On the consistency of supervised learning with missing values

Julie Josse, Nicolas Prost, Erwan Scornet, Gael Varoquaux
CMAP, INRIA Parietal
15 March 2019
https://hal.archives-ouvertes.fr/hal-02024202
Research activities

- Dimensionality reduction methods to visualize complex data (PCA based): multi-sources, textual, arrays, questionnaire
- Low rank estimation, selection of regularization parameters
- Missing values - matrix completion
- Causal inference
- Fields of application: bio-sciences (agronomy, sensory analysis), health data (hospital data)
- R community: book R for Stat, R foundation, taskforce, packages: 
  - FactoMineR explore continuous, categorical, multiple contingency tables (correspondence analysis), combine clustering and PC, ..
  - MissMDA for single and multiple imputation, PCA with missing
  - denoiseR to denoise data with low-rank estimation
  - R-miss-tastic missing values plateform
1. Introduction

2. Handling missing values

3. Supervised learning with missing values

4. Decision trees

5. Discussion
Introduction
Collaborators

PhD students : G. Robin, W. Jiang, I. Mayer, N. Prost, polytechnique
Colleagues : J-P Nadal (EHESS), E. Scornet (X), G. Varoquaux (INRIA), S. Wager (Stanford), B. Naras (Stanford)
Traumabase (hospital) : T. Gauss, S. Hamada, J-D Moyer
Capgemini
### Traumabase

15000 patients/ 250 variables/ 11 hospitals, from 2011 (4000 new patients/ year)

<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>Fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>Lille</td>
<td>Other</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>130</td>
<td>62</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>Gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>62</td>
</tr>
<tr>
<td>Beaujon</td>
<td>AVP moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>145</td>
<td>89</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP bicycle</td>
<td>33</td>
<td>m</td>
<td>75</td>
<td>NR</td>
<td>NR</td>
<td>104</td>
<td>86</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP pedestrian</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>66</td>
</tr>
<tr>
<td>HEGP</td>
<td>White weapon</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>1.92</td>
<td>26.58</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>Toulon</td>
<td>White weapon</td>
<td>20</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>124</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpO2</th>
<th>Temperature</th>
<th>Lactates</th>
<th>Hb</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>35.6</td>
<td>&lt;NA&gt;</td>
<td>12.7</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>36.5</td>
<td>4.8</td>
<td>11.1</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>36</td>
<td>3.9</td>
<td>11.4</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>36.7</td>
<td>1.66</td>
<td>13</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>36</td>
<td>NM</td>
<td>14.4</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>36.6</td>
<td>NM</td>
<td>14.3</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>37.5</td>
<td>13</td>
<td>15.9</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>36.9</td>
<td>NM</td>
<td>13.7</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ **Estimate causal effect** : administration of the treatment "tranexamic acid" (within the first 3 hours after the accident) on mortality (outcome) for traumatic brain injury (TBI) patients.
15000 patients/ 250 variables/ 11 hospitals, from 2011 (4000 new patients/ year)

<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>Fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>Lille</td>
<td>Other</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>130</td>
<td>62</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>Gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>62</td>
</tr>
<tr>
<td>Beaujon</td>
<td>AVP moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>145</td>
<td>89</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP bicycle</td>
<td>33</td>
<td>m</td>
<td>75</td>
<td>NR</td>
<td>NR</td>
<td>104</td>
<td>86</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP pedestrian</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>66</td>
</tr>
<tr>
<td>HEGP</td>
<td>White weapon</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>1.92</td>
<td>26.58</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>Toulon</td>
<td>White weapon</td>
<td>20</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>124</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpO2</th>
<th>Temperature</th>
<th>Lactates</th>
<th>Hb</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>35.6</td>
<td>&lt;NA&gt;</td>
<td>12.7</td>
<td>12</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.5</td>
<td>4.8</td>
<td>11.1</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>3.9</td>
<td>11.4</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.7</td>
<td>1.66</td>
<td>13</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>NM</td>
<td>14.4</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.6</td>
<td>NM</td>
<td>14.3</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>37.5</td>
<td>13</td>
<td>15.9</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.9</td>
<td>NM</td>
<td>13.7</td>
<td>15</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ **Predict** whether to start a blood transfusion, to administer fresh frozen plasma, etc...

⇒ (Logistic) regressions with **missing** categorical/continuous values
Missing values

Percentage of missing values

variable
null.data
na.data
nr.data
nf.data
imp.data

Percentage
0
25
50
75
100
Handling missing values
Missing values

are everywhere: unanswered questions in a survey, lost data, damaged plants, machines that fail...

*The best thing to do with missing values is not to have any*” Gertrude Mary Cox.

⇒ Still an issue with "big data"

Data integration: data from different sources

Multilevel data: sporadically - systematic (one variable missing in one hospital)
Solutions to handle missing values


⇒ Modify the estimation process to deal with missing values. Maximum likelihood : **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) ; Louis for their variability
    Difficult to establish ?
    Not many implementations, even for simple models
    One specific algorithm for each statistical method...

⇒ **Imputation** (multiple) to get a completed data set on which you can perform any statistical method (Rubin, 1976)

⇒ Aim : estimate parameters and their variance from an incomplete data
Dealing with missing values

⇒ Imputation to get a completed data set

$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

$\hat{\mu}_y = 0.01$
$\hat{\sigma}_y = 0.5$
$\hat{\rho} = 0.30$
Wright IJ, et al. (2004). The worldwide leaf economics spectrum. Nature, 69 000 species - LMA (leaf mass per area), LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), Rmass (dark respiration rate)
Imputation methods

\[ \mu_y = 0 \]

\[ \sigma_y = 1 \]

\[ \rho = 0.6 \]
Imputation methods

- Impute by regression take into account the relationship: estimate $\beta$ - impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow$ variance underestimated and correlation overestimated.

\[
\begin{array}{l}
\mu_y = 0 & 0.01 \\
\sigma_y = 1 & 0.5 \\
\rho = 0.6 & 0.30 \\
\end{array}
\]

\[
\begin{array}{l}
\mu_y = 0 & 0.01 \\
\sigma_y = 1 & 0.72 \\
\rho = 0.6 & 0.78 \\
\end{array}
\]
Imputation methods

- Impute by regression take into account the relationship: estimate $\beta$ - impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow$ variance underestimated and correlation overestimated.
- Impute by stochastic reg: estimate $\beta$ and $\sigma$ - impute from the predictive $y_i \sim \mathcal{N}(x_i \hat{\beta}, \hat{\sigma}^2) \Rightarrow$ preserve distribution

| $\mu_y$ | 0.01 | 0.01 | 0.01 |
| $\sigma_y$ | 0.5 | 0.72 | 0.99 |
| $\rho$ | 0.30 | 0.78 | 0.59 |
Imputation methods: joint modeling

Assuming a
- a gaussian distribution: \( x_i \sim \mathcal{N}(\mu, \Sigma) \) packages Amelia
- low rank: \( X_{n \times d} = \mu_{n \times d} + \varepsilon \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \) with \( \mu \) of low rank \( k \)

\[ \Rightarrow \text{Iterative (shrinked) SVD algo arg min}_{\mu} \left\{ \| W \odot (X - \mu) \|_2^2 + \lambda \| \mu \|_* \right\} \]

packages softimpute (Hastie, Mazumber), missMDA.

Model makes sense (Udell & Townsend)

Mixed data: iterative FAMD

Multilevel mixed imputation: hospital effect with patient nested in hospital. (J., Husson, Robin & Balasu., 2018, Imputation of mixed data with multilevel SVD. *JCGS*)
### Random forests versus PCA

<table>
<thead>
<tr>
<th>Feat1</th>
<th>Feat2</th>
<th>Feat3</th>
<th>Feat4</th>
<th>Feat5...</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C7</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C9</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C10</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C11</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C12</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C13</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>C14</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Igor</td>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td>8</td>
</tr>
<tr>
<td>Frank</td>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td>8</td>
</tr>
<tr>
<td>Bertrand</td>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td>9</td>
</tr>
<tr>
<td>Alex</td>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td>9</td>
</tr>
<tr>
<td>Yohann</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>10</td>
</tr>
<tr>
<td>Jean</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feat1</th>
<th>Feat2</th>
<th>Feat3</th>
<th>Feat4</th>
<th>Feat5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>2.00</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>2.00</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>3.00</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>3.00</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>4.00</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>4.00</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>5.00</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>5.00</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td>6.00</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td>6.00</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7.0</td>
<td>7.00</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7.0</td>
<td>7.00</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6.87</td>
<td>6.87</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>6.87</td>
<td>6.87</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>6.87</td>
<td>6.87</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>6.87</td>
<td>6.87</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>6.87</td>
<td>6.87</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>6.87</td>
<td>6.87</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

⇒ Missing
⇒ Random forests
⇒ PCA

⇒ Imputation inherits from the method : RF (computationally costly)
good for non linear relationship / PCA linear relation

⇒ Aim is not to impute as well as possible but estimate parameters and their variance (multiple imputation).
Logistic regression with missing covariates: parameter estimation, model selection and prediction. (Jiang, J., Lavielle, Gauss, Hamada, 2018)

\( x = (x_{ij}) \) a \( n \times d \) matrix of quantitative covariates
\( y = (y_i) \) an \( n \)-vector of binary responses \( \{0, 1\} \)

**Logistic regression model**

\[
\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}
\]

**Covariables**

\[ x_i \sim \mathcal{N}_p(\mu, \Sigma) \]

**Log-likelihood** for complete-data with \( \theta = (\mu, \Sigma, \beta) \)

\[
\mathcal{L}\mathcal{L}(\theta; x, y) = \sum_{i=1}^{n} \left( \log(p(y_i | x_i; \beta)) + \log(p(x_i; \mu, \Sigma)) \right).
\]

**Decomposition**: \( x = (x_{\text{obs}}, x_{\text{mis}}) \)

Under MAR, possibility to ignore the missing value mechanism

**Observed likelihood** \( \arg \max \mathcal{L}\mathcal{L}(\theta; x_{\text{obs}}, y) = \int \mathcal{L}\mathcal{L}(\theta; x, y) dx_{\text{mis}} \)
Stochastic Approximation EM

• **E-step**: Evaluate the quantity

\[ Q_k(\theta) = \mathbb{E}[\mathcal{L}(\theta; x, y)|x_{obs}, y; \theta_{k-1}] \]

\[ = \int \mathcal{L}(\theta; x, y)p(x_{mis}|x_{obs}, y; \theta_{k-1}) \, dx_{mis} \]

• **M-step**: \( \theta_k = \arg \max_\theta Q_k(\theta) \)

⇒ *Unfeasible computation of expectation*

MCEM ([Wei & Tanner, 1990]) : generate samples of missing data from 
\( p(x_{mis}|x_{obs}, y; \theta_{k-1}) \) and replaces the expectation by an empirical mean.

⇒ *Require a huge number of samples*

SAEM ([Lavielle, 2014]) almost sure convergence to MLE. (Metropolis Hasting - Variance estimation with Louis).

Unbiased estimates : \( \hat{\beta}_1, \ldots, \hat{\beta}_d \) - \( \hat{V}(\hat{\beta}_1), \ldots, \hat{V}(\hat{\beta}_d) \) - good coverage
Supervised learning with missing values
Supervised learning

- A feature matrix $X$ and a response vector $Y$
- Find a prediction function that minimizes the expected risk.
  Bayes rule: $f^* \in \arg \min_{f: \mathcal{X} \to \mathcal{Y}} \mathbb{E}[\ell(f(X), Y)] \Rightarrow f^*(X) = \mathbb{E}[Y|X]$
- Empirical risk minimization:
  $$\hat{f}_{D_{n,\text{train}}} \in \arg \min_{f: \mathcal{X} \to \mathcal{Y}} \left( \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right)$$

A new data $D_{n,\text{test}}$ to estimate the generalization error rate
- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(X), Y)] \xrightarrow{n \to \infty} \mathbb{E}[\ell(f^*(X), Y)]$
Supervised learning

- A feature matrix $X$ and a response vector $Y$
- Find a prediction function that minimizes the expected risk.
  
  Bayes rule: $f^* \in \arg\min_{f: \mathcal{X} \rightarrow \mathcal{Y}} \mathbb{E}[\ell(f(X), Y)]$  
  $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$
- Empirical risk minimization:

  $$
  \hat{f}_{D_n,\text{train}} \in \arg\min_{f: \mathcal{X} \rightarrow \mathcal{Y}} \left( \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right)
  $$

  A new data $D_{n,\text{test}}$ to estimate the generalization error rate
- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(X), Y)] \xrightarrow{n \rightarrow \infty} \mathbb{E}[\ell(f^*(X), Y)]$

**Difference with classical literature**

- response variable $Y$ - Aim: Prediction
- two data sets (out of sample) with missing values: train & test sets

$\Rightarrow$ Is it possible to use previous approaches (EM - impute), consistent?
$\Rightarrow$ Do we need to design new ones?
EM and out-of sample prediction

\[ P(y_i = 1|x_i; \beta) = \frac{\exp(\sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{d} \beta_j x_{ij})} \]

After EM: \( \hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_d, \hat{\mu}, \hat{\Sigma}) \)

New obs: \( x_{n+1} = (x_{(n+1)1}, NA, NA, x_{(n+1)4}, \ldots, x_{(n+1)d}) \)

Predict \( Y \) on a test set with missing entries \( x = (x_{obs}, x_{miss}) \)
EM and out-of sample prediction

\[
P(y_i = 1|x_i; \beta) = \frac{\exp(\sum_{j=1}^{d} \beta_j x_{ij})}{1+\exp(\sum_{j=1}^{d} \beta_j x_{ij})}
\]

After EM: \(\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_d, \hat{\mu}, \hat{\Sigma})\)

New obs: \(x_{n+1} = (x_{(n+1)1}, NA, NA, x_{(n+1)4}, \ldots, x_{(n+1)d})\)

Predict \(Y\) on a test set with missing entries \(x = (x_{obs}, x_{miss})\)

\[
\hat{y} = \arg \max_y p_{\hat{\theta}}(y|x_{obs})
\]

\[
= \arg \max_y \int p_{\hat{\theta}}(y|x)p_{\hat{\theta}}(x_{mis}|x_{obs})dx_{mis}
\]

\[
= \arg \max_y \mathbb{E}_{p_{x_m|x_o=x_o}} p_{\hat{\theta}}(y|X_m, x_o) \approx \arg \max_y \sum_{m=1}^{M} p_{\hat{\theta}} \left( y|x_{obs}, x_{mis}^{(m)} \right).
\]

Logistic regression

\[
\hat{y}^1 \quad \hat{y}^M \quad \hat{y} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}^m
\]
Prediction on test incomplete data with a full data model

- Let a Bayes-consistent predictor $f$ for complete data: $f(X) = \mathbb{E}[Y|X]$
- Note the data: $\tilde{X} = X \odot (1 - M) + NA \odot M$ (takes value in $\mathbb{R} \cup \{NA\}$)
- Perform multiple imputation:

$$f^*_{\text{mult imput}}(\tilde{x}) = \mathbb{E}_{X_m|X_o = x_o}[f(X_m, x_o)]$$

same as out-of sample EM but assuming know $f$

**Theorem**

Consider the regression model $Y = f(X) + \varepsilon$, where

- we assume MAR $\forall S \subset \{1, \ldots, d\}, (M_j)_{j \in S} \perp (X_j)_{j \in S} \mid (X_k)_{k \in S^c}$
- $\varepsilon \perp (M_1, X_1, \ldots, M_d, X_d)$ is a centred noise

Then multiple imputation is consistent:

$$f^*_{\text{mult imput}}(\tilde{x}) = \mathbb{E}[Y|\tilde{X} = \tilde{x}]$$
Proof

Let \( \tilde{x} \in (\mathbb{R} \cup \text{NA})^d \). Without loss of generality, assume only \( \tilde{x}_1, \ldots, \tilde{x}_j \) are NA.

\[
f_{\text{mult imput}}^*(\tilde{x}) = \mathbb{E}_{X_m|X_o=x_o}[f(X_m, X_o = x_o)] \\
= \mathbb{E}[f(X_m, X_o = x_o)|X_o = x_o] \\
= \mathbb{E}[Y|X_o = x_o] \\
= \mathbb{E}[Y|\tilde{X}_{j+1} = \tilde{x}_{j+1}, \ldots, \tilde{X}_d = \tilde{x}_d]
\]

\[
\mathbb{E}[Y|\tilde{X} = \tilde{x}] = \mathbb{E}[Y|\tilde{X}_1 = \text{NA}, \ldots, \tilde{X}_j = \text{NA}, \tilde{X}_{j+1} = \tilde{x}_{j+1}, \ldots, \tilde{X}_d = \tilde{x}_d] \\
= \mathbb{E}[Y|M_1 = 1, \ldots, M_j = 1, \tilde{X}_{j+1} = \tilde{x}_{j+1}, \ldots, \tilde{X}_d = \tilde{x}_d] \\
= \mathbb{E}[Y|\tilde{X}_{j+1} = \tilde{x}_{j+1}, \ldots, \tilde{X}_d = \tilde{x}_d]
\]
Imputation prior to learning

Impute the train, learn a model with $\hat{X}_{\text{train}}, Y_{\text{train}}$. Impute the test with the same imputation and predict with $\hat{X}_{\text{test}}$ and $\hat{f}_{\text{train}}$. 

Same imputation
## Imputation prior to learning

### Imputation with the same model
Easy to implement for univariate imputation: the means \((\hat{\mu}_1, ..., \hat{\mu}_d)\) of each column of the train. Also OK for Gaussian imputation.
Issue: many methods are "black-boxes" and take as an input the incomplete data and output the completed data (mice, missForest)

### Separate imputation
Impute train and test separately (with a different model)
Issue: depends on the size of the test set? one observation?

### Group imputation
Impute train and test simultaneously but the predictive model is learned only on the training imputed data set
Issue: sometimes not the training set
Imputation with the same model: mean imputation is consistent

Learn on the mean-imputed training data, impute the test set with the same means and predict is optimal if the missing data are MAR and the learning algorithm is universally consistent.

Framework - assumptions

- $Y = f(X) + \varepsilon$
- $X = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0, 1]^d$
- $\|f\|_{\infty} < \infty$
- Missing data on $X_1$ with $M_1 \perp X_1 | X_2, \ldots, X_d$.
- $(x_2, \ldots, x_d) \mapsto \mathbb{P}[M_1 = 1 | X_2 = x_2, \ldots, X_d = x_d]$ is continuous
- $\varepsilon$ is a centered noise independent of $(X, M_1)$
- imputed entry $x' = (x'_1, x_2, \ldots, x_d)$ where $x'_1 = x_1 \mathbb{1}_{M_1 = 0} + \mathbb{E}[X_1] \mathbb{1}_{M_1 = 1}$
- $\tilde{X} = X'$ if $X'_1 \neq \mathbb{E}[X_1]$ and $\tilde{X} = (\text{NA}, X_2, \ldots, X_d)$ if $X'_1 = \mathbb{E}[X_1]$
Imputation with the same model: mean imputation is consistent

Learn on the mean-imputed training data, impute the test set with the **same** means and predict is optimal if the missing data are MAR and the learning algorithm is universally consistent.

**Theorem**

\[
\begin{align*}
f_{\text{impute}}^*(x') &= \mathbb{E}[Y|X_2 = x_2, \ldots, X_d = x_d, M_1 = 1] \\
&= \mathbb{1}_{x_1' = \mathbb{E}[X_1]} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]>0} \\
&\quad + \mathbb{E}[Y|X = x'] \mathbb{1}_{x_1' = \mathbb{E}[X_1]} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]=0} \\
&\quad + \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d, M_1 = 0] \mathbb{1}_{x_1' \neq \mathbb{E}[X_1]}.
\end{align*}
\]

Prediction with mean is equal to the Bayes function almost everywhere

\[
f_{\text{impute}}^*(x') = \tilde{f}^*(\tilde{X}) = \mathbb{E}[Y|\tilde{X} = \tilde{x}]
\]

(remains valid when missing values occur for variables \(X_1, \ldots, X_j\))
Learn on the mean-imputed training data, impute the test set with the same means and predict is optimal if the missing data are MAR and the learning algorithm is universally consistent

**Rationale**

The learning algorithm learns the imputed value (here the mean) and use that information to detect that the entry was initially missing. If the imputed value changes from train to test set the learning algorithm may fail, since imputed data distribution differs between train and test sets.

⇒ Other values than the mean are possible. Mean not a bad choice despite its drawbacks.
Trees
End-to-end learning with missing values

\[ \hat{Y}_{\text{test}} \rightarrow \text{Decision trees} \]
CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg \min_{(j, z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y | X_j \leq z])^2 \cdot 1_{X_j \leq z} ight. 
+ \left. (Y - \mathbb{E}[Y | X_j > z])^2 \cdot 1_{X_j > z} \right].$$
CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg \min_{(j, z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z])^2 \cdot 1_{X_j \leq z} + (Y - \mathbb{E}[Y|X_j > z])^2 \cdot 1_{X_j > z} \right].$$
Built recursively by splitting the current cell into two children: find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg \min_{(j,z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z])^2 \cdot 1_{X_j \leq z} \right. \left. + (Y - \mathbb{E}[Y|X_j > z])^2 \cdot 1_{X_j > z} \right].$$
CART with missing values: split on available cases

\[
X_1 \quad X_2
\]

\[
\begin{array}{c|c|c}
\hline
& \text{X}_1 \leq \text{s}_1 & \text{X}_1 > \text{s}_1 \\
\hline
\text{X}_2 \leq \text{s}_2 & \text{left} & \text{right} \\
\hline
\text{X}_2 > \text{s}_2 & & \\
\hline
\end{array}
\]

\[
\mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z, M_j = 0])^2 \cdot 1_{X_j \leq z, M_j = 0} + (Y - \mathbb{E}[Y|X_j > z, M_j = 0])^2 \cdot 1_{X_j > z, M_j = 0} \right].
\]
CART with missing values: split on available cases

\[ X_1 \leq s_1 \]
\[ X_1 > s_1 \]

\[ \mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z, M_j = 0])^2 \cdot \mathbb{1}_{X_j \leq z, M_j = 0} + (Y - \mathbb{E}[Y|X_j > z, M_j = 0])^2 \cdot \mathbb{1}_{X_j > z, M_j = 0} \right]. \]
CART with missing values: split on available cases

\[
\mathbb{E}\left[ (Y - \mathbb{E}[Y|X_j \leq z, M_j = 0])^2 \cdot \mathbb{I}_{X_j \leq z, M_j = 0} + (Y - \mathbb{E}[Y|X_j > z, M_j = 0])^2 \cdot \mathbb{I}_{X_j > z, M_j = 0} \right].
\]

Propagate missing values?

Probabilistic splits: \textit{Bernoulli}\left( \frac{\#L}{\#L+\#R} \right) (C4.5 algorithm)

Block propag: send all to a side by minimizing the error (xgboost, lightgbm)

Surrogate split: search for a split on another variable that induces a partition close to the original one (rpart)

Implicit impute by an interval: missing values assigned to the left or right
Split on available observations

⇒ Bias in variable selection: tendency to underselect variables with missing values (favor variables where many splits are available)

⇒ Conditional tree (Hothorn, 2006) Ctree selects variables with a test

\[
\begin{align*}
X_1 & \perp X_2 \sim \mathcal{N}(0, 1) \\
Y &= 0.25 X_1 + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0, 1)
\end{align*}
\]

Frequency of selection of \(X_1\) when there are missing values on \(X_1\):

CART selects the non-informative variable \(X_2\) more frequently
Missing incorporated in attribute, Twala et al 2008

Selection of the variable, threshold and propagation of missing values

\[ f^* \in \arg \min_{f \in \mathcal{P}_{c,\text{miss}}} \mathbb{E} \left[ (Y - f(\tilde{X}))^2 \right], \]

where \( \mathcal{P}_{c,\text{miss}} = \mathcal{P}_{c,\text{miss},L} \cup \mathcal{P}_{c,\text{miss},R} \cup \mathcal{P}_{c,\text{miss},\text{sep}} \) with

- \( \mathcal{P}_{c,\text{miss},L} \rightarrow \{\{\tilde{X}_j \leq z \lor \tilde{X}_j = \text{NA}\}, \{\tilde{X}_j > z\}\} \)
- \( \mathcal{P}_{c,\text{miss},R} \rightarrow \{\{\tilde{X}_j \leq z\}, \{\tilde{X}_j > z \lor \tilde{X}_j = \text{NA}\}\} \)
- \( \mathcal{P}_{c,\text{miss},\text{sep}} \rightarrow \{\{\tilde{X}_j \neq \text{NA}\}, \{\tilde{X}_j = \text{NA}\}\} \).

⇒ Missing values treated like a category (well to handle \( \mathbb{R} \cup \text{NA} \))

⇒ Target \( \mathbb{E} \left[ Y \bigg| \tilde{X} \right] = \sum_{m \in \{0,1\}^d} \mathbb{E} \left[ Y \bigg| o(X, m), M = m \right] \mathbb{1}_{M=m} \)

⇒ Good for informative pattern

⇒ Implementation: duplicate the incomplete columns, and replace the missing entries once by \( +\infty \) and once by \( -\infty \).
Split comparison

\[
\begin{cases}
Y = X_1 \\
X_1 \sim U([0, 1])
\end{cases}
\quad \begin{cases}
\mathbb{P}[M_1 = 0] = 1 - \rho \\
\mathbb{P}[M_1 = 1] = \rho
\end{cases},
\]

The best split CART \( s^* = 1/2 \)

The split chosen by the MIA
Consider the regression model
\[
\begin{align*}
Y &= X_1 \\
X_1 &\sim U([0, 1]), \\
X_2 &= X_1 \mathbb{1}_{W=1}
\end{align*}
\]
where \((M_1, W) \perp (X_1, Y)\).

\[
\begin{align*}
\mathbb{P}[W = 0] &= \eta \\
\mathbb{P}[W = 1] &= 1 - \eta \\
\mathbb{P}[M_1 = 0] &= 1 - p \\
\mathbb{P}[M_1 = 1] &= p
\end{align*}
\]
Simulations: 20% missing values

Quadratic: \( Y = X_1^2 + \varepsilon, \ x_i. \in \mathcal{N}(\mu_3, \Sigma_{3 \times 3}), \ \rho = 0.5, \ n = 1000 \)

**MCAR (MAR)**

<table>
<thead>
<tr>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. ctree (surrogates)</td>
<td>8. ctree (surrogates) + mask</td>
<td>7. rpart (surrogates)</td>
</tr>
<tr>
<td>6. rpart (surrogates) + mask</td>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian</td>
</tr>
<tr>
<td>3. impute mean</td>
<td>2. impute mean + mask</td>
<td>1. MIA</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian + mask</td>
<td>3. impute mean</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>2. impute mean + mask</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
<tr>
<td>1. MIA</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
</tbody>
</table>

Relative explained variance

**MNAR**

<table>
<thead>
<tr>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. ctree (surrogates)</td>
<td>8. ctree (surrogates) + mask</td>
<td>7. rpart (surrogates)</td>
</tr>
<tr>
<td>6. rpart (surrogates) + mask</td>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian</td>
</tr>
<tr>
<td>3. impute mean</td>
<td>2. impute mean + mask</td>
<td>1. MIA</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian + mask</td>
<td>3. impute mean</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>2. impute mean + mask</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
<tr>
<td>1. MIA</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
</tbody>
</table>

Relative explained variance

**Predictive**

<table>
<thead>
<tr>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
<th>DECISION TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. ctree (surrogates)</td>
<td>8. ctree (surrogates) + mask</td>
<td>7. rpart (surrogates)</td>
</tr>
<tr>
<td>6. rpart (surrogates) + mask</td>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian</td>
</tr>
<tr>
<td>3. impute mean</td>
<td>2. impute mean + mask</td>
<td>1. MIA</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>5. impute Gaussian</td>
<td>4. impute Gaussian + mask</td>
<td>3. impute mean</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
<td>RANDOM FOREST</td>
</tr>
<tr>
<td>2. impute mean + mask</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
<tr>
<td>1. MIA</td>
<td>RANDOM</td>
<td>RANDOM</td>
</tr>
</tbody>
</table>

Relative explained variance

\[ M_{i,1} \sim \mathcal{B}(p) \]

\[ M_{i,1} = \mathbb{1}_{X_{i,1} > [X_1](1 - \rho_n)} \]

\[ Y = X_1^2 + 3M_1 + \varepsilon \]
Consistency: 40% missing values MCAR

Linear problem (high noise)

Friedman problem (high noise)

Non-linear problem (low noise)

Sample size

Explained variance

Linear problem (high noise)

Friedman problem (high noise)

Non-linear problem (low noise)

Sample size

Explained variance

DECISION TREE

RANDOM FOREST

rpart (surrogates)
rpart (surrogates) + mask
Mean imputation
Mean imputation + mask
Gaussian imputation
Gaussian imputation + mask
MIA
ctree (surrogates)
Bayes rate
Discussion
Discussion

Take-home

- Consistent learner for the fully observed data → multiple imputation on the test set
- Incomplete train and test → same imputation model
- Single mean imputation is consistent, provided a powerful learner
- Tree-based models → Missing Incorporated in Attribute optimizes not only the split but also the handling of the missing values
- Empirically, good imputation methods reduce the number of samples required to reach good prediction
- Informative missing data? Adding the mask helps imputation!

To be done

- Nonasymptotic results
- Prove the usefulness of methods in MNAR
- Uncertainty associated with the prediction
- Distributional shift: no missing values in the test set?
Major trauma: any injury that endangers the life or the functional integrity of a person. Road traffic accidents, interpersonal violence, self-harm, falls, etc → hemorrhage and traumatic brain injury.

Major source of mortality and handicap in France and worldwide (3rd cause of death, 1st cause for 16-45 - 2-3th cause of disability)

⇒ A public health challenge

Patient prognosis can be improved: standardized and reproducible procedures but personalized for the patient and the trauma system.

Trauma decision making: rapid and complex decisions under time pressure in a dynamic and multi-player environment (fragmentation: loss or distortion of information) with high levels of uncertainty and stress. Issues: patient management exceeds time frames, diagnostic errors, decisions not reproducible, etc

⇒ Can Machine Learning, AI help?
Decision support tool for the management of severe trauma: Traumamatrix

PATIENT CENTERED PROBABILISTIC DECISION SUPPORT

112 Dispatch
EMS SAMU
Triage ALS
Resus Room
Stabilisation Damage Control
Scan Theatre Angio
Immediate interventions
ICU Complete care

REAL TIME ADAPTIVE AND LEARNING INFORMATION MANAGEMENT FOR ALL PROVIDERS

Audio-Visual Cockpit style Decision Support

Decision => Management
Causal inference for traumatic brain injury with missing values

- 3050 patients with a brain injury (a lesion visible on the CT scan)
- Treatment: tranexamic acid (binary)
- Outcome: in-ICU death (binary), causes: brain death, withdrawal of care, head injury and multiple organ failure.
- 45 *quantitative & categorical* covariates selected by experts (Delphi process). Pre-hospital (blood pressure, patients reactivity, type of accident, anamnensis, etc.) and hospital data
based on Gaussian assumption: $x_i \sim \mathcal{N}(\mu, \Sigma)$

- Bivariate with missing on $x_{i1}$ (stochastic reg): estimate $\beta$ and $\sigma$ - impute from the predictive $x_{i1} \sim \mathcal{N}(x_{i2}\hat{\beta}, \hat{\sigma}^2)$
- Extension to multivariate case: estimate $\mu$ and $\Sigma$ from an incomplete data with EM - impute by drawing from $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ equivalence conditional expectation and regression (complement Schur)

packages Amelia, mice (conditional)
PCA reconstruction

\[ X \approx F \hat{\mu} \]

⇒ Minimizes distance between observations and their projection
⇒ Approx \( X_{n \times p} \) with a low rank matrix \( k < p \) \( \| A \|_2^2 = \text{tr}(A A^\top) \):

\[
\arg\min_{\mu} \left\{ \| X - \mu \|_2^2 : \text{rank } (\mu) \leq k \right\}
\]

SVD \( X \):

\[
\hat{\mu}_\text{PCA} = U_{n \times k} D_{k \times k} V'_{p \times k} = F_{n \times k} V'_{p \times k}
\]

\( F = UD \) PC - scores

\( V \) principal axes - loadings
PCA reconstruction

\[
\begin{bmatrix}
-2.00 & -2.74 \\
NA & -0.77 \\
-1.11 & -1.59 \\
-0.67 & -1.13 \\
-0.22 & NA \\
0.22 & -0.52 \\
0.67 & 1.46 \\
NA & 0.63 \\
1.56 & -1.10 \\
2.00 & 1.00 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
-2.16 & -2.58 \\
-0.96 & -1.35 \\
-1.15 & -1.55 \\
-0.70 & -1.09 \\
-0.53 & -0.92 \\
0.64 & -0.34 \\
1.24 & 0.69 \\
1.05 & 0.69 \\
1.50 & 1.15 \\
1.67 & 1.33 \\
\end{bmatrix}
\]

\[
F \approx X \hat{\mu}
\]

⇒ Minimizes distance between observations and their projection
⇒ Approx $X_{n \times p}$ with a low rank matrix $k < p \parallel A \parallel_2^2 = \text{tr}(AA^\top)$:

\[
\arg \min_\mu \left\{ \|X - \mu\|_2^2 : \text{rank}(\mu) \leq k \right\}
\]

SVD $X$:

\[
\hat{\mu}_{\text{PCA}}^{\text{PC}} = U_{n \times k} D_{k \times k} V'_{p \times k} = F_{n \times k} V'_{p \times k}
\]

$F = UD$ PC - scores
$V$ principal axes - loadings
Missing values in PCA

⇒ PCA : least squares

\[
\arg \min_\mu \left\{ \| X_{n \times p} - \mu_{n \times p} \|_2^2 : \text{rank} (\mu) \leq k \right\}
\]

⇒ PCA with missing values : weighted least squares

\[
\arg \min_\mu \left\{ \| W_{n \times p} \odot (X - \mu) \|_2^2 : \text{rank} (\mu) \leq k \right\}
\]

with \( w_{ij} = 0 \) if \( x_{ij} \) is missing, \( w_{ij} = 1 \) otherwise ; \( \odot \) elementwise multiplication

Many algorithms :
Gabriel & Zamir, 1979 : weighted alternating least squares (without explicit imputation)
Kiers, 1997 : iterative PCA (with imputation)
Iterative PCA

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>-2.01</td>
</tr>
<tr>
<td>-1.5</td>
<td>-1.48</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.01</td>
</tr>
<tr>
<td>1.5</td>
<td>NA</td>
</tr>
<tr>
<td>2.0</td>
<td>1.98</td>
</tr>
</tbody>
</table>
Iterative PCA

Initialization $\ell = 0 : X^0$ (mean imputation)
Iterative PCA

$\begin{array}{cc}
-2.0 & -2.01 \\
-1.5 & -1.48 \\
0.0 & -0.01 \\
1.5 & \text{NA} \\
2.0 & 1.98 \\
\end{array}$

$\begin{array}{cc}
-2.0 & -2.01 \\
-1.5 & -1.48 \\
0.0 & -0.01 \\
1.5 & 0.00 \\
2.0 & 1.98 \\
\end{array}$

$\begin{array}{cc}
-1.98 & -2.04 \\
-1.44 & -1.56 \\
0.15 & -0.18 \\
1.00 & 0.57 \\
2.27 & 1.67 \\
\end{array}$

PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, D^\ell)$;
Iterative PCA

\[ \hat{\mu}^\ell = U^\ell D^\ell V^\ell \]

Missing values imputed with the fitted matrix \( \hat{\mu}^\ell = U^\ell D^\ell V^\ell \).
The new imputed dataset is $\hat{X}^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
Iterative PCA

\[
\begin{array}{llllllll}
  x1 & x2 \\
-2.0 & -2.01 \\
-1.5 & -1.48 \\
0.0 & -0.01 \\
1.5 & \text{NA} \\
2.0 & 1.98 \\
\end{array}
\]
Iterative PCA

\begin{align*}
\begin{array}{llllllll}
  x_1 & x_2 \\
-2.0 & -2.01 \\
-1.5 & -1.48 \\
0.0 & -0.01 \\
1.5 & \text{NA} \\
2.0 & 1.98 \\
\end{array}
\end{align*}

\begin{align*}
\begin{array}{llllllll}
  x_1 & x_2 \\
-2.00 & -2.01 \\
-1.47 & -1.52 \\
0.09 & -0.11 \\
1.20 & 0.90 \\
2.18 & 1.78 \\
\end{array}
\end{align*}

\begin{align*}
\begin{array}{llllllll}
  x_1 & x_2 \\
-2.0 & -2.01 \\
-1.5 & -1.48 \\
0.0 & -0.01 \\
1.5 & 0.90 \\
2.0 & 1.98 \\
\end{array}
\end{align*}
Steps are repeated until convergence
Iterative PCA

PCA on the completed data set $\rightarrow (U^\ell, D^\ell, V^\ell)$

Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell D^\ell V^{\ell'}$
Iterative PCA

1. initialization $\ell = 0 : X^0$ (mean imputation)
2. step $\ell$ :
   (a) PCA on the completed data $\rightarrow (U^\ell, D^\ell, V^\ell)$; $k$ dim kept
   (b) $\hat{\mu}^{\text{PCA}} = \sum_{q=1}^{k} d_q u_q v_q'$
      $X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

$\Rightarrow$ Overfitting : nb param $(U_{n \times k}, V_{k \times p})$/obs values : $k$ large - NA ; noisy

Regularized versions. Imputation is replaced by

$$(\hat{\mu})_\lambda = \sum_{q=1}^{p} (d_q - \lambda)_+ u_q v_q'$

$\text{arg min}_\mu \left\{ \| W \odot (X - \mu) \|_2^2 + \lambda \| \mu \|_* \right\}$


$\Rightarrow$ Iterative SVD algo good to impute data (matrix completion, Netflix)

$\Rightarrow$ Model makes sense : data $= \text{rank } k \text{ signal} + \text{ noise}$

$X = \mu + \varepsilon \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ with $\mu$ of low rank

(Udell & Townsend, 2017)
Iterative SVD

⇒ Imputation with FAMD for mixed data:

<table>
<thead>
<tr>
<th>age</th>
<th>weight</th>
<th>size</th>
<th>alcohol</th>
<th>sex</th>
<th>snore</th>
<th>tobacco</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>100</td>
<td>190</td>
<td>NA</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>70</td>
<td>96</td>
<td>186</td>
<td>1-2 gl/d</td>
<td>M</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>NA</td>
<td>104</td>
<td>194</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>62</td>
<td>68</td>
<td>165</td>
<td>1-2 gl/d</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>weight</th>
<th>size</th>
<th>alcohol</th>
<th>sex</th>
<th>snore</th>
<th>tobacco</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>100</td>
<td>190</td>
<td>1-2 gl/d</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>70</td>
<td>96</td>
<td>186</td>
<td>1-2 gl/d</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>104</td>
<td>194</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>62</td>
<td>68</td>
<td>165</td>
<td>1-2 gl/d</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
</tbody>
</table>

⇒ Multilevel imputation: hospital effect with patient nested in hospital.

(J., Husson, Robin & Balasu., 2018, Imputation of mixed data with multilevel SVD. *JCGS* )

package MissMDA.
Imputation methods: conditional model

Imputation with fully conditional specification (FCS). Impute with a joint model defined implicitly through the conditional distributions (mice).

⇒ Imputation model for each variable is a forest.

1. Initial imputation: mean imputation - random category
2. for \( t \) in 1 : \( T \) loop through iterations \( t \)
3. for \( j \) in 1 : \( p \) loop through variables \( j \)

Define currently complete data set except \( X^t_{-j} = (X^t_1, X^t_{j-1}, X^t_{j+1}, X^t_{p-1}) \), then \( X^t_j \) is obtained by

- fitting a RF \( X^t_j^{obs} \) on the other variables \( X^t_{-j} \)
- predicting \( X^t_j^{miss} \) using the trained RF on \( X^t_{-j} \)

package missForest (Stekhoven & Buhlmann, 2011)
Mechanism

\( M = (M_1, \ldots, M_d) \): indicator of missing values in \( X = (X_1, \ldots, X_d) \).

**Missing value mechanisms (Rubin, 1976)**

- **MCAR**  
  \( \forall \phi, \forall m, x, g_\phi(m|x) = g_\phi(m) \)

- **MAR**  
  \( \forall \phi, \forall i, \forall x', o(x', m_i) = o(x_i, m_i) \implies g_\phi(m_i|x') = g_\phi(m_i|x_i) \)
  
  (e.g. \( g_\phi((0, 0, 1, 0) \mid (3, 2, 4, 8)) = g_\phi((0, 0, 1, 0) \mid (3, 2, 7, 8)) \))

- **MNAR**  
  Not MAR

\( \implies \) useful for likelihoods

**Missing value mechanisms – variable level**

- **MCAR**  
  \( M \perp\!\!\!\!\!\!\!\!\!\| X \)

- **MAR (bis)**  
  \( \forall S \subset \{1, \ldots, d\}, (M_j)_{j \in S} \perp\!\!\!\!\!\!\!\!\!\| (X_j)_{j \in S} \mid (X_k)_{k \in S^c} \)

- **MNAR**  
  Not MAR

\( \implies \) useful for our results
Let \( X \sim f_{\theta^*} \).

Observed log-likelihood

\[
\ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_{\theta}(x) \, d\delta_{o(\cdot, m_i) = o(x_i, m_i)}(x).
\]

(inspired by Seaman 2013)

**Example**

\[
X_1, X_2 \sim f_{\theta}(x_1)g_{\theta}(x_2|x_1)
\]

\[
M_{1,2}, \ldots, M_{r,2} = 1
\]

\[
\ell_{\text{obs}}(\theta) = \sum_{i=1}^{r} \log f_{\theta}(x_1) + \sum_{i=r+1}^{n} \log f_{\theta}(x_1)g_{\theta}(x_2|x_1).
\]
Let $X \sim f_{\theta^*}$.

Observed log-likelihood

$$\ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(x) \, d\delta_{o(\cdot,m_i)=o(x_i,m_i)}(x).$$

Full log-likelihood

$$\ell_{\text{full}}(\theta, \phi) = \sum_{i=1}^{n} \log \int f_\theta(x) g_\phi(m_i|x) \, d\delta_{o(\cdot,m_i)=o(x_i,m_i)}(x).$$

**Theorem (Theorem 7.1 in Rubin 1976)**

$\theta$ can be inferred from $\ell_{\text{obs}}$, assuming MAR.
Ignorable mechanism

Full log-likelihood:

\[ \ell_{\text{full}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(x) g_\phi(m_i|x) \, d\delta_{o(\cdot,m_i)=o(x_i,m_i)}(x). \]

Observed log-likelihood:

\[ \ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(x) \, d\delta_{o(\cdot,m_i)=o(x_i,m_i)}(x). \]

Assuming MAR,

\[ \ell_{\text{full}}(\theta, \phi) = \sum_{i=1}^{n} \log \int f_\theta(x) g_\phi(m_i|x_i) \, d\delta_{o(\cdot,m_i)=o(x_i,m_i)}(x_i) \]

\[ = \ell_{\text{obs}}(\theta) + \sum_{i=1}^{n} \log g_\phi(m_i|x_i). \]
Let $X \sim f_{\theta^*}$.

**Observed log-likelihood**

$$\ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_{\theta}(x) \, d\delta_{o(\cdot, m_i) = o(x_i, m_i)}(x).$$

**EM algorithm (Dempster, 1977)**

Starting from an initial parameter $\theta^{(0)}$, the algorithm alternates the two following steps,

**(E-step)**

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \int (\log f_{\theta}(x)) f_{\theta^{(t)}}(x) \, d\delta_{o(\cdot, m_i) = o(x_i, m_i)}(x).$$

**(M-step)**

$$\theta^{(t+1)} \in \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t)}).$$

The likelihood is guaranteed to increase.
\[ \tilde{X} = X \odot (1 - M) + \text{NA} \odot M \text{ takes value in } \mathbb{R} \cup \{\text{NA}\} \]

The (unobserved) complete sample \( D_n = (X_i, M_i, Y_i)_{1 \leq i \leq n} \sim (X, M, Y) \)

\[
d_n = \begin{bmatrix}
2 & 3 & 1 & 0 & 0 & 0 & 1 & 0 & 15 \\
1 & 0 & 3 & 5 & 0 & 1 & 0 & 0 & 13 \\
9 & 4 & 2 & 5 & 0 & 0 & 0 & 1 & 18 \\
7 & 6 & 3 & 2 & 0 & 0 & 1 & 1 & 10 \\
\end{bmatrix},
\]

The observed training set \( \tilde{D}_{n,\text{train}} = (\tilde{X}_i, Y_i)_{1 \leq i \leq n} \)

\[
\tilde{d}_n = \begin{bmatrix}
2 & 3 & \text{NA} & 0 & 15 \\
1 & \text{NA} & 3 & 5 & 13 \\
9 & 4 & 2 & \text{NA} & 18 \\
7 & 6 & \text{NA} & \text{NA} & 10 \\
\end{bmatrix}.
\]