Missing values imputation for mixed data based on principal component methods

Vincent Audigier, François Husson & Julie Josse

Agrocampus Rennes

Compstat’ 2012, Limassol (Cyprus), 28-08-2012
### A real dataset

<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 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>70</td>
<td>96</td>
<td>186</td>
<td>1 or 2 glasses/day</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 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>91</td>
<td>180</td>
<td>No</td>
<td>W</td>
<td>yes</td>
<td>&gt;1</td>
</tr>
<tr>
<td>50</td>
<td>109</td>
<td>195</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>68</td>
<td>98</td>
<td>188</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>49</td>
<td>90</td>
<td>179</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>55</td>
<td>57</td>
<td>163</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&gt;1</td>
</tr>
<tr>
<td>61</td>
<td>61</td>
<td>167</td>
<td>1 or 2 glasses/day</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>63</td>
<td>108</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>34</td>
<td>92</td>
<td>181</td>
<td>1 or 2 glasses/day</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>44</td>
<td>91</td>
<td>180</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>57</td>
<td>97</td>
<td>187</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>46</td>
<td>117</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>45</td>
<td>104</td>
<td>194</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>69</td>
<td>107</td>
<td>198</td>
<td>No</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>58</td>
<td>98</td>
<td>188</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>65</td>
<td>105</td>
<td>196</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>43</td>
<td>108</td>
<td>194</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>38</td>
<td>69</td>
<td>166</td>
<td>1 or 2 glasses/day</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
</tbody>
</table>
A real dataset

<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>NA</td>
<td>172</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 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>NA</td>
<td>164</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 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>91</td>
<td>180</td>
<td>No</td>
<td>W</td>
<td>yes</td>
<td>&gt;1</td>
</tr>
<tr>
<td>50</td>
<td>109</td>
<td>NA</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>68</td>
<td>98</td>
<td>188</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>49</td>
<td>NA</td>
<td>179</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>65</td>
<td>57</td>
<td>163</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&gt;1</td>
</tr>
<tr>
<td>NA</td>
<td>61</td>
<td>167</td>
<td>1 or 2 glasses/day</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>63</td>
<td>108</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>34</td>
<td>NA</td>
<td>181</td>
<td>NA</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>44</td>
<td>91</td>
<td>NA</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>57</td>
<td>97</td>
<td>NA</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>46</td>
<td>117</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>NA</td>
<td>104</td>
<td>168</td>
<td>No</td>
<td>W</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>69</td>
<td>107</td>
<td>198</td>
<td>No</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>58</td>
<td>98</td>
<td>NA</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>65</td>
<td>NA</td>
<td>186</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>43</td>
<td>108</td>
<td>174</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>38</td>
<td>69</td>
<td>166</td>
<td>NA</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
</tbody>
</table>
### A real dataset

<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>NA</td>
<td>172</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 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>NA</td>
<td>164</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 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>48</td>
<td>91</td>
<td>180</td>
<td>No</td>
<td>W</td>
<td>yes</td>
<td>&gt;1</td>
</tr>
<tr>
<td>50</td>
<td>109</td>
<td>NA</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>68</td>
<td>98</td>
<td>188</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>49</td>
<td>NA</td>
<td>179</td>
<td>No</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>65</td>
<td>57</td>
<td>163</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&gt;1</td>
</tr>
<tr>
<td>NA</td>
<td>61</td>
<td>167</td>
<td>1 or 2 glasses/day</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>63</td>
<td>108</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>34</td>
<td>NA</td>
<td>181</td>
<td>NA</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>44</td>
<td>91</td>
<td>NA</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>57</td>
<td>97</td>
<td>NA</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>46</td>
<td>117</td>
<td>194</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>NA</td>
<td>104</td>
<td>168</td>
<td>No</td>
<td>W</td>
<td>NA</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>69</td>
<td>107</td>
<td>198</td>
<td>No</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>58</td>
<td>98</td>
<td>NA</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>65</td>
<td>NA</td>
<td>186</td>
<td>1 or 2 glasses/day</td>
<td>M</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>43</td>
<td>108</td>
<td>174</td>
<td>&gt;2 glasses/day</td>
<td>M</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>38</td>
<td>69</td>
<td>166</td>
<td>NA</td>
<td>W</td>
<td>no</td>
<td>&lt;=1</td>
</tr>
</tbody>
</table>

⇒ Popular approach to deal with missing values: Single imputation
**Single imputation methods**

**Continuous variables:** k-nearest neighbours, normal distribution (joint modelling), iterative regression (fully conditional specification), etc.

**Categorical variables:** hot-deck imputation, multinomial model, latent class model (Vermunt et al., 2008), etc.

**Mixed data:**

- transform the categorical variables into dummy variables and deal as continuous variables (package Amelia)
- MICE (multivariate imputation by chained equations, van Bureen 1999): a model must be specified for each variable
- random forest (Stekhoven & Bühlmann, 2011)

⇒ New imputation method based on principal component methods
Imputation with PCA for continuous variables

PCA minimizes:

\[ C = \| X_{I \times J} - F_{I \times S} U_{S \times J}^t \|^2 \]

With missing values:

\[ C = \| W \ast (X - FU^t) \|^2, \]

with \( w_{ij} = 0 \) if \( x_{ij} \) is missing, \( w_{ij} = 1 \) otherwise.

\[ \Rightarrow \text{Iterative PCA (Kiers, 1997)} \]
Iterative PCA algorithm

The data set

<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 algorithm

Initialization step: mean imputation
Iterative PCA algorithm

PCA performed on the completed data set; 1 dimension is kept

```
x1   x2
-2.0 -2.01
-1.5 -1.48
0.0  -0.01
1.5  NA
2.0  1.98

|x1   x2
-2.0 -2.01
-1.5 -1.48
0.0  -0.01
1.5  0.00
2.0  1.98

[^]   ^
x1   x2
-1.98 -2.04
-1.44 -1.56
0.15  -0.18
1.00  0.57
2.27  1.67
```
Iterative PCA algorithm

Calculation of the model prediction
**Iterative PCA algorithm**

Imputation step: \( \mathbf{X}^\ell = \mathbf{W} \ast \mathbf{X} + (1 - \mathbf{W}) \ast \hat{\mathbf{X}}^\ell \)
Missing values in PCA

Iterative PCA algorithm

PCA is performed; 1 dimension is kept
Iterative PCA algorithm

Imputation step: \( \mathbf{X}^\ell = \mathbf{W} \mathbf{X} + (1 - \mathbf{W}) \mathbf{\hat{X}}^\ell \)
Iterative PCA algorithm

Iterate until convergence

<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>0.00</td>
</tr>
<tr>
<td>2.0</td>
<td>1.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.98</td>
<td>-2.04</td>
</tr>
<tr>
<td>-1.44</td>
<td>-1.56</td>
</tr>
<tr>
<td>0.15</td>
<td>-0.18</td>
</tr>
<tr>
<td>1.00</td>
<td>0.57</td>
</tr>
<tr>
<td>2.27</td>
<td>1.67</td>
</tr>
</tbody>
</table>

<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>0.57</td>
</tr>
<tr>
<td>2.0</td>
<td>1.98</td>
</tr>
</tbody>
</table>
Iterative PCA - convergence

Imputed values are obtained at convergence

```
x1  x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  NA
 2.0  1.98
```

```
x1  x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  1.46
 2.0  1.98
```
Iterative PCA

1. initialization \( \ell = 0: X^0 \) (mean imputation)
2. step \( \ell \):
   (a) PCA on the completed matrix \( X^{\ell-1} \rightarrow \hat{F}^\ell_{I \times S}, \hat{U}^\ell_{K \times S} \)
       \( S \) dimensions are kept; \( \hat{X}^\ell = \hat{F}^\ell \hat{U}^{\ell'} \)
   (estimation)
   (b) \( X^\ell = W \ast X + (1 - W) \ast \hat{F}^\ell \hat{U}^{\ell'} \)
       (imputation)
3. Estimation and imputation are repeated until convergence

- The number of dimensions \( S \) has to be chosen \textit{a priori}
- An imputation is performed during the algorithm
  \( \Rightarrow \) PCA can be seen as an imputation method
- Overfitting problems are managed with a regularized algorithm
Imputation with MCA for categorical variables

MCA can be seen as the PCA on an indicator matrix $X$ with specific weights depending on $D_\Sigma$

\[
X = D_\Sigma \Sigma J
\]

\[
X = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 & 1 & \cdots & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & \cdots & 0 & 1 \\
NA & NA & NA & 0 & 1 & 0 & 0 & \cdots & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \cdots & 0 & 1 \\
\end{bmatrix}
\]

\[
D_\Sigma = \begin{bmatrix}
0 & 0 & 1 & NA & NA & 0 & \cdots & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \cdots & 0 & 1 \\
\end{bmatrix}
\]

\[
J = \begin{bmatrix}
I_1 \\
J \\
I_J \\
I_K \\
\end{bmatrix}
\]

\[
D_\Sigma = \begin{bmatrix}
0 & I_k \\
\cdots & \cdots \\
I_K & 0 \\
\end{bmatrix}
\]
Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) MCA on the completed indicator matrix $\Rightarrow \hat{F}, \hat{U}$ (Estimation)
   (b) imputation of the missing values with the model matrix
   (c) column margins are updated

$\Rightarrow$ imputed values can be seen as degree of membership
Principal component method for mixed data

The core of principal component methods is PCA on particular matrices

"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize" (Benzécri)
**Principal component method for mixed data (complete case)**

Factorial Analysis on Mixed Data (Escofier, 1979), PCAMIX (Kiers, 1991)

<table>
<thead>
<tr>
<th>Continuous variables</th>
<th>Categorical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>51 100 190</td>
<td>0 1 0</td>
</tr>
<tr>
<td>70 96 196</td>
<td>0 1 0 10</td>
</tr>
<tr>
<td>38 69 166</td>
<td>0 1 0 1 0 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicator matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$ $I_2$ $I_k$</td>
</tr>
</tbody>
</table>

- Centring & scaling
- Division by $\sqrt{I/I_k}$ and centring

Matrix which balances the influence of each variable

A PCA is performed on the weighted matrix
Properties of the method

• The distance between individuals is:

\[
d^2(i, l) = \sum_{k=1}^{K_{cont}} (x_{ik} - x_{lk})^2 + \sum_{q=1}^{Q} \sum_{k=1}^{K_q} \frac{1}{I_{kq}} (x_{iq} - x_{lq})^2
\]

• The principal component \( F_s \) maximises:

\[
\sum_{k=1}^{K_{cont}} r^2(F_s, v_k) + \sum_{q=1}^{Q_{cat}} \eta^2(F_s, v_q)
\]
## Imputation of mixed data

The same kind of iterative algorithm as before

<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>

<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>1.0</td>
<td>0.7</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>96</td>
<td>186</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td>48</td>
<td>104</td>
<td>194</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td>62</td>
<td>68</td>
<td>165</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>NA</td>
</tr>
</tbody>
</table>

### imputeAFDM

<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>0.2</td>
<td>0.7</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>96</td>
<td>186</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.8</td>
</tr>
<tr>
<td>48</td>
<td>104</td>
<td>194</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>62</td>
<td>68</td>
<td>165</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Properties on the imputations

- Imputation based on scores and loadings ⇒ similarities between individuals and relationships between variables
- Relationships between continuous and categorical variables are taken into account
- Categorical variables evolve within many dimensions \((k_q - 1)\) if they have \(k_q\) categories) so they need many dimensions to be well predicted
- Compared to a PCA on the (unweighted) indicator matrix, small categories are better imputed
Simulations

• Simulation pattern
  • 2 independent variables are drawn from a normal distribution
  • 1 variable is replicated 4 times, the other 8 ⇒ 2 dimensions
  • Random noise is added
  • Half of the variables in each dimension are split in 3 clusters
  • 10%, 20% or 30% of missing values are chosen at random
  ⇒ Data are constructed (expected) to be in 4 dimensions

• Criterion
  • for continuous data:
    \[
    N2RMSE = \sqrt{\sum_{i \in \text{missing}} \frac{\text{mean} \left( \left( X_{i}^{\text{true}} - X_{i}^{\text{imp}} \right)^2 \right)}{\text{var} (X_{i}^{\text{true}})}}
    \]
  • for categorical data: proportion of falsely classified entries
Simulations

Imputation using continuous data only

Imputation using both continuous and categorical data

Error on continuous data
Simulations

Imputation using continuous data only
Imputation using both continuous and categorical data

Error on continuous data

Categorical data improved the imputation on continuous data ...
Simulations

Imputation using continuous data only
Imputation using categorical data only
Imputation using both continuous and categorical data

Categorical data improved the imputation on continuous data ...
Simulations

Imputation using continuous data only
Imputation using categorical data only
Imputation using both continuous and categorical data

Categorical data improved the imputation on continuous data...

... and continuous data improved the imputation on categorical data

The error on the estimation of the number of dimensions has not an important impact on the imputation error ... if the estimation is not too bad.
Comparison with random forest on real data sets

Imputations obtained with random forest & iterative algorithm

GBSG2

N2RMSE

RF 10% AFDM 10% RF 20% AFDM 20% RF 30% AFDM 30%

2.2 2.4 2.6 2.8

Ozone

N2RMSE

RF 10% AFDM 10% RF 20% AFDM 20% RF 30% AFDM 30%

0.2 0.3 0.4 0.5

PFC

0.25 0.30 0.35
Comparison with random forest on real data sets

Imputations obtained with random forest & iterative algorithm

GBSG2

N2RMSE

Ozone

N2RMSE

PFC

PFC
Comparison with random forest

Compared to random forest, imputations are quite similar

Imputations are slightly better:
- for categorical variables
- especially for rare categories

and imputations are worse:
- when there are non-linear relationships between continuous variables
- when there are interactions
Conclusion

- A new way to impute missing values
- is efficient when strong linear relationships between variables (you learn from the other variables) ...
- ... but needs tuning parameters, cv approximation?
- is available in the missMDA package. This package:
  - handles missing values in principal component methods (PCA, MCA, MIXPCA, MFA)
  - impute missing values for continuous, categorical and mixed variables
  - performs multiple imputation for continuous variables
How to perform a statistical analysis from an incomplete dataset?

- we can modify the estimation process to apply it on an incomplete dataset (not always easy!)
- we can predict the missing entries with a single imputation method, but BE CAREFUL using the usual methods leads to underestimate the standard errors

⇒ An alternative is to use multiple imputation ... and single imputation is a first step towards multiple imputation