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

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1 or 2 glasses/day	M	yes	no
70	96	186	1 or 2 glasses/day	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1 or 2 glasses/day	M	no	<=1
48	91	180	No	W	yes	>1
50	109	195	>2 glasses/day	M	yes	no
68	98	188	1 or 2 glasses/day	M	yes	<=1
49	90	179	No	W	no	<=1
65	57	163	>2 glasses/day	M	no	>1
61	61	167	1 or 2 glasses/day	W	no	<=1
63	108	194	1 or 2 glasses/day	M	no	no
34	92	181	1 or 2 glasses/day	W	no	<=1
44	91	180	1 or 2 glasses/day	M	yes	<=1
57	97	187	>2 glasses/day	M	yes	<=1
46	117	194	1 or 2 glasses/day	M	no	<=1
45	104	194	No	W	no	<=1
69	107	198	No	M	no	<=1
58	98	188	1 or 2 glasses/day	M	yes	<=1
65	105	196	1 or 2 glasses/day	M	yes	no
43	108	194	>2 glasses/day	M	no	<=1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	69	166	1 or 2 glasses/day	W	no	<=1

A real dataset

age	weight	size	alcohol	sex	snore	tobacco
51	NA	172	NA	M	yes	no
70	96	186	1 or 2 glasses/day	M	NA	<=1
48	NA	164	No	W	no	NA
62	68	165	1 or 2 glasses/day	M	no	<=1
48	91	180	No	W	yes	>1
50	109	NA	>2 glasses/day	M	yes	no
68	98	188	1 or 2 glasses/day	M	NA	NA
49	NA	179	No	W	no	<=1
65	57	163	>2 glasses/day	M	NA	>1
NA	61	167	1 or 2 glasses/day	W	no	<=1
63	108	194	1 or 2 glasses/day	M	no	no
34	NA	181	NA	W	no	<=1
44	91	NA	1 or 2 glasses/day	M	yes	<=1
57	97	NA	>2 glasses/day	M	NA	<=1
46	117	194	1 or 2 glasses/day	M	no	NA
NA	104	168	No	W	NA	<=1
69	107	198	No	M	no	<=1
58	98	NA	1 or 2 glasses/day	M	NA	NA
65	NA	186	1 or 2 glasses/day	M	yes	no
43	108	174	>2 glasses/day	M	no	<=1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	69	166	NA	W	no	<=1

A real dataset

age	weight	size	alcohol	sex	snore	tobacco
51	NA	172	NA	M	yes	no
70	96	186	1 or 2 glasses/day	M	NA	<=1
48	NA	164	No	W	no	NA
62	68	165	1 or 2 glasses/day	M	no	<=1
48	91	180	No	W	yes	>1
50	109	NA	>2 glasses/day	M	yes	no
68	98	188	1 or 2 glasses/day	M	NA	NA
49	NA	179	No	W	no	<=1
65	57	163	>2 glasses/day	M	NA	>1
NA	61	167	1 or 2 glasses/day	W	no	<=1
63	108	194	1 or 2 glasses/day	M	no	no
34	NA	181	NA	W	no	<=1
44	91	NA	1 or 2 glasses/day	M	yes	<=1
57	97	NA	>2 glasses/day	M	NA	<=1
46	117	194	1 or 2 glasses/day	M	no	NA
NA	104	168	No	W	NA	<=1
69	107	198	No	M	no	<=1
58	98	NA	1 or 2 glasses/day	M	NA	NA
65	NA	186	1 or 2 glasses/day	M	yes	no
43	108	174	>2 glasses/day	M	no	<=1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	69	166	NA	W	no	<=1

$\Rightarrow$ 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)

$\Rightarrow$ New imputation method based on principal component methods

Imputation with PCA for continuous variables

PCA minimizes:

$$\mathcal{C} \quad = \quad \|\mathbf{X}_{I\times J} - \mathbf{F}_{I\times S}\mathbf{U}^t_{S\times J}\|^2$$

With missing values:

$$\mathcal{C} = \|\mathbf{W} * (\mathbf{X} - \mathbf{F}\mathbf{U}^t)\|^2,$$

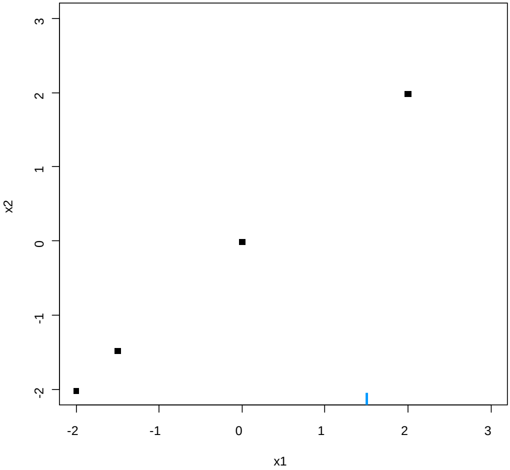
with $w_{ij} = 0$ if x_{ij} is missing, $w_{ij} = 1$ otherwise.

$\Rightarrow$ Iterative PCA (Kiers, 1997)

Iterative PCA algorithm

The data set

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

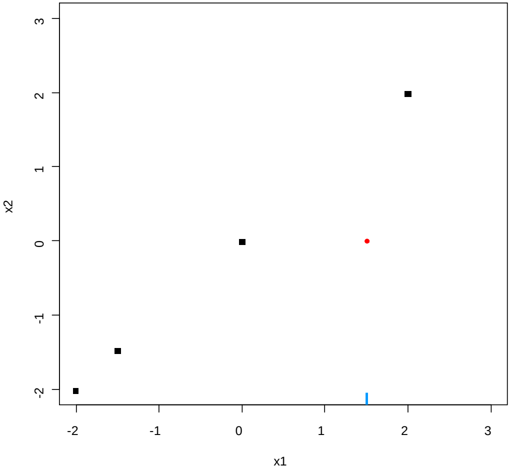


Iterative PCA algorithm

Initialization step: mean imputation

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



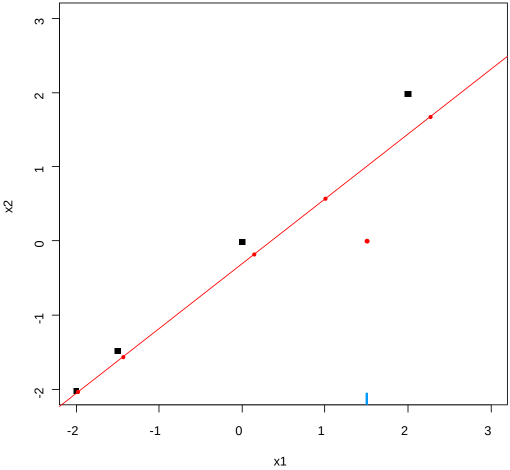
Iterative PCA algorithm

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

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



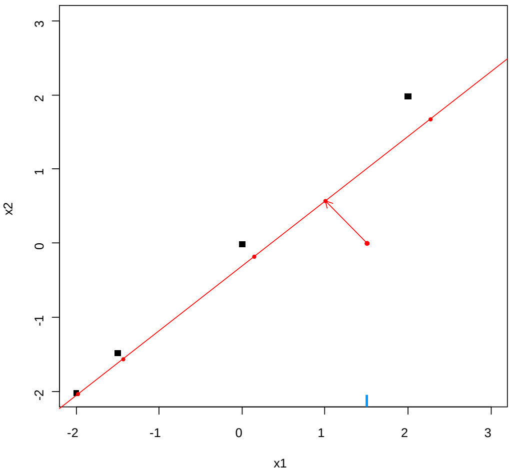
Iterative PCA algorithm

Calculation of the model prediction

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

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67



Iterative PCA algorithm

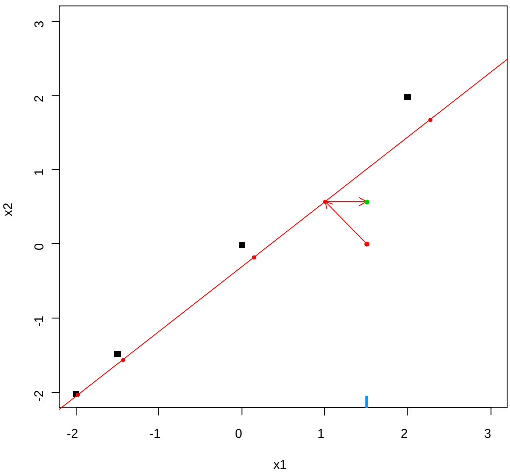
Imputation step: $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$

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

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



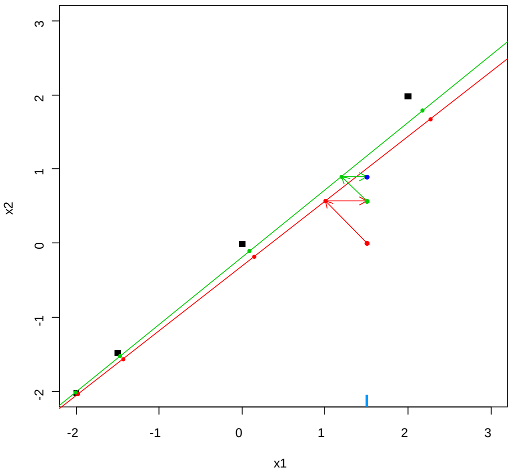
Iterative PCA algorithm

PCA is performed; 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	**0.57**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Iterative PCA algorithm

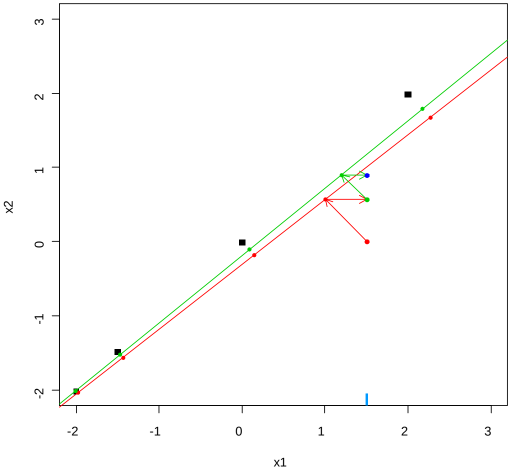
Imputation step: $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$

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	**0.57**
2.0	1.98

$\hat{x1}$	$\hat{x2}$
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	**0.90**
2.18	1.78

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.90**
2.0	1.98



Iterative PCA algorithm

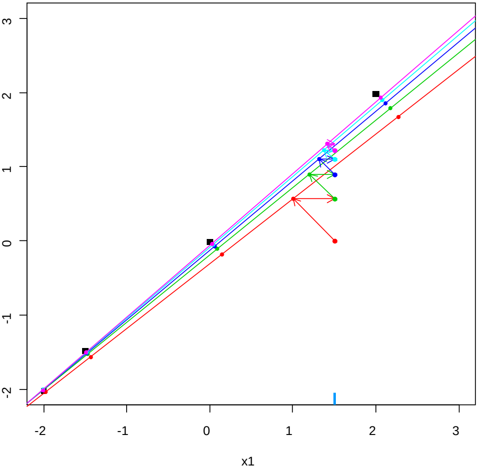
Iterate until convergence

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	**0.00**
2.0	1.98

$\hat{x}1$	$\hat{x}2$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67

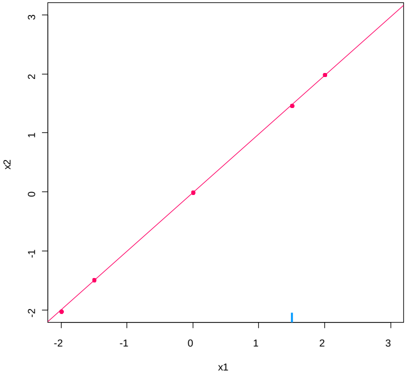
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



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

- The number of dimensions S has to be chosen *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 $\mathbf{D}_\Sigma$

$$
X = \begin{array}{c}
\left[\begin{array}{ccc|cc|cccc|cc}
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots & & \text{NA} & \text{NA} \\
\text{NA} & \text{NA} & \text{NA} & 0 & 1 & 0 & 0 & \ldots & & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & & 0 & 1 \\
 & & & & & & x_{ik} & & & & \\
0 & 0 & 1 & \text{NA} & \text{NA} & 0 & \ldots & 0 & & 1 & \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & & 0 & 1
\end{array}\right]
\begin{array}{c} J \\ \\ \\ \\ J \\ \\ J \end{array} \\
\begin{array}{cccc} I_1 & I_k & I_K & IJ \end{array}
\end{array}
\qquad
D_\Sigma = \begin{bmatrix}
I_1 & & & & 0 \\
 & \ddots & & & \\
 & & I_k & & \\
 & 0 & & \ddots & \\
 & & & & I_K
\end{bmatrix}
$$

Iterative MCA

Iterative MCA algorithm:

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

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

⇒ imputed values can be seen as degree of membership

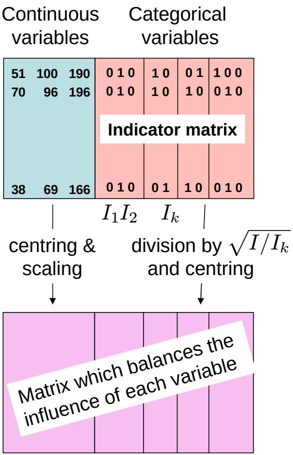
Principal component method for mixed data

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)



A PCA is performed on the weighted matrix

Properties of the method

- The distance between individuals is:

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

- The principal component $\mathbf{F}_s$ maximises:

$$\sum_{k=1}^{K_{cont}} r^2(\mathbf{F}_s, v_k) + \sum_{q=1}^{Q_{cat}} \eta^2(\mathbf{F}_s, v_q)$$

Imputation of mixed data

The same kind of iterative algorithm as before

age	weight	size	alcohol	sex	snore	tobacco
NA	100	190	NA	M	yes	no
70	96	186	1-2 gl/d	M	NA	<=1
NA	104	194	No	W	no	NA
62	68	165	1-2 gl/d	M	no	<=1

NA	100	190	NA	NA	NA	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	NA	NA	0	1	0
NA	104	194	1	0	0	0	1	1	0	NA	NA	NA
62	68	165	0	1	0	1	0	1	0	0	1	0

imputeAFDM

51	100	190	0.2	0.7	0.1	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	0.8	0.2	0	1	0
48	104	194	1	0	0	0	1	1	0	0.1	0.8	0.1
62	68	165	0	1	0	1	0	1	0	0	1	0

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1-2 gl/d	M	yes	no
70	96	186	1-2 gl/d	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1-2 gl/d	M	no	<=1

Properties on the imputations

- Imputation based on scores and loadings $\Rightarrow$ 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 $\Rightarrow$ 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

$\Rightarrow$ Data are constructed (expected) to be in 4 dimensions

- Criterion
  - for continuous data:

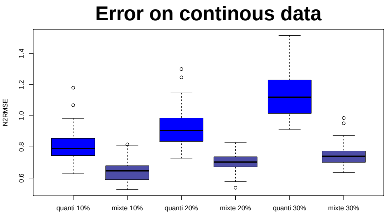
$$N2RMSE = \sqrt{\sum_{i \in \text{missing}} \frac{mean\left(\left(X_i^{true} - X_i^{imp}\right)^2\right)}{var\left(X_i^{true}\right)}}$$

  - for categorical data: proportion of falsely classified entries

Simulations

Imputation using continuous data only

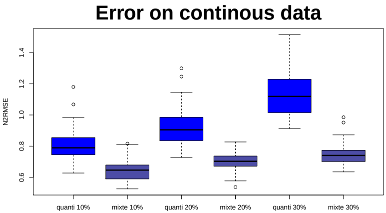
Imputation using both continuous and categorical data



Simulations

Imputation using continuous data only

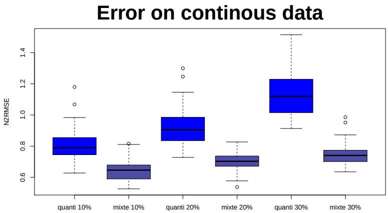
Imputation using both continuous and categorical data

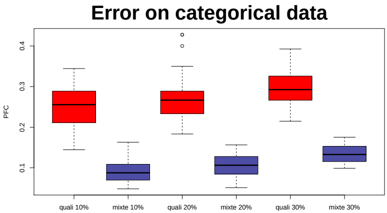


Categorical data improved the imputation on continuous data ...

Simulations

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

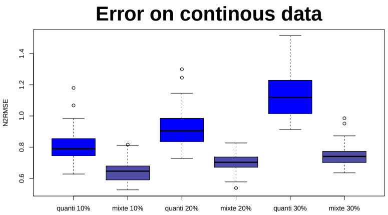


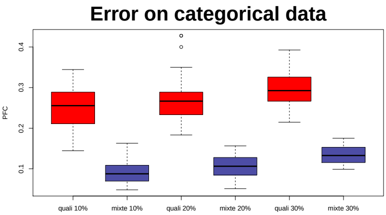


Categorical data improved the imputation on continuous data ...

Simulations

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

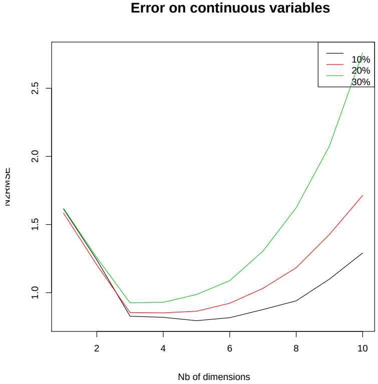


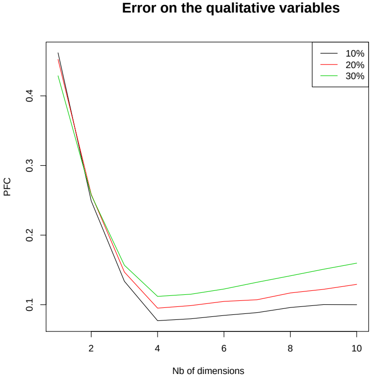


Categorical data improved the imputation on continuous data ...

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

Simulations

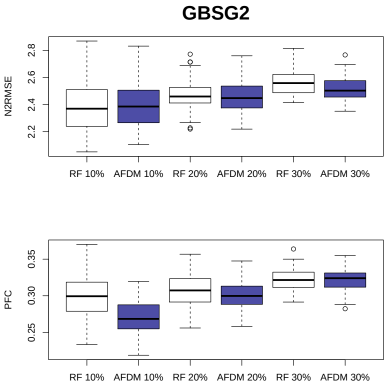




$\Rightarrow$ The error on the estimation of the number of dimensions has not an important impact on the imputation error ... if the estimation is not too bad

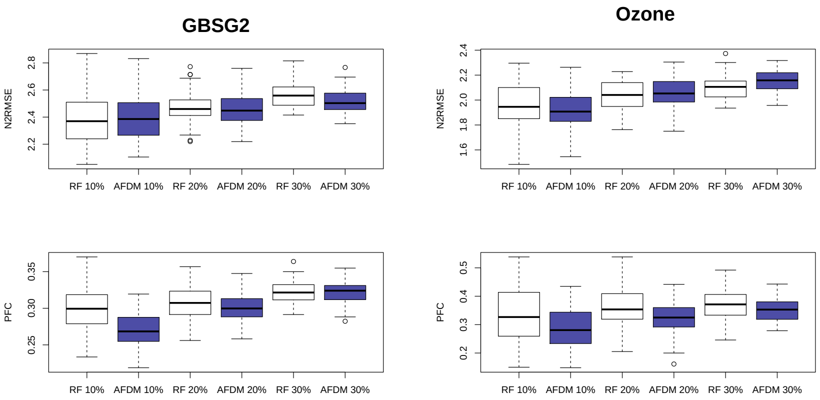
Comparison with random forest on real data sets

Imputations obtained with random forest & iterative algorithm



Comparison with random forest on real data sets

Imputations obtained with random forest & iterative algorithm



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

Perspective

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

$\Rightarrow$ An alternative is to use multiple imputation ... and single imputation is a first step towards multiple imputation