Imputation for mixed data: Random Forest versus PCA

Vincent Audigier, François Husson & Julie Josse

Agrocampus Rennes

ERCIM 2013, London, 14-12-2013

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

Little & Rubin (2002), Shafer (1997)

Single imputation methods

Continuous variables: k-nearest neighbors, joint modeling: normal distribution, conditional modeling (van Bureen 1999): iterative regressions, etc.

Categorical variables: k-nn, joint modeling: log-linear model, latent class model (Vermunt, 2008), conditional modeling: iterative logistic regressions, etc.

Mixed data:

- General location model (Schaefer, 1997)
- Transform the categorical variables into dummy variables and deal as continuous variables (package Amelia)
- MICE (conditional multivariate imputation by chained equations, van Bureen 1999): a model must be specified for each variable - iterative linear and logistic regressions (package mice)

$\Rightarrow$ Random forests (Stekhoven & Bühlmann, 2011)
$\Rightarrow$ Principal component method (Audigier, Husson & Josse, 2013)

Iterative Random Forests imputation

1. Initial imputation: mean imputation - frequent category
   Sort the variables according to the amount of missing values
2. Fit a RF X_j^{obs} on the other variables X_{-j}^{obs}
   Predict X_j^{miss} using the trained RF on X_{-j}^{miss}
3. Cycling through variables
4. Repeat step 2 and 3 until convergence

$\Rightarrow$ Conditional modeling based on RF

Iterative Random Forests imputation

1. Initial imputation: mean imputation - frequent category
   Sort the variables according to the amount of missing values
2. Fit a RF X_j^{obs} on the other variables X_{-j}^{obs}
   Predict X_j^{miss} using the trained RF on X_{-j}^{miss}
3. Cycling through variables
4. Repeat step 2 and 3 until convergence

$\Rightarrow$ Conditional modeling based on RF

- number of trees/variable: 100
- number of variables randomly selected at each node: $\sqrt{p}$
- computational time (linear in the number of trees)
- number of iteration: 4-5

Iterative Random Forests imputation

⇒ Properties:

- Non-linear relations
- Complex interactions
- n«p
  (difficult with MICE: ridge regression per variable)
- OOB: approximation of the imputation error

⇒ Outperforms k-nn and MICE

PCA with missing values

$\Rightarrow$ PCA: least squares

$$\|\mathbf{X}_{n\times p} - \mathbf{U}_{n\times S}\mathbf{\Lambda}^{\frac{1}{2}}_{S\times S}\mathbf{V}'_{p\times S})\|^2$$

- $\mathbf{F} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ principal components - scores
- $\mathbf{V}$ principal axes - loadings

$\Rightarrow$ PCA with missing values: weighted least squares

$$\|\mathbf{W}_{n\times p} * (\mathbf{X}_{n\times p} - \mathbf{U}_{n\times S}\mathbf{\Lambda}^{\frac{1}{2}}_{S\times S}\mathbf{V}'_{p\times S})\|^2$$

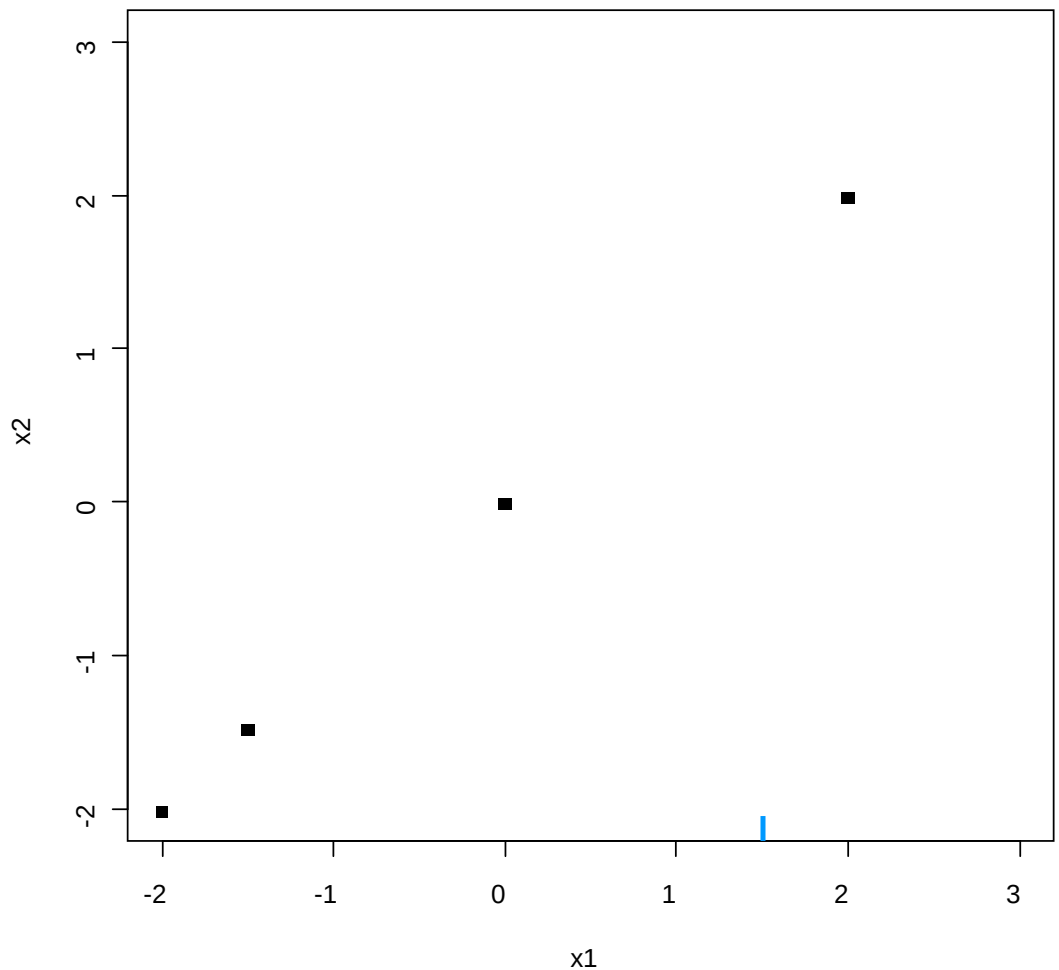
with $w_{ij} = 0$ if x_{ij} is missing, $w_{ij} = 1$ otherwise

Many algorithms: weighted alternating least squares (Gabriel & Zamir, 1979); 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

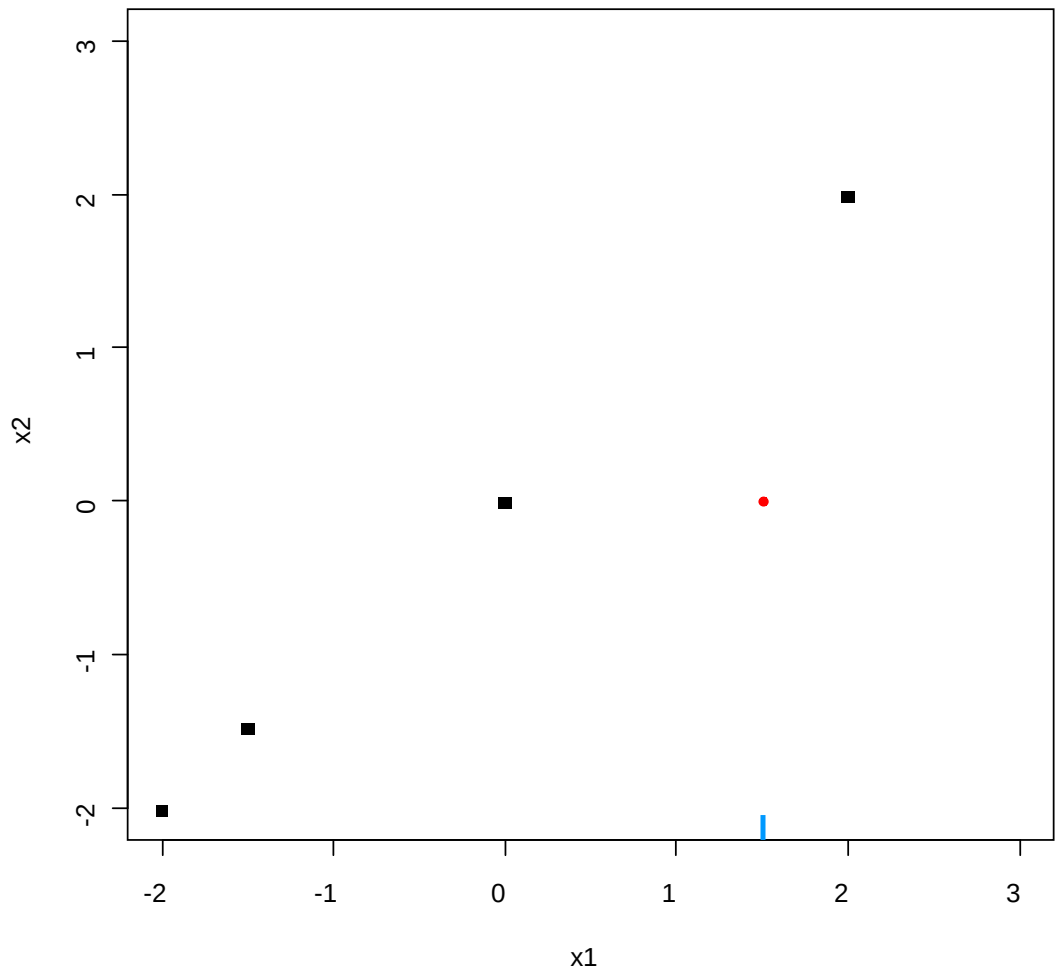


Iterative PCA algorithm

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



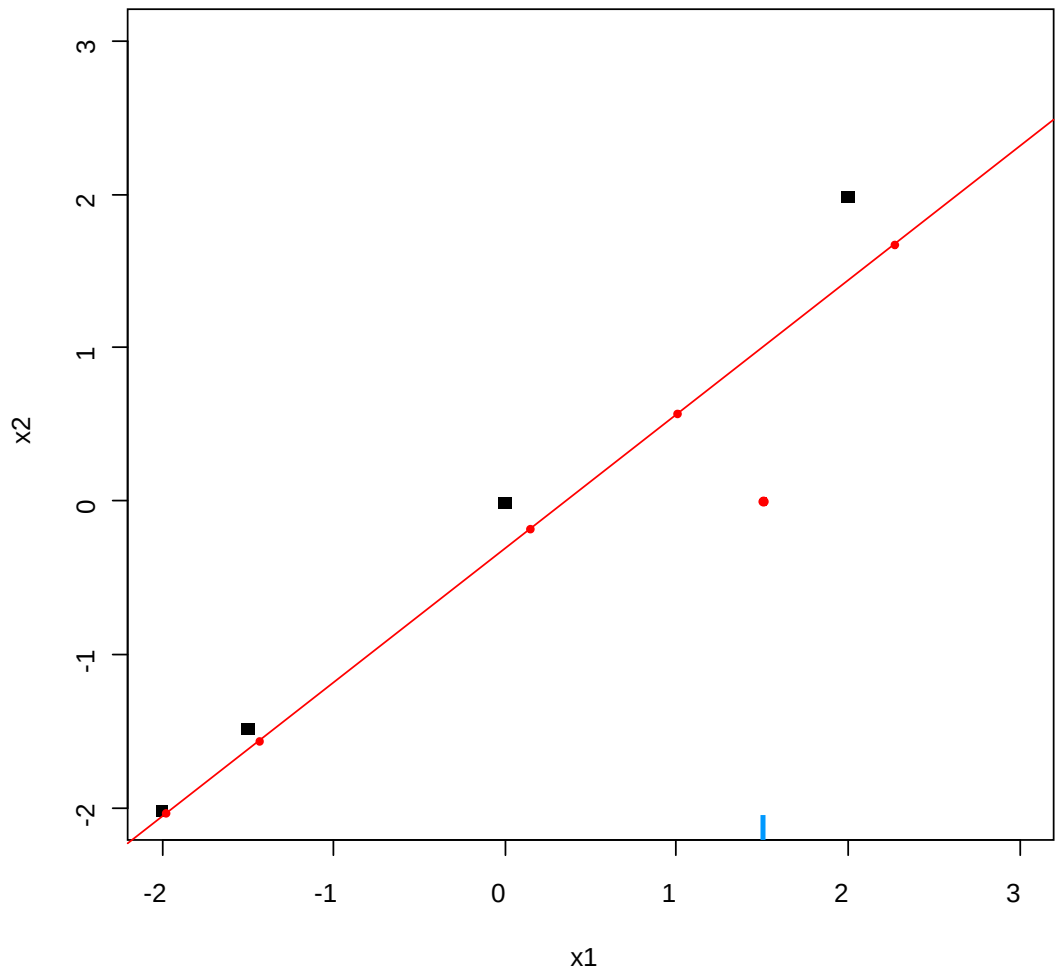
Iterative PCA algorithm

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

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



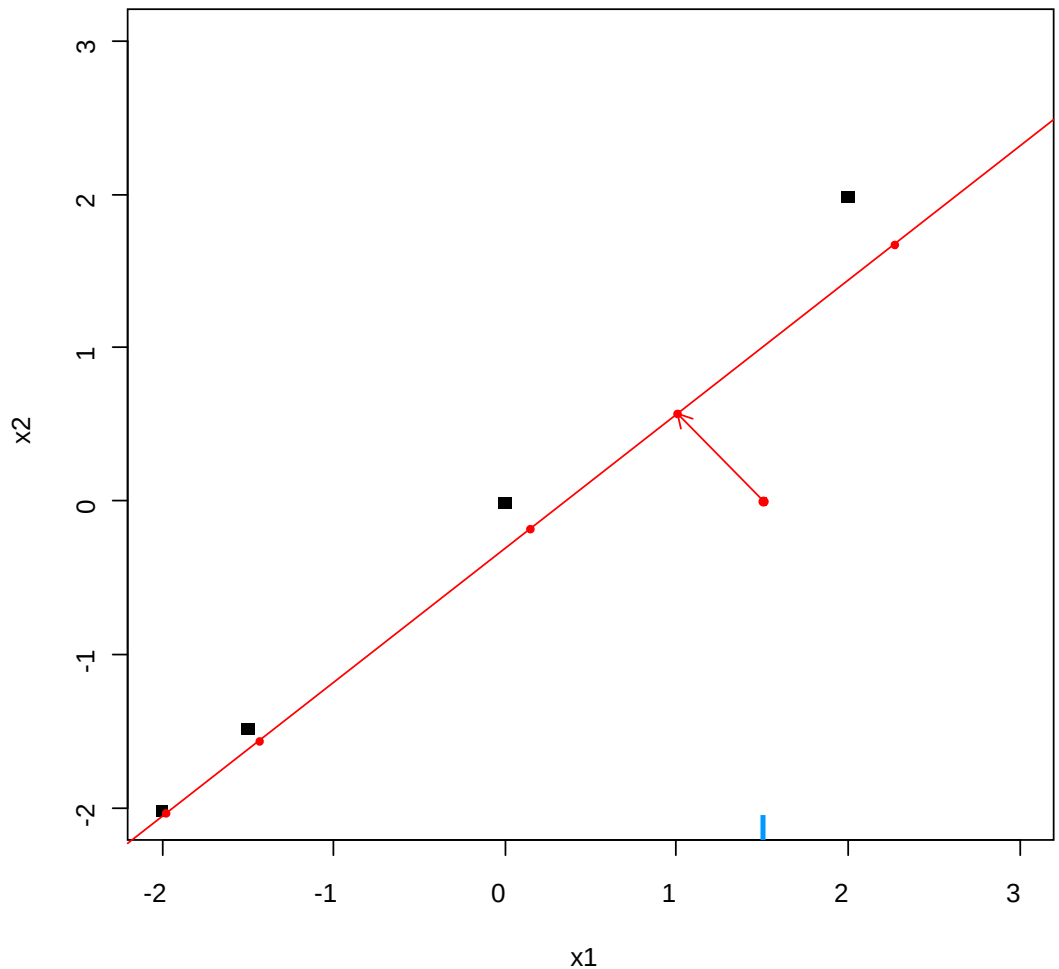
Iterative PCA algorithm

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

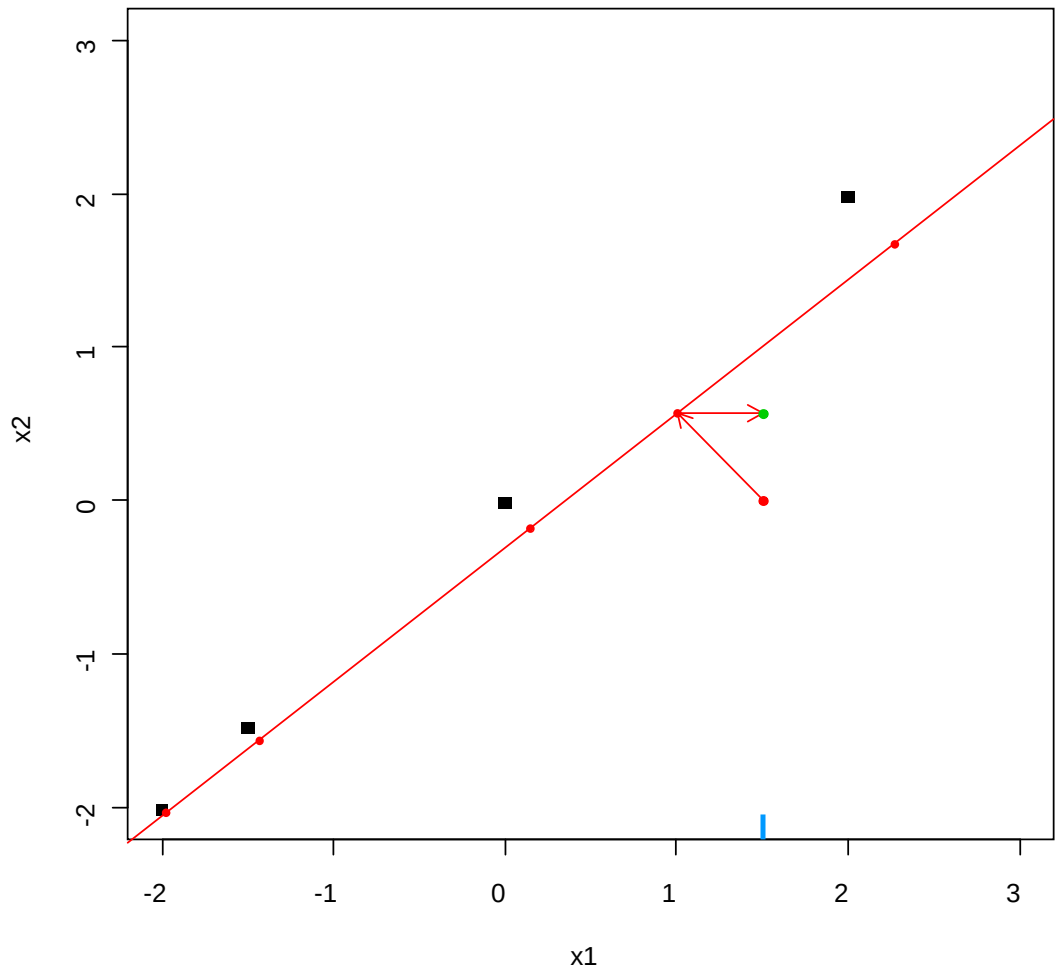
Imputation step: $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$

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

$\hat{x1}$	$\hat{x2}$
-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



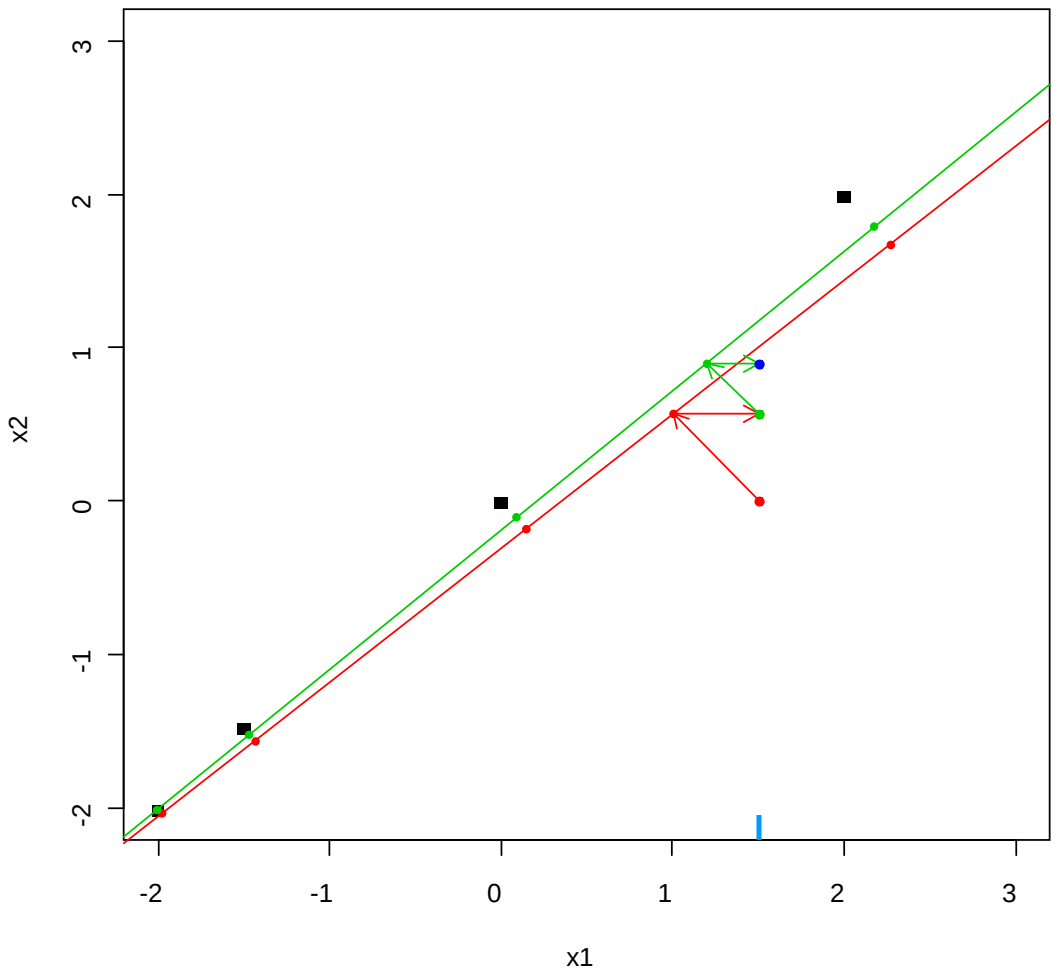
Iterative PCA algorithm

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

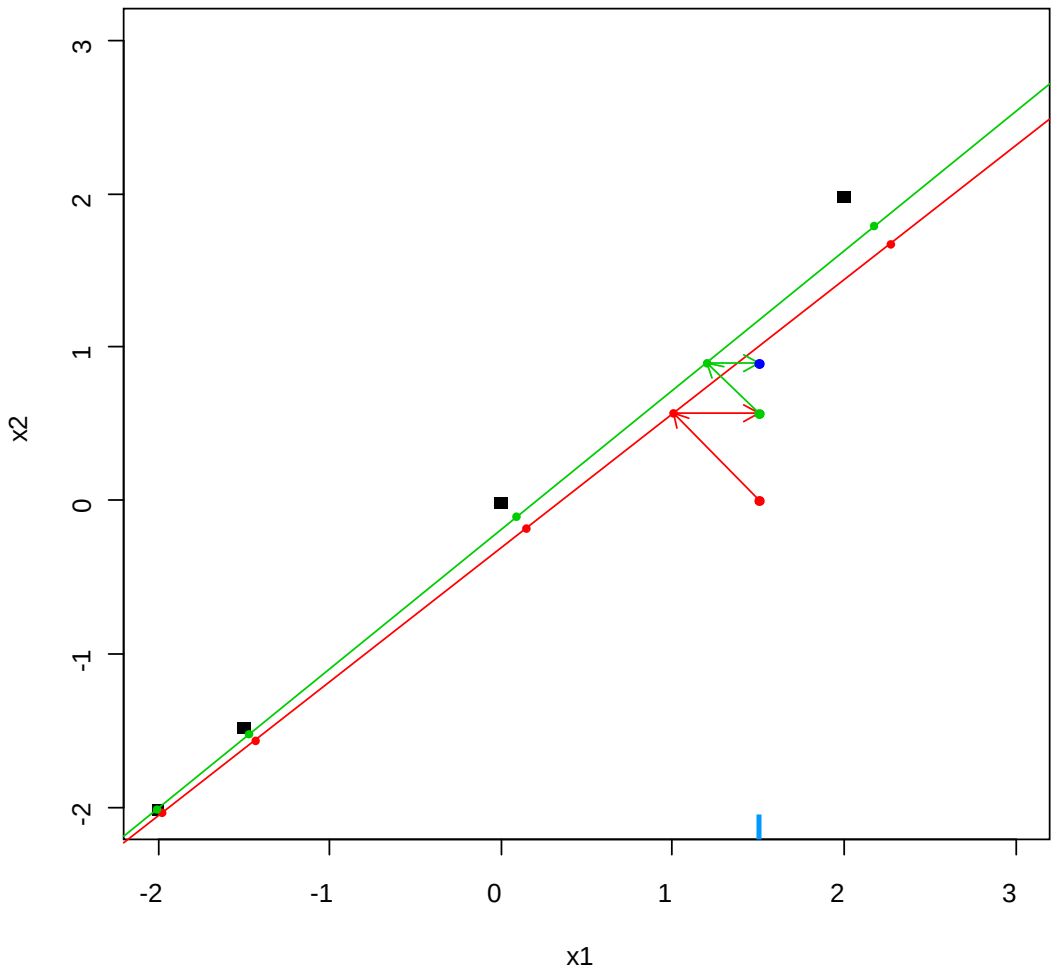
Imputation step: $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$

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

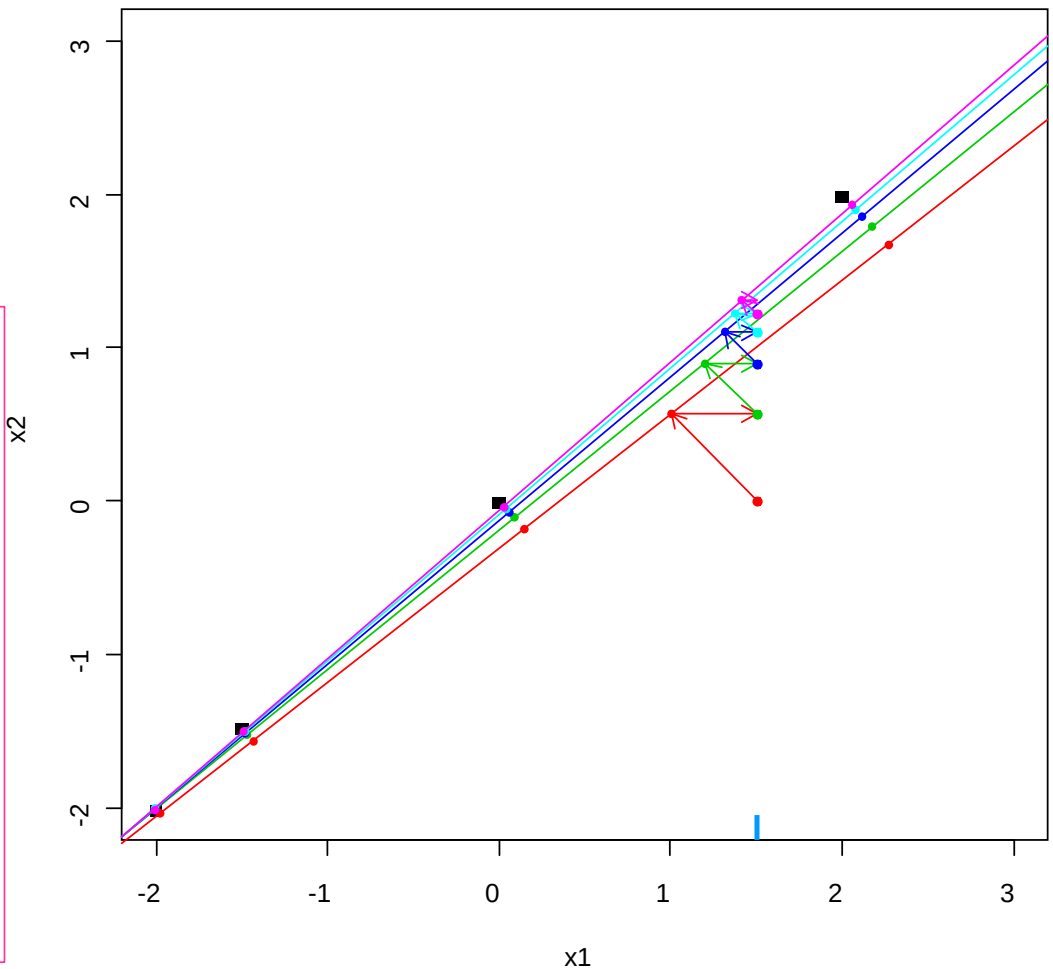
Iterate until convergence

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

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

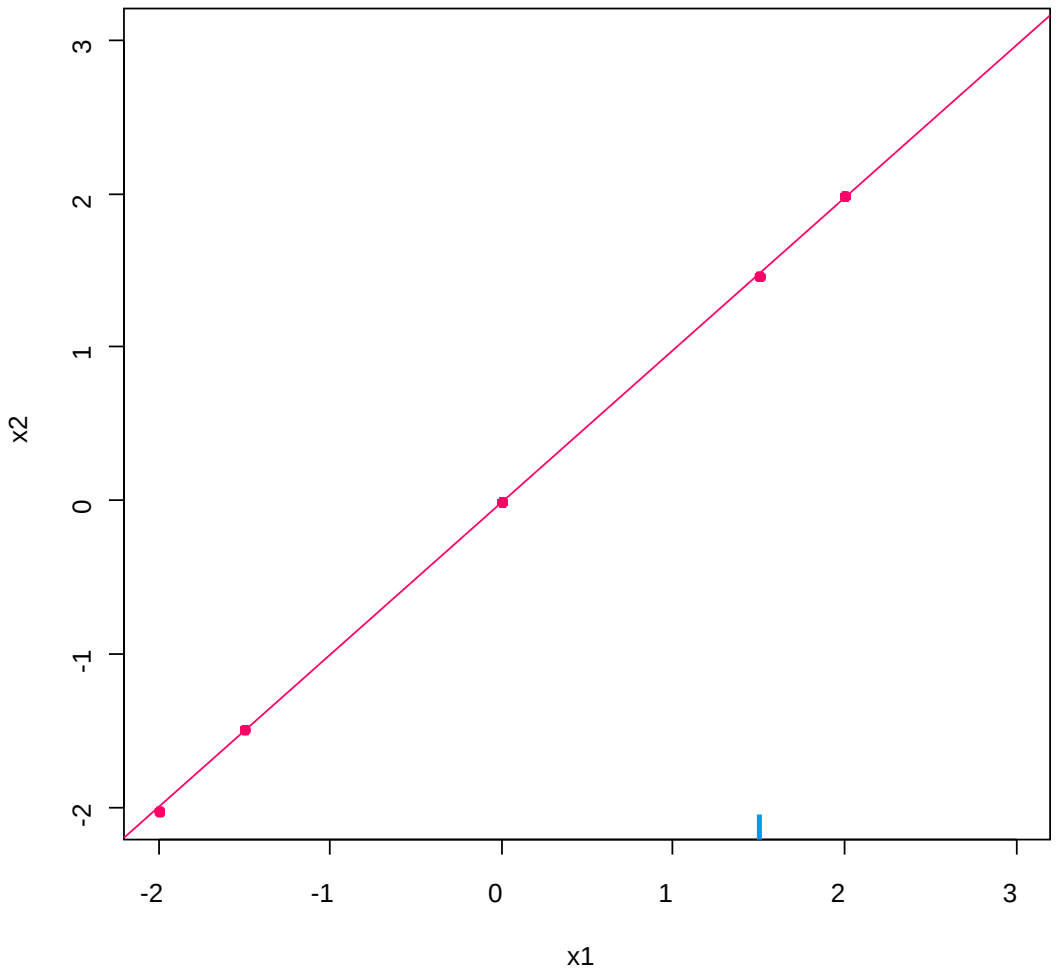
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Iterative PCA - 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 $\mathbf{X}^{\ell-1} \rightarrow (\mathbf{U}^\ell, \mathbf{\Lambda}^\ell, \mathbf{V}^\ell)$
   S dimensions are kept; $\hat{\mathbf{X}}^\ell = \mathbf{U}^\ell \mathbf{\Lambda}^{1/2^\ell} \mathbf{V}^{\ell\prime}$ (estimation)
   (b) $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$ (imputation)
3. Estimation and imputation are repeated until convergence

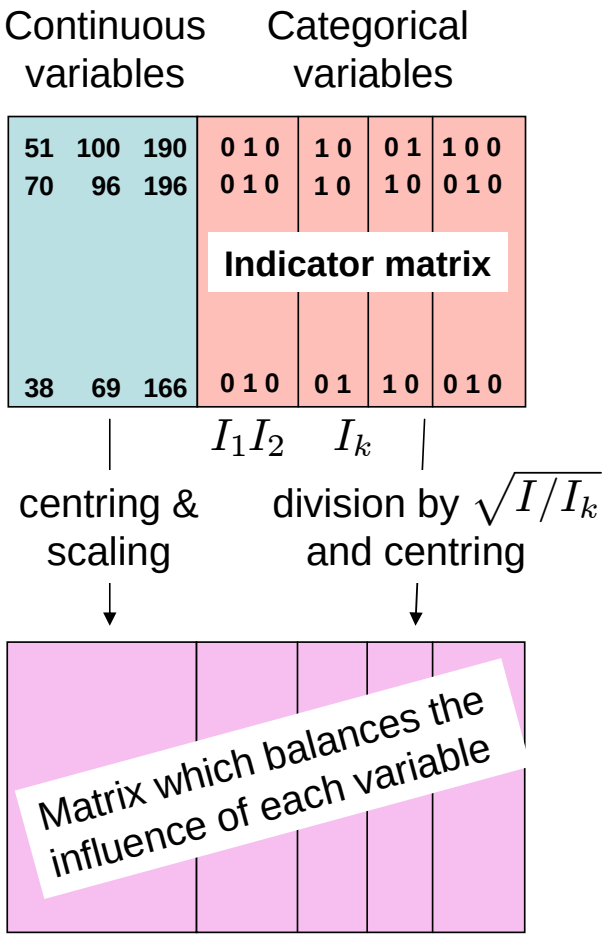
- The number of dimensions S has to be chosen *a priori*
- An imputation is performed during the algorithm

$\Rightarrow$ PCA can be seen as an imputation method

- Overfitting problems are dealt with a regularized algorithm

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:

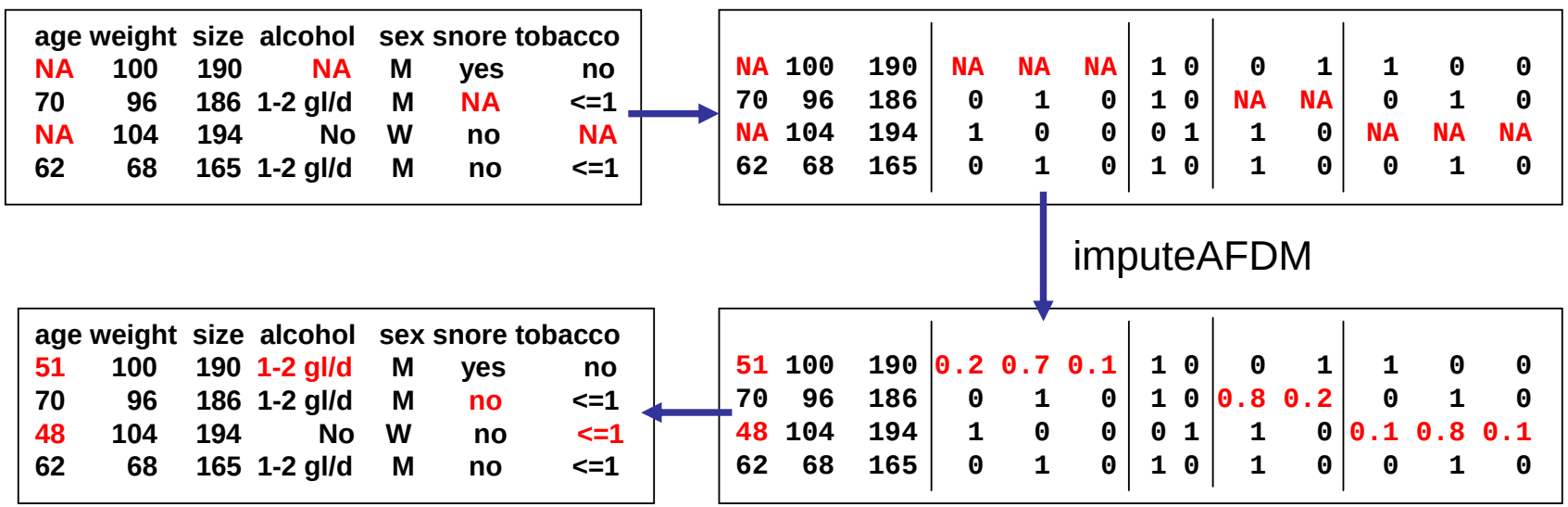
$$d^2(i,l) = \sum_{k=1}^{K_{cont}} (x_{ik} - x_{lk})^2 + \sum_{q=1}^{Q} \sum_{k=1}^{K_q} \frac{1}{l_{kq}} (x_{iq} - x_{lq})^2$$

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

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

Iterative FAMD algorithm

1. initialization: imputation mean (continuous) and proportion (dummy)
2. iterate until convergence
   (a) estimation: FAMD on the completed data $\Rightarrow \mathbf{U}, \mathbf{\Lambda}, \mathbf{V}$
   (b) imputation of the missing values with the model matrix
   (c) means, standard deviations and column margins are updated



$\Rightarrow$ Imputed values can be seen as degree of membership

Iterative FAMD

$\Rightarrow$ Properties:

- Imputation based on scores and loadings $\Rightarrow$ similarities between individuals and relationships between continuous and categorical variables
- Linear relationships
- Compared to a PCA on the (unweighted) indicator matrix, small categories are better imputed
- The number of dimensions is a tuning parameter

Simulations

- number of continuous - categorical variables
- number of categories, individuals/categories
- Signal to noise ratio
- 10%, 20% or 30% of missing values are chosen at random

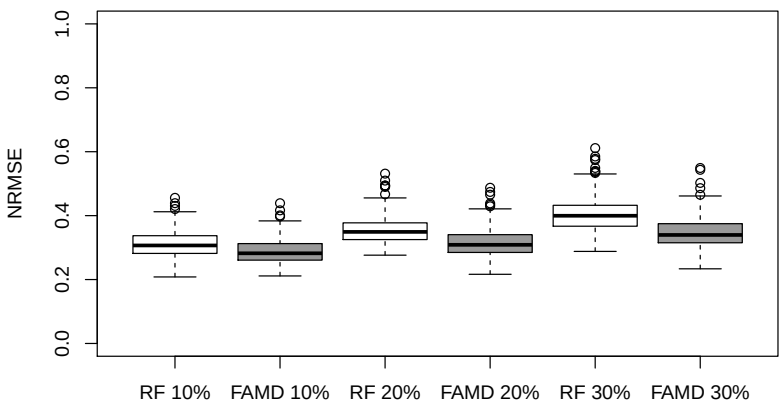
$\Rightarrow$ Criterion

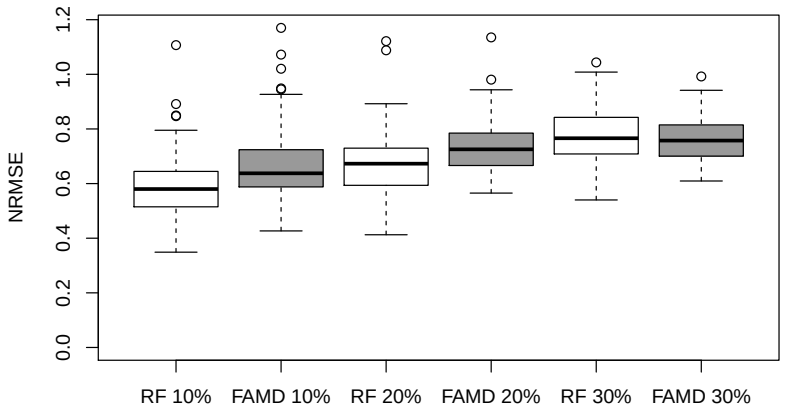
- for continuous data:

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

- for categorical data: proportion of falsely classified entries

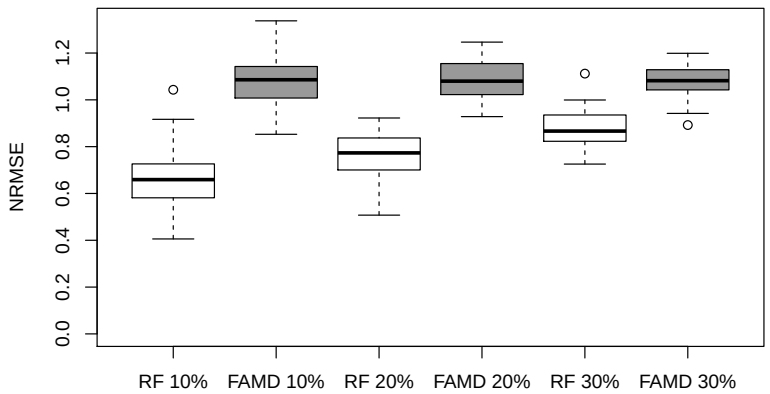
Linear - non-linear relations

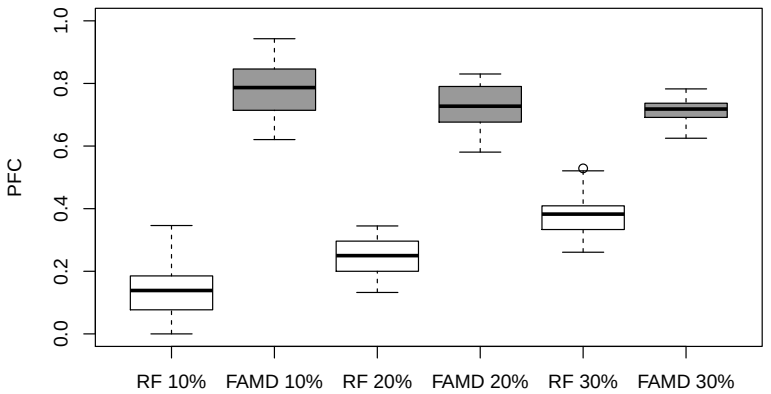




$\Rightarrow$ Solution FAMD: cut continuous variables into categories

Interactions





$\Rightarrow$ FAMD based on relationships between pairs of variables

$\Rightarrow$ The quality of imputation is poor - close mean imputation

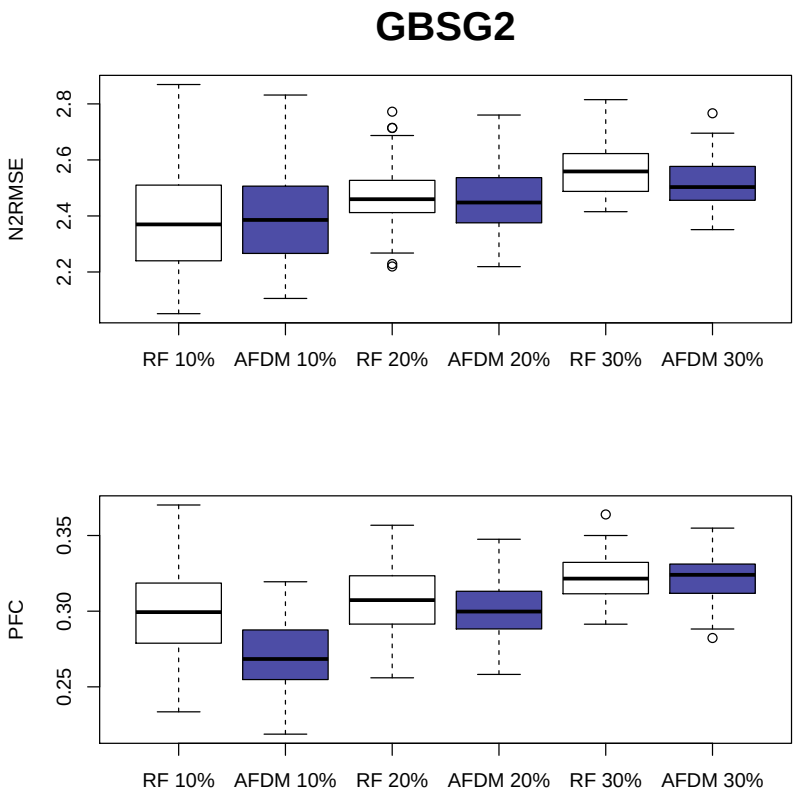
$\Rightarrow$ Solution FAMD: add a variable corresponding to interaction

Rares categories

Number of rows	f	FAMD	Random forest
100	10%	0.060	0.096
100	4%	0.082	0.173
1000	10%	0.042	0.041
1000	4%	0.060	0.071
1000	1%	0.074	0.167
1000	0.4%	0.107	0.241

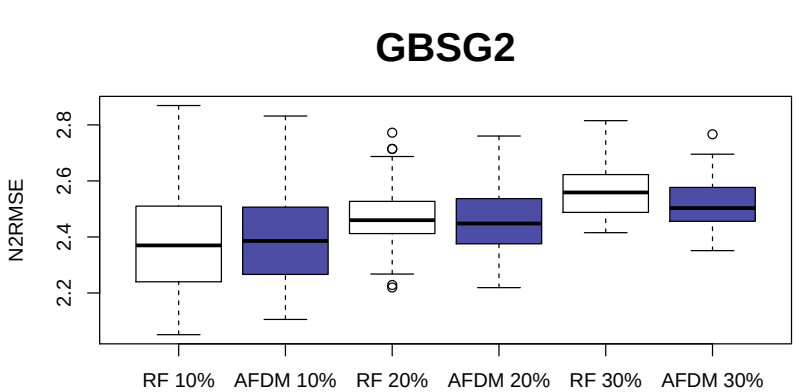
Comparison with random forest on real data sets

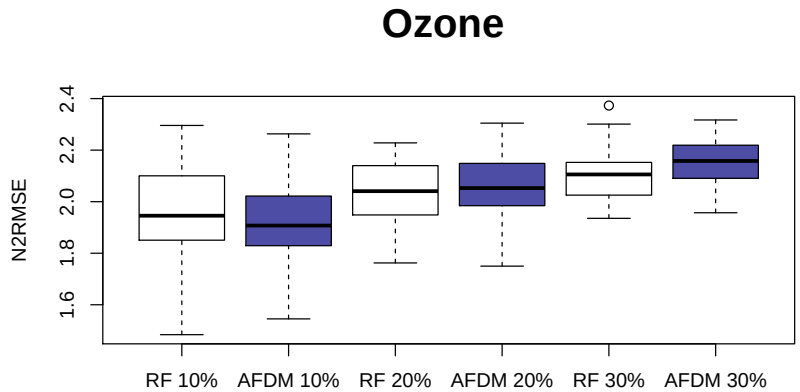
Imputations obtained with random forest & iterative algorithm

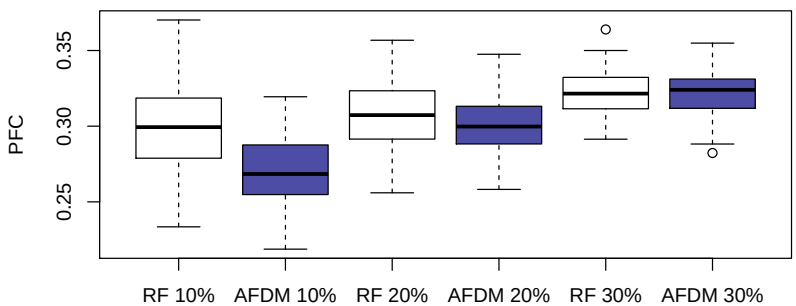


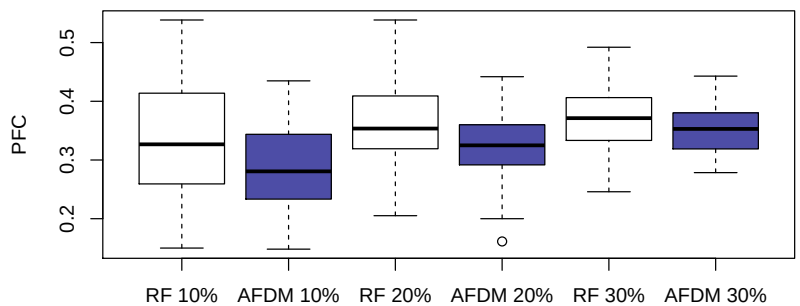
Comparison with random forest on real data sets

Imputations obtained with random forest & iterative algorithm









Conclusion

Random Forests:

- non-linear relationships between continuous variables
- interactions

$\Rightarrow$ no tuning parameters?
$\Rightarrow$ package `missForest`

Principal component:

- linear relationships
- categorical variables especially rare categories

$\Rightarrow$ tuning parameters: number of dimensions, cv? approximation?
$\Rightarrow$ package `missMDA`:

- handles missing values in PC methods (PCA, MCA, FAMD, MFA)
- impute continuous, categorical and mixed data

Perspectives

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