

\[ FI = F(\text{concentration}) \times \varepsilon, \quad \varepsilon \sim \log N(0, 0.1) \]

Sample concentration estimation

3 variables (X1, X2, X3)
500 observations
3 groups
30% of missing data on X1

MCAR
MAR (X2)
MNAR (X1)

Fluorescence (FI)

Curve fitting

\[ FI = F(\text{concentration}) \]

Parameter estimation

Fluorescence (FI)

Assays pipeline
Data simulation

Simulation of samples exact concentration

\[ FI = F(\text{concentration}) \times \varepsilon, \quad \varepsilon \sim \text{LogN}(0, 0.1) \]

Exact concentration

Fluorescence (FI)

3 variables (X1, X2, X3)
500 observations
3 groups
30% of missing data on X1
Left-censored data

Sample concentration estimation

\[ \text{Concentration} = \hat{F}^{-1}(FI) \]

Fluorescence (FI)

Parameter estimation

Curve fitting

\[ FI = F(\text{concentration}) \]

C1
C2
\ldots
C7

Fluorescence (FI)
Data simulation: 3 scenarios

Scenario 1: Large separation
Scenario 2: Intermediate separation
Scenario 3: Diagonal separation

Missing data mechanism
Proportion of left-censored data
Outline

• Simulation framework

• Clustering procedures using MI

• Performances on simulations
Clustering with multiple imputation

Independent analyses

Regression

- $a_1$
- $a_2$
- $a_3$

Combine results

Rubin’s rules

- $a$

Partition Generation

Consensus clustering
Consensus clustering

**Principle:** Combining multiple clustering results to reveal consistency.

Consensus clustering

Principle:
Combining multiple clustering results to reveal consistency.
Consensus clustering

Principle:
Combining multiple clustering results to reveal consistency.

Consensus clustering

Consensus clustering for MI

K - MEANS

Consensus function
Clustering with incomplete data using MI

- Imputation of missing and left-censored data
- K selection & consensus
- K means
Clustering with incomplete data using MI

Methods for missing data
- deletion
- SSI (mice)
- MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
- SI: standard curve
- SSI for left-censored data\(^1\) (mice)
- MI for left-censored data\(^1\) (mice)

Criteria for K selection
- \(\text{CH}^2\)
- \(\text{CritCF}^3\)

Consensus methods
- None
- Multicons\(^4\)
- Combinatorial optimisation (Bruckers\(^5\))
- Object co-occurrence (Basagaña\(^6\))

References:

MI = 5 imputations
Clustering with incomplete data using MI

Imputation of missing and left-censored data

K selection & consensus

Methods for missing data
- deletion
- SSI (mice)
- MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
- SI: standard curve
- SSI for left-censored data¹ (mice)
- MI for left-censored data¹ (mice)

Criteria for K selection
- \( CH^2 \)
- \( CritCF^3 \)

Consensus methods
- None
- Multicons⁴
- Combinatorial optimisation (Bruckers⁵)
- Object co-occurrence (Basagaña⁶)

\[
CH = \frac{BSS}{WSS} \cdot \frac{n - k}{k - 1}
\]

\[
CritCF = \left( \frac{2p}{p + 1} \cdot \frac{1}{1 + \frac{W}{B}} \right)^{log_2(k+1)+1}
\]

MI = 5 imputations
Clustering with incomplete data using MI

Methods for missing data
- deletion
- SSI (mice)
- MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
- SI: standard curve
- SSI for left-censored data\(^1\) (mice)
- MI for left-censored data\(^1\) (mice)

Criteria for K selection
- CH\(^2\)
- CritCF\(^3\)

Consensus methods
- None
- Multicons\(^4\)
- Combinatorial optimisation (Bruckers\(^5\))
- Object co-occurrence (Basagaña\(^6\))


MI = 5 imputations
Example of proposed method

Methods for missing data
- deletion
- SSI (mice)
- MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
- SI: standard curve
- SSI for left-censored data (mice)
- MI for left-censored data (mice)

Criteria for K selection
- CH
- CritCF

Consensus methods
- None
- Multicons
  - Combinatorial optimisation (Bruckers)
  - Object co-occurrence (Basagaña)

MI Multicons CH
Stochastic simple imputation

Imputation of missing and left-censored data

K selection & consensus

Methods for missing data
- deletion
- SSI (mice)
- MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
- SI: standard curve
- SSI for left-censored data (mice)
- MI for left-censored data (mice)

Criteria for K selection
- CH
- CritCF

Consensus methods
- None
- Multicons
- Combinatorial optimisation (Bruckers)
- Object co-occurrence (Basagaña)

SSI CH
Complete cases analysis

- Imputation of missing and left-censored data
- K selection & K means
- Consensus

Methods for missing data
- Deletion
  - SSI (mice)
  - MI (mice)

Methods for left-censored data
- SI: 1/2 LOD
  - SI: standard curve
  - SSI for left-censored data (mice)
  - MI for left-censored data (mice)

Criteria for K selection
- CH
- CritCF

Consensus methods
- None
- Multicons
- Combinatorial optimisation (Bruckers)
- Object co-occurrence (Basagaña)

CCA CH
Outline

• Simulation framework

• Clustering procedures using MI

• Performances on simulations
How performance of clustering was evaluated

Adjusted Rand Index (ARI)
frequency of agreement over pairs, adjusted for the chance of grouping elements

External validation criterion: ARI

$$RI(P_1, P_{ref}) = \frac{\text{Number of concording pairs}}{\text{Number of pairs}}$$

$$ARI(P_1, P_{ref}) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$

Concording pair AB
Non concording pair AB

1000 simulations
Performances on complete data (all scenarios)

Correct number of clusters : 98.7%

\[ \hat{k} = 4 : 3.6\% \quad \hat{k} = 4 : 0.1\% \quad \hat{k} = 2 : 0.3\% \]
How performance of clustering was also evaluated

Adjusted Rand Index (ARI) \[\Delta \text{ARI}\]
Left-censored data & no missing data

ARI $- ARI_{Complete Data}$
Left-censored data & no missing data

$\Delta_{\text{Complete Data}}$

Scenario 3: poor performances with CritCF

$\hat{k} = 2: 2\%$  
$\hat{k} \geq 4: 4\%$  

$\hat{k} = 2: 79\%$  
$\hat{k} \geq 4: 7\%$  

$\hat{k} = 2: 94\%$  
$\hat{k} = 2: 99\%$
Left-censored data & no missing data

\[ ARI \quad - \quad ARI_{Complete\ Data} \]

Scenario 1: same performances of all methods
Scenario 2: Best results with calibration curve imputation, same performances of other methods

![Box plots comparing different scenarios across methods](image-url)
Left-censored data & no missing data

\[ ARI - ARI_{Complete\ Data} \] 

Scenario 3: best performances of standard curve imputation & MI with Multicons
Left-censored data & no missing data

Scenario 3: best performances of standard curve imputation & MI with Multicons
Missing data & no Left-censored data

$ARI - ARI_{Complete Data}$
Missing data & no Left-censored data

$ARI - ARI_{\text{Complete Data}}$

**Scenario 3: Poor performances with CritCF**

<table>
<thead>
<tr>
<th>Method</th>
<th>MCAR</th>
<th>MAR</th>
<th>MNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA.ch</td>
<td><img src="image1" alt="Boxplot" /></td>
<td><img src="image2" alt="Boxplot" /></td>
<td><img src="image3" alt="Boxplot" /></td>
</tr>
<tr>
<td>ssi.ch</td>
<td><img src="image4" alt="Boxplot" /></td>
<td><img src="image5" alt="Boxplot" /></td>
<td><img src="image6" alt="Boxplot" /></td>
</tr>
<tr>
<td>ml.MultiCons.ch</td>
<td><img src="image7" alt="Boxplot" /></td>
<td><img src="image8" alt="Boxplot" /></td>
<td><img src="image9" alt="Boxplot" /></td>
</tr>
<tr>
<td>ml.Brouckers.ch</td>
<td><img src="image10" alt="Boxplot" /></td>
<td><img src="image11" alt="Boxplot" /></td>
<td><img src="image12" alt="Boxplot" /></td>
</tr>
<tr>
<td>ml.MultiCons.CritCF</td>
<td><img src="image13" alt="Boxplot" /></td>
<td><img src="image14" alt="Boxplot" /></td>
<td><img src="image15" alt="Boxplot" /></td>
</tr>
<tr>
<td>ml.Brouckers.CritCF</td>
<td><img src="image16" alt="Boxplot" /></td>
<td><img src="image17" alt="Boxplot" /></td>
<td><img src="image18" alt="Boxplot" /></td>
</tr>
<tr>
<td>ml.Basagana.CritCF</td>
<td><img src="image19" alt="Boxplot" /></td>
<td><img src="image20" alt="Boxplot" /></td>
<td><img src="image21" alt="Boxplot" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completeness</th>
<th>MCAR</th>
<th>MAR</th>
<th>MNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Data</td>
<td>32%</td>
<td>15%</td>
<td>5%</td>
</tr>
<tr>
<td>MAR</td>
<td>26%</td>
<td>14%</td>
<td>4%</td>
</tr>
<tr>
<td>MNAR</td>
<td>20%</td>
<td>17%</td>
<td>4%</td>
</tr>
</tbody>
</table>

$k = 3$
Missing data & no Left-censored data

$ARI - ARI_{Complete Data}$

Scenario 1: under-performance of CCA with MAR mechanism

$k = 3$

$73$

$k = 4$

$27$

$\hat{k} = 3$

$\hat{k} = 4$
Missing data & no Left-censored data

$ARI - ARI_{Complete \ Data}$

Scenario 2 & 3: Better performances under MAR than MCAR or MNAR
Scenario 2 & 3: Better performances under MAR than MCAR or MNAR

Missing data & no Left-censored data

$ARI - ARI_{Complete\ Data}$
Missing data & no Left-censored data

$ARI - ARI_{Complete\ Data}$

Scenario 2 & 3: Best performances with CCA & MI with Multicons
Missing data & Left-censored data

\[ ARI - ARI_{Complete\ Data} \]
### Missing data & Left-censored data

**ARI – ARI\_Complete Data**

**Sc. 1**: no difference;

**Sc. 2 & 3**: MAR >> MCAR & MNAR, Multicons >> Bruckers & Basagana;

**Sc. 3**: CH >> CritCF

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCAR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MAR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MNAR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Missing data & Left-censored data

$ARI - ARI_{Complete\ Data}$

Scenario 2: Same performances of $\frac{1}{2}$ LOD, Standard curve and MI
Missing data & Left-censored data

$ARI - ARI_{\text{Complete Data}}$

Scenario 2: Same performances of $\frac{1}{2}$ LOD, Standard curve and MI
Missing data & Left-censored data

\( ARI - ARI_{Complete\ Data} \)

Scenario 3: More variability in MI performances compared to SI
Missing data & Left-censored data

$ARI - ARI_{Complete Data}$

Scenario 3: More variability in MI performances compared to SI
Results

Left-censored data & no missing data
- CH >> CritfCF
- MI with Multicons & standard curve >> other methods

No Left-censored data & missing data
- CH >> CritfCF (∀ missing mechanism)
- MI with Multicons & CCA >> SSI, Bruckers consensus, Basagana consensus
- MAR >> MCAR & MNAR

Left-censored data & missing data
- CH >> CritfCF (∀ missing mechanism)
- MI with Multicons & CCA >> SSI, Bruckers consensus, Basagana consensus
- ½ LOD & standard curve > MI for censored data (more variability)
- MAR >> MCAR & MNAR

Recommended method: Multicons
with imputation by standard curve for left censored data
and CH criterion for K selection
Results extendable to any dataset with missing data

Understand group structure distortion by incomplete data
  MAR distortion seems to cause less difficulty than MNAR and MCAR

Performances compared to likelihood based methods