Multiple imputation with principal component methods

Vincent Audigier[1], François Husson[2], Julie Josse[3]

1. CNAM, MSDMA team, Paris
2. Agrocampus Ouest, Rennes
3. Ecole Polytechnique, Paris

Telecom ParisTech, November 16, 2017

Missing values

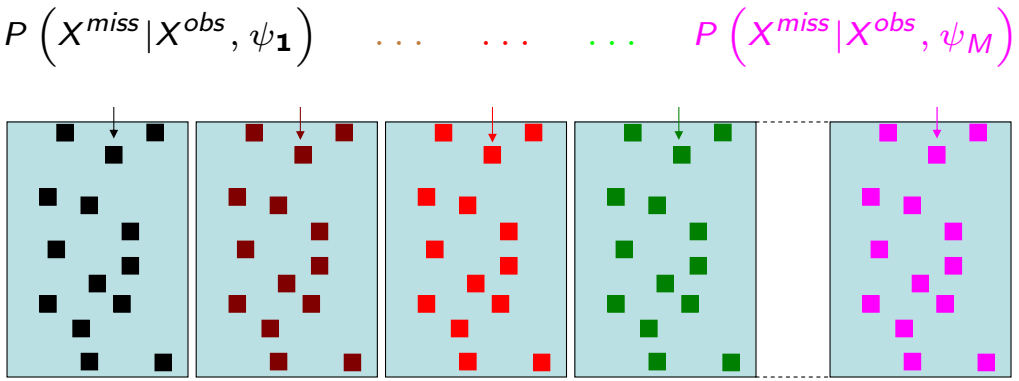
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

- Aim: inference on a quantity θ from incomplete data
  $\rightarrow$ point estimate $\hat{\theta}$ and associated variability T

- Complete-case analysis
- Likelihood approaches
- Multiple Imputation

Multiple imputation (Rubin, 1987)

1. Provide a set of M parameters to generate M plausible imputed data sets



2. Perform the analysis on each imputed data set: $\hat{\theta}_m, \widehat{Var}\left(\hat{\theta}_m\right)$

3. Combine the results: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$

$$T = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\theta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2$$

$\Rightarrow$ Aim: provide estimation of the parameters and of their variability

Principal component methods: aims

From multidimensional data, principal component methods:

- summarize
- describe
- visualize

Several objectives:

- identify the similarities between individuals
- highlight the relationships between variables
- describe some groups of individuals by a set of relevant variables

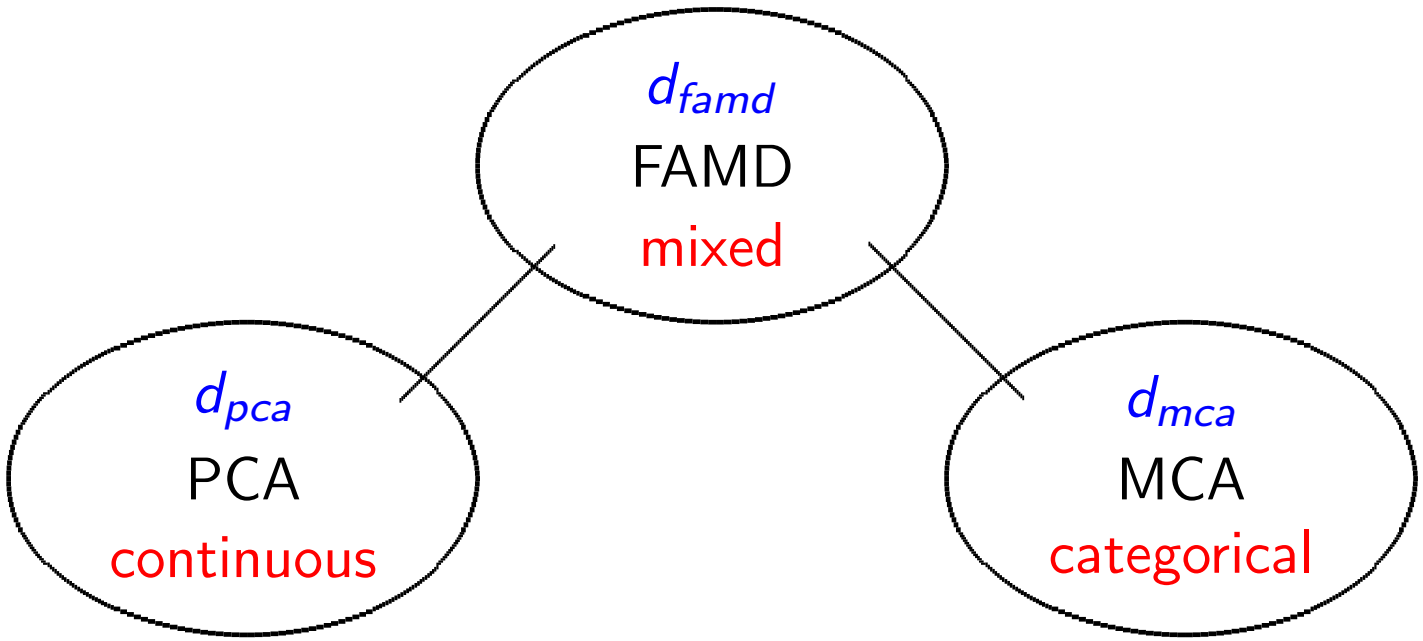
How does it work?

- The I individuals are seen as elements of $\mathbb{R}^K$

- A distance d on $\mathbb{R}^K$ to define proximities between individuals
- $Vect(v_1, ..., v_S)$ maximising the projected inertia

$\Rightarrow$ Dimensionality reduction methods

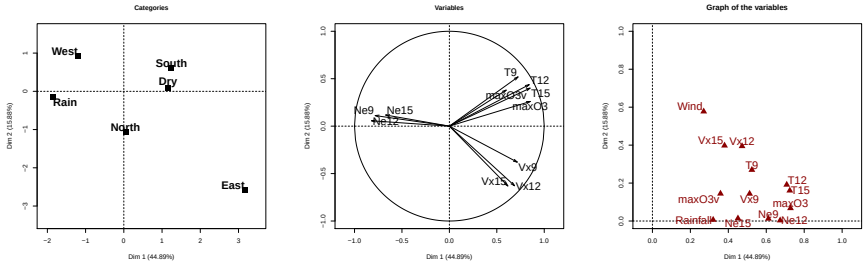
A set of methods

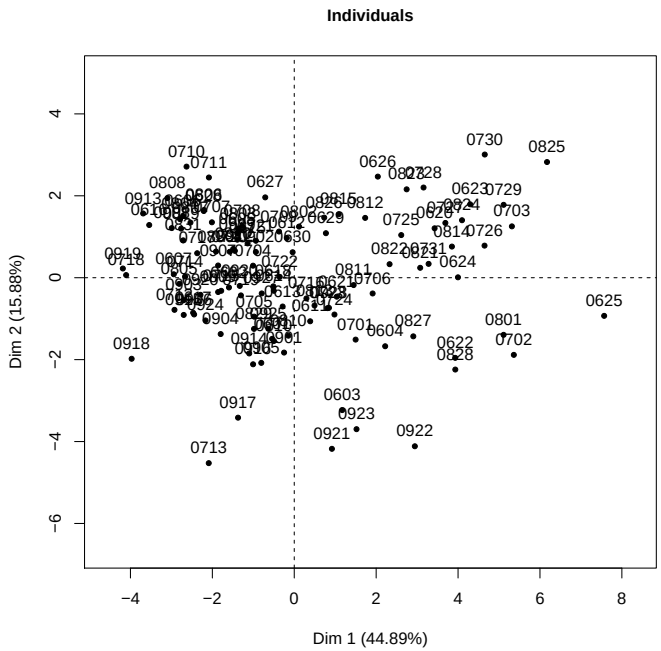


$$d^2_{famd} = d^2_{pca} + d^2_{mca}$$

FAMD

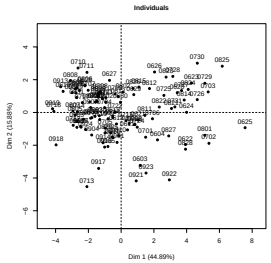
	Wind	Rainfall	maxO3	T9	T12	...
0602	North	Dry	82	17.0	18.4	...
0603	East	Dry	92	15.3	17.6	...
0604	North	Dry	114	16.2	19.7	...
0605	West	Dry	94	17.4	20.5	...
0606	West	Rain	80	17.7	19.8	...
...	...	...	...	...	...	...

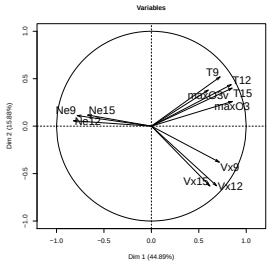


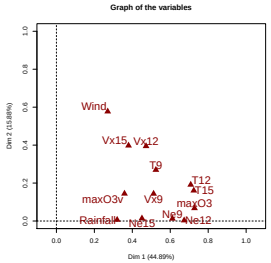


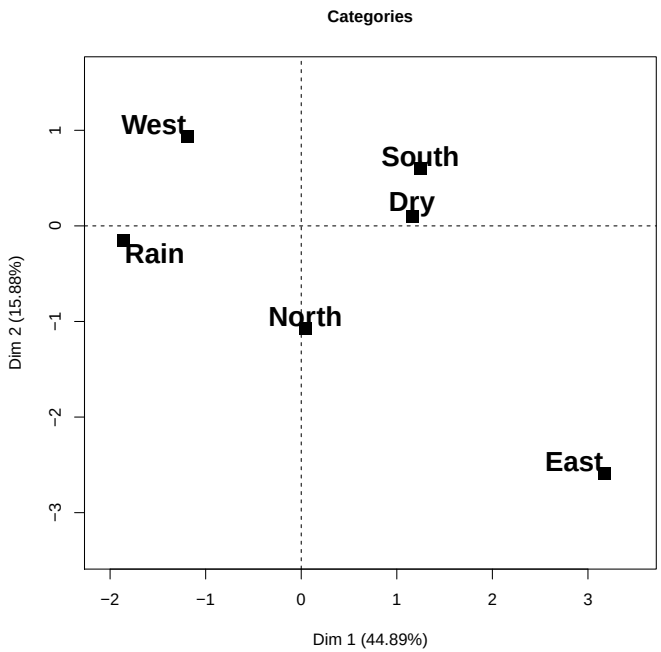
FAMD

	Wind	Rainfall	maxO3	T9	T12	...
0602	North	Dry	82	17.0	18.4	...
0603	East	Dry	92	15.3	17.6	...
0604	North	Dry	114	16.2	19.7	...
0605	West	Dry	94	17.4	20.5	...
0606	West	Rain	80	17.7	19.8	...
...	...	...	...	...	...	...



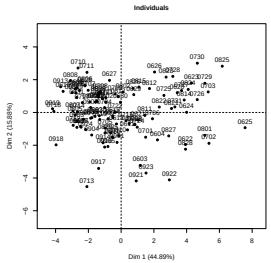


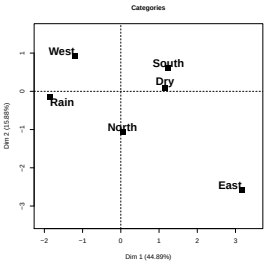


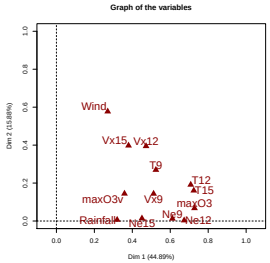


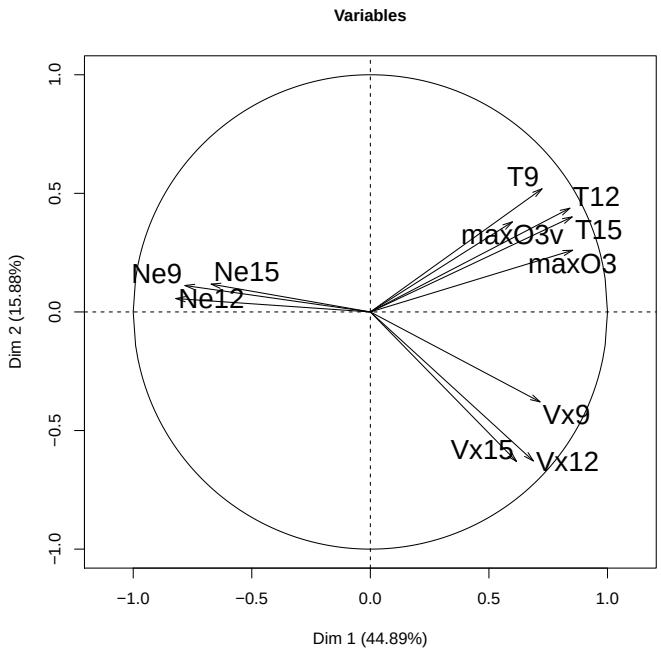
FAMD

	Wind	Rainfall	maxO3	T9	T12	...
0602	North	Dry	82	17.0	18.4	...
0603	East	Dry	92	15.3	17.6	...
0604	North	Dry	114	16.2	19.7	...
0605	West	Dry	94	17.4	20.5	...
0606	West	Rain	80	17.7	19.8	...
...	...	...	...	...	...	...



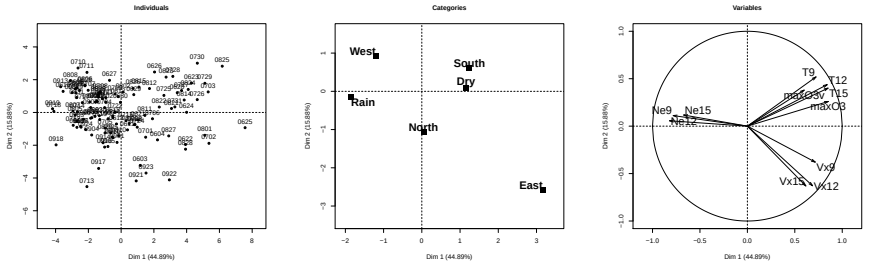


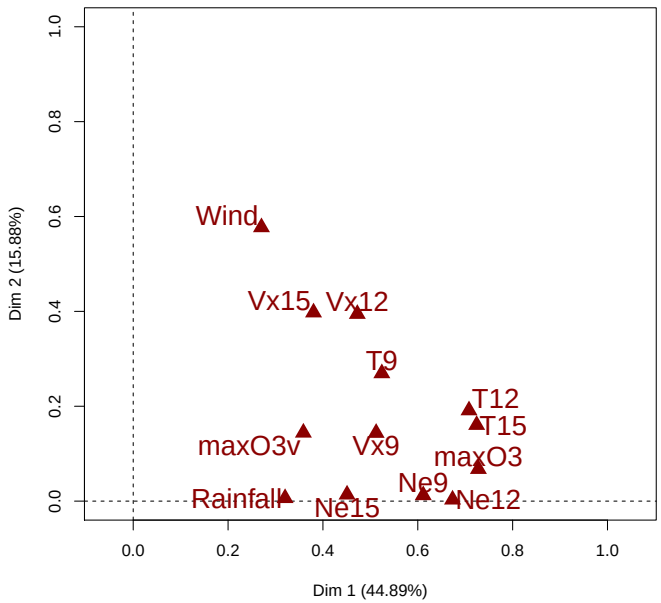




FAMD

	Wind	Rainfall	maxO3	T9	T12	...
0602	North	Dry	82	17.0	18.4	...
0603	East	Dry	92	15.3	17.6	...
0604	North	Dry	114	16.2	19.7	...
0605	West	Dry	94	17.4	20.5	...
0606	West	Rain	80	17.7	19.8	...
...	...	...	...	...	...	...





Why principal component methods for MI?

1. Lack of imputation models for non-continuous data
   - mixed data: various relationships between variables
   - categorical data: combinatorial explosion

   A large range of models

2. Imputation models with too many parameters:
   - overfitting
   - storage issue

   A small number of parameters

3. Models based on regressions:
   - number of individuals less than the number of variables
   - collinearity

   No inversion issues

How to perform FAMD?

FAMD can be seen as the SVD of $\mathbf{X}$ with weights for

- the continuous variables and categories: $(\mathbf{D}_\Sigma)^{-1}$
- the individuals: $\frac{1}{I}\mathbb{1}_I$

$$\longrightarrow SVD\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$$

$$X = \begin{array}{|ccc||ccc|ccc|}
\hline
11.04 & \dots & 2.07 & 1 & 0 & \dots & 1 & 0 & 0 \\
10.76 & \dots & 1.86 & 1 & 0 & \dots & 1 & 0 & 0 \\
11.02 & \dots & 2.04 & 1 & 0 & \dots & 1 & 0 & 0 \\
11.02 & \dots & 1.92 & 0 & 1 & \dots & 0 & 1 & 0 \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
11.06 & & 2.01 & 0 & 1 & & 0 & 0 & 1 \\
10.95 & & 1.67 & 0 & 1 & & 0 & 1 & 0 \\
\hline
\end{array}$$

$$D_\Sigma = \begin{array}{|cc|}
\hline
\sigma_{x_1} & \\
 \ddots & \mathbf{0} \\
 & \sigma_{x_k} \\
\hline
 & I_{k+1} \\
\mathbf{0} & \ddots \\
 & I_K \\
\hline
\end{array}$$

How to perform FAMD?

$$SVD\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \tfrac{1}{I}\mathbb{1}_I\right) \longrightarrow \begin{array}{ll} \mathbf{X}_{I\times K} & = \mathbf{U}_{I\times K}\Lambda^{1/2}_{K\times K}\mathbf{V}^\top_{K\times K} \\ \text{with} & \mathbf{U}^\top\left(\tfrac{1}{I}\mathbb{1}_I\right)\mathbf{U} = \mathbb{1}_K \\ & \mathbf{V}^\top\mathbf{D}_\Sigma^{-1}\mathbf{V} = \mathbb{1}_K \end{array}$$

- principal components: $\hat{\mathbf{F}}_{I\times S} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}$
- loadings: $\hat{\mathbf{V}}^\top_{K\times S}$
- fitted matrix: $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^\top_{K\times S}$

$$\| \hat{\mathbf{X}} - \mathbf{X} \|^2_{\mathbf{D}_\Sigma^{-1}\otimes\frac{1}{I}\mathbb{1}} = tr\left(\left(\hat{\mathbf{X}} - \mathbf{X}\right)\mathbf{D}_\Sigma^{-1}\left(\hat{\mathbf{X}} - \mathbf{X}\right)^\top \tfrac{1}{I}\mathbb{1}_I\right)$$

minimized under the constraint of rank S

FAMD with missing values

$\Rightarrow$ FAMD: minimize

$$\|\mathbf{X}_{I\times K} - \hat{\mathbf{X}}_{I\times K}\|^2_{\mathbf{D}_\Sigma^{-1}\otimes\frac{1}{I}\mathbb{1}}$$

$\Rightarrow$ FAMD with missing values:

$$\|\mathbf{W}_{I\times K} * (\mathbf{X}_{I\times K} - \hat{\mathbf{X}}_{I\times K})\|^2_{\mathbf{D}_\Sigma^{-1}\otimes\frac{1}{I}\mathbb{1}}$$

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

Many algorithms developed for PCA such as NIPALS (Christoffersson, 1970) or iterative PCA (Kiers, 1997)

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation by mean/proportion
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (b) imputation of the missing values with
   $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{K\times S}$
   (c) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

$\rightarrow$

NA	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	NA	1	0	...	NA	NA	NA
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	NA	NA		0	0	1
NA		1.67	0	1		0	1	0

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation by mean/proportion
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (b) imputation of the missing values with
   $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{K\times S}$
   (c) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

$\rightarrow$

11.01	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	1.89	1	0	...	0.61	0.19	0.20
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.32	0.68		0	0	1
11.01		1.67	0	1		0	1	0

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation by mean/proportion
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (b) imputation of the missing values with
   $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{K\times S}$
   (c) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

$\rightarrow$

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	2.04	1	0	...	0.81	0.05	0.14
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.25	0.75		0	0	1
10.95		1.67	0	1		0	1	0

Single imputation with FAMD (Audigier et al., 2016)

Iterative FAMD algorithm:

1. initialization: imputation by mean/proportion
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (b) imputation of the missing values with
   $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{K\times S}$
   (c) $\mathbf{D}_\Sigma$ is updated

11.04	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	**2.04**	A	...	**A**
11.02	...	1.92	B	...	B
11.06		2.01	**B**	...	C
10.95		1.67	B	...	B

$\leftarrow$

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	**2.04**	1	0	...	**0.81**	**0.05**	**0.14**
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	**0.25**	**0.75**		0	0	1
10.95		1.67	0	1		0	1	0

$\Rightarrow$ the imputed values can be seen as degree of membership

Single imputation with FAMD (Audigier et al., 2016)

Iterative FAMD algorithm:

1. initialization: imputation by mean/proportion
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (b) imputation of the missing values with
   $\hat{\mathbf{X}}_{I\times K} = \hat{\mathbf{U}}_{I\times S} f\left(\hat{\Lambda}^{1/2}_{S\times S}\right)\hat{\mathbf{V}}^{\top}_{K\times S}$ $\qquad f\left(\lambda_s^{1/2}\right) = \hat{\lambda}_s^{1/2} - \frac{\hat{\sigma}^2}{\hat{\lambda}_s^{1/2}}$
   (c) $\mathbf{D}_\Sigma$ is updated

11.04	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	2.04	A	...	A
11.02	...	1.92	B	...	B
11.06		2.01	B	...	C
10.95		1.67	B	...	B

$\leftarrow$

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	2.04	1	0	...	0.81	0.05	0.14
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.25	0.75		0	0	1
10.95		1.67	0	1		0	1	0

$\Rightarrow$ the imputed values can be seen as degree of membership

Imputation with Random Forests (Bühlmann & Stekhoven, 2011)

A random forest:

- Each tree is built from a subset of observations
- Fitted values are pooled

Imputation with RF:

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

Implemented: `R` package `missForest` (Stekhoven)

Simulation study

Several data sets

- relationships between variables
- number of categories
- percentage of missing values (10%,20%,30%)

Criteria:

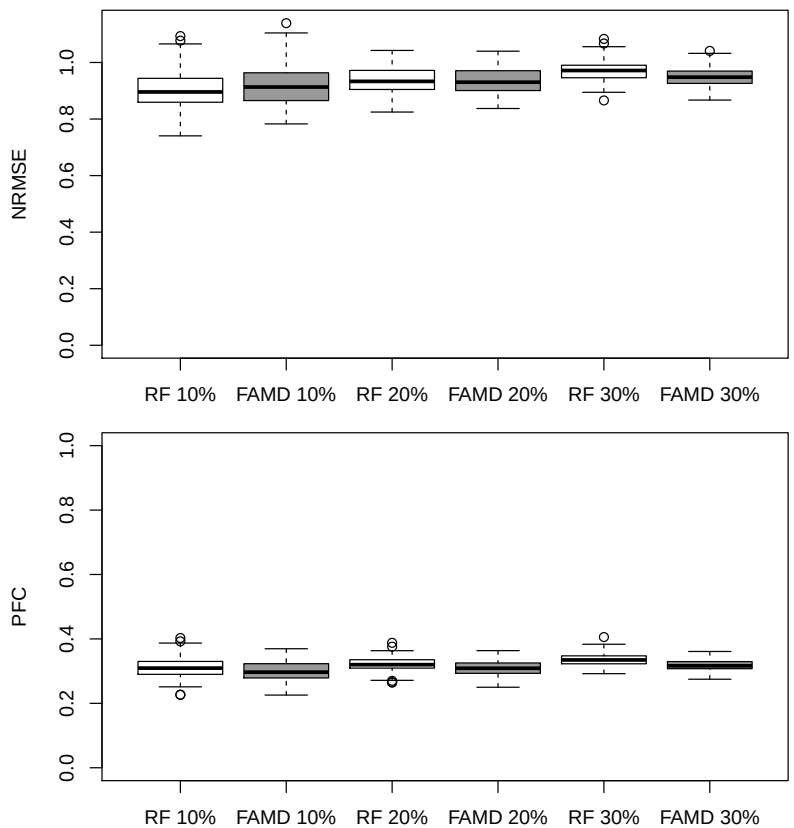
- for continuous data:

$$NRMSE = \sqrt{\frac{\sum_k \sum_{i=1}^{I} (1 - w_{ik}) \left(\frac{x_{ik} - \hat{x}_{ik}}{\hat{\sigma}_{\mathbf{x}_k}} \right)^2}{\sum_k \sum_{i=1}^{I} (1 - w_{ik})}}$$

- for categorical data: proportion of falsely classified entries

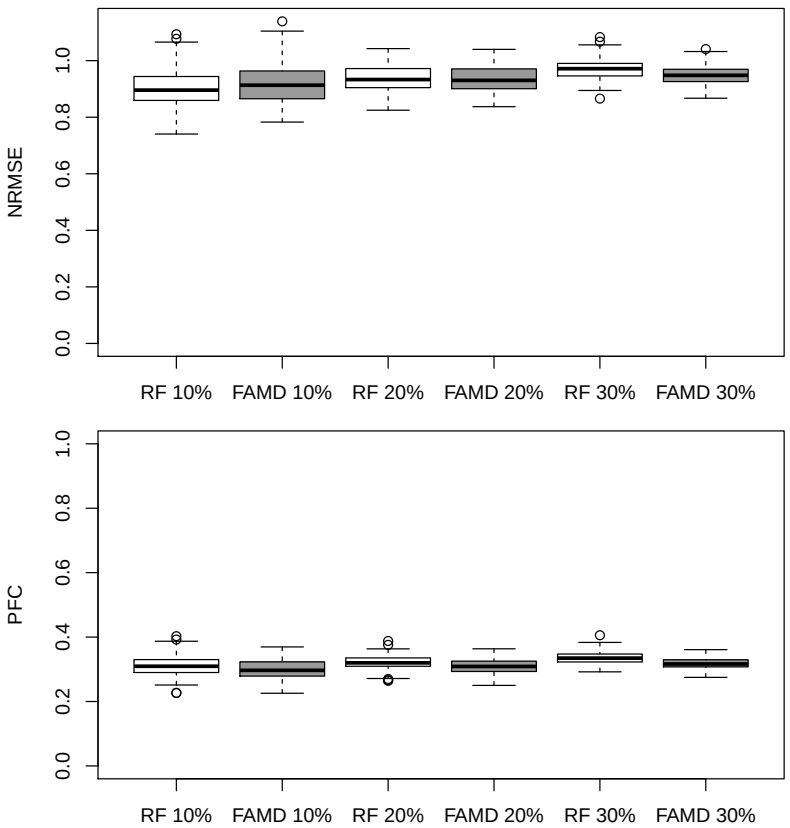
Comparisons from real data sets

GBSG2

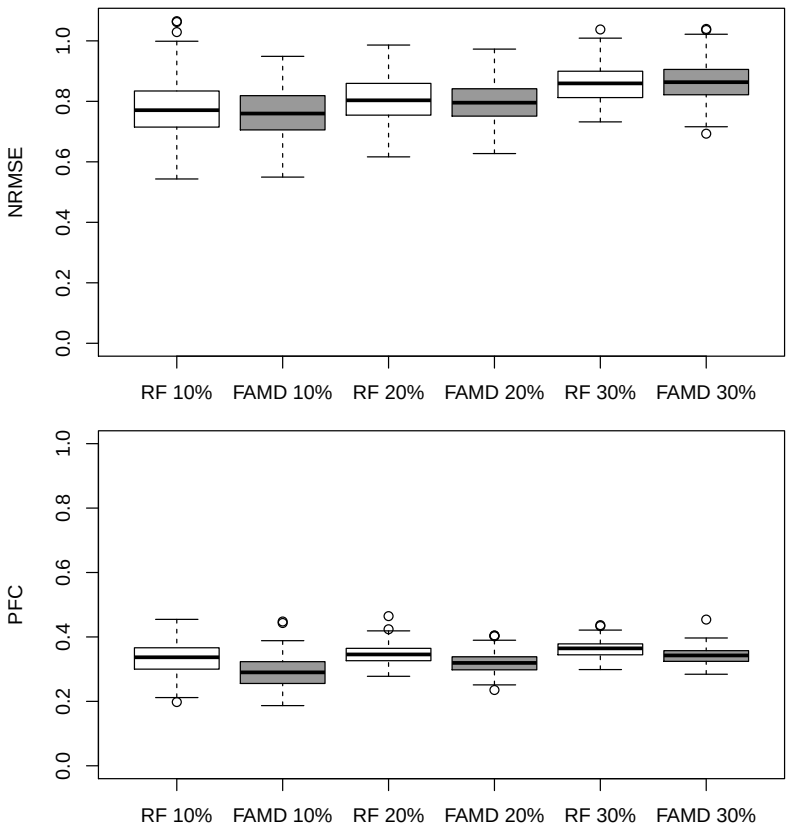


Comparisons from real data sets

GBSG2



Tips



Simulation results

Single imputation with FAMD shows a high quality of prediction compared to random forests (Stekhoven and Bühlmann, 2012)

- on real data
- when the relationships between continuous variables are linear
- for rare categories
- with MAR/MCAR mechanism

Can impute mixed, continuous or categorical data

Simulation results

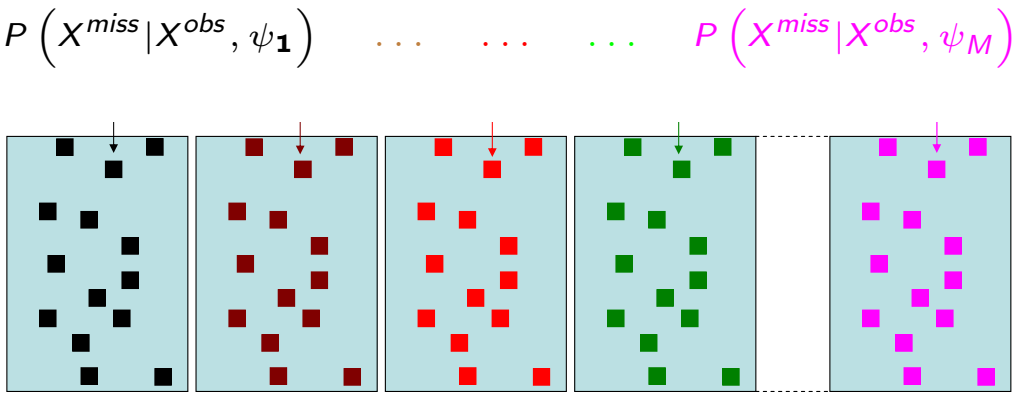
Single imputation with FAMD shows a high quality of prediction compared to random forests (Stekhoven and Bühlmann, 2012)

- on real data
- when the relationships between continuous variables are linear
- for rare categories
- with MAR/MCAR mechanism

Can impute mixed, continuous or categorical data

But a single imputation method only

From single imputation to multiple imputation



1. Reflect the variability on the parameters of the imputation model

   $\rightarrow \left(\left(\hat{\mathbf{U}}_{I\times S},\hat{\Lambda}^{1/2}_{S\times S},\hat{\mathbf{V}}^{\top}_{K\times S}\right)_1,\ldots,\left(\hat{\mathbf{U}}_{I\times S},\hat{\Lambda}^{1/2}_{S\times S},\hat{\mathbf{V}}^{\top}_{K\times S}\right)_M\right)$

   Bayesian or Bootstrap

2. Add a disturbance on the prediction by $\hat{X}_m = \hat{\mathbf{U}}_m\hat{\Lambda}^{1/2}_m\hat{\mathbf{V}}^{\top}_m$

   $\rightarrow$ need to distinguish continuous and categorical data

Stochastic single imputation

π_b	0.4
π_a	0.6

$\pi_{b\|A}$	0.2
$\pi_{a\|A}$	0.8
$\pi_{a\|B}$	0.4
$\pi_{b\|B}$	0.6

$\rightarrow$

V_1	V_2
A	a
B	b
B	a
B	b
$\vdots$	$\vdots$

$\rightarrow$

V_1	V_2
A	a
B	NA
B	a
B	NA
$\vdots$	$\vdots$

MCA majority

$\pi_{b\|A}$	0.14
$\pi_{a\|A}$	0.86
$\pi_{a\|B}$	0.27
$\pi_{b\|B}$	0.73

$cov_{95\%}(\pi_b) = 51.5$

MCA draw

$\pi_{b\|A}$	0.18
$\pi_{a\|A}$	0.82
$\pi_{a\|B}$	0.41
$\pi_{b\|B}$	0.59

$cov_{95\%}(\pi_b) = 89.9$

$\Rightarrow$ Standard errors of the parameters $(\hat{\sigma}_{\hat{\pi}_b})$ calculated from the imputed data set are underestimated

Multiple imputation with MCA (Audigier et al., 2017)

1. Variability of the parameters of MCA $(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{K\times S})$ using a non-parametric bootstrap:
   - define M weightings $(R_m)_{1\leq m\leq M}$ for the individuals
   - estimate MCA parameters using SVD of $\left(\mathbf{X}, \frac{1}{K}(\mathbf{D}_\Sigma)^{-1}, R_m\right)$
2. Imputation:

$\hat{\mathbf{X}}_1$

1	0	...	1	0
1	0	...	1	0
1	0	...	**0.81**	**0.19**
0.25	**0.75**		0	1
0	1		0	1

$\hat{\mathbf{X}}_2$

1	0	...	1	0
1	0	...	1	0
1	0	...	**0.60**	**0.40**
0.26	**0.74**		0	1
0	1		0	1

...

$\hat{\mathbf{X}}_M$

1	0	...	1	0
1	0	...	1	0
1	0	...	**0.74**	**0.16**
0.20	**0.80**		0	1
0	1		0	1

Draw categories from the values of $\left(\hat{\mathbf{X}}_m\right)_{1\leq m\leq M}$

A	...	A
A	...	A
A	...	**B**
B	...	C
B	...	B

A	...	A
A	...	A
A	...	**A**
B	...	C
B	...	B

...

A	...	A
A	...	A
A	...	**B**
B	...	C
B	...	B

Properties

MI for categorical data is challenging for a moderate number of variables

- estimation issues
- storage issues

MCA address the categorical data challenge by

- requiring a small number of parameters
- preserving the essential data structure
- using a regularisation strategy

MIMCA can be applied on various data sets

- small or large number of variables/categories
- small or large number of individuals

MI using the loglinear model (Schafer, 1997)

- Hypothesis on $X = (x_{ijk})_{i,j,k}$: $X|\psi \sim \mathcal{M}(n, \psi)$

$$log(\psi_{ijk}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

1. Variability of the parameter ψ: Bayesian formulation
2. Imputation using the set of M parameters

- Implemented: `R` package `cat` (J.L. Schafer)

Properties:

- Captures all the data relationships
- A number of parameters very large $\rightarrow$ fails on large data sets

MI using a latent class model (Si and Reiter, 2013)

- Hypothesis: $\mathbb{P}(X = (x_1, \ldots, x_K); \psi) = \sum_{\ell=1}^{L} \left(\psi_\ell \prod_{k=1}^{K} \psi_{x_k}^{(\ell)} \right)$

1. Variability of the parameters ψ_L and ψ_X : Bayesian formulation
2. Imputation using the set of M parameters

- Implemented: `R` package `mi` (Gelman *et al.*)

Properties:

- Local independence assumption
- Captures complex relationships
- A small number of parameters

Conditional modelling (van Buuren, 2006)

- A recent one: one **random forest**/variable (Doove *et al.*, 2014)

  Properties:
  - non-parametric modelling
  - captures complex relationships between variables

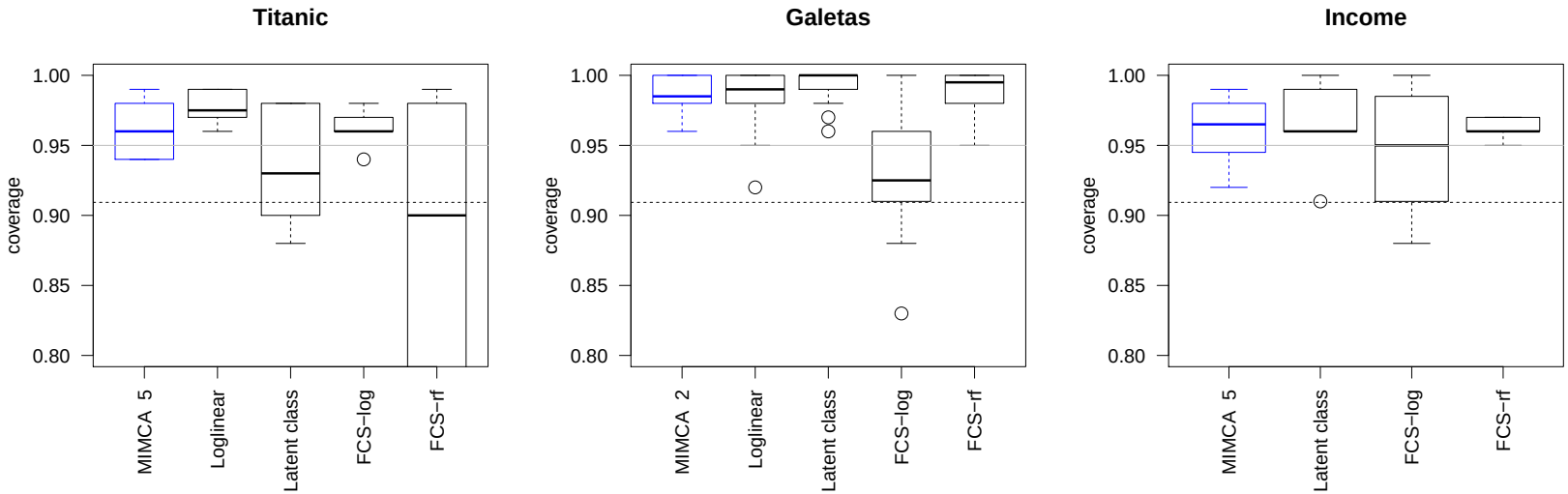
- A standard one: one **logistic regression** model/variable without interaction

  Properties: captures relationships between pairs of variables

Simulations from real data sets

- Quantities of interest: θ = parameters of a logistic model
- Simulation design (repeated 200 times)
  - the real data set is considered as a population
  - drawn one sample from the data set
  - generate 20% of missing values
  - multiple imputation using $M = 5$ imputed arrays
- Criteria
  - bias
  - CI width, coverage
- Comparison with :
  - JM: log-linear model, latent class model
  - FCS: logistic regression, random forests

Results - Inference



	Titanic	Galetas	Income
Number of variables	4	4	14
Number of categories	≤ 4	≤ 11	≤ 9

Results - Time

	Titanic	Galetas	Income
MIMCA	2.750	8.972	**58.729**
Loglinear	0.740	4.597	NA
Latent class model	10.854	17.414	143.652
FCS logistic	4.781	38.016	881.188
FCS forests	265.771	112.987	6329.514

Table: Time consumed in second

	Titanic	Galetas	Income
Number of individuals	2201	1192	6876
Number of variables	4	4	14

Conclusion

MI methods using dimensionality reduction method

- captures the relationships between variables
- captures the similarities between individuals
- requires a small number of parameters

Address some imputation issues:

- can be applied on various data sets
- provide correct inferences for analysis model based on relationships between pairs of variables

Available in the R package missMDA, with a user guide on http://vincentaudigier.weebly.com/links.html

Perspectives

- MI for mixed data using FAMD
- uncertainty on the number of dimensions S
- promising methods for very large datasets
  - dimensionality reduction methods
  - SVD can be quickly performed on large matrices
  - should we preferably use single imputation?

References I

V. Audigier, F. Husson, and J. Josse. Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2):501–518, 2017. ISSN 1573-1375.

V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156, 2016a.

V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26, 2016b.

D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.

How to choose the number of dimensions?

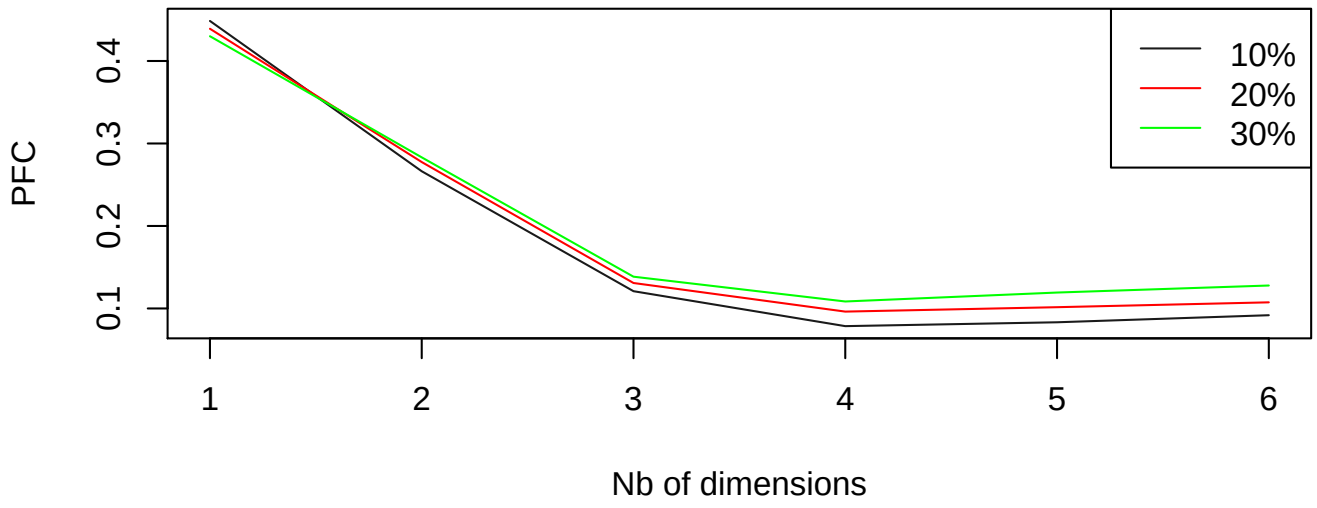
By cross-validation procedures:

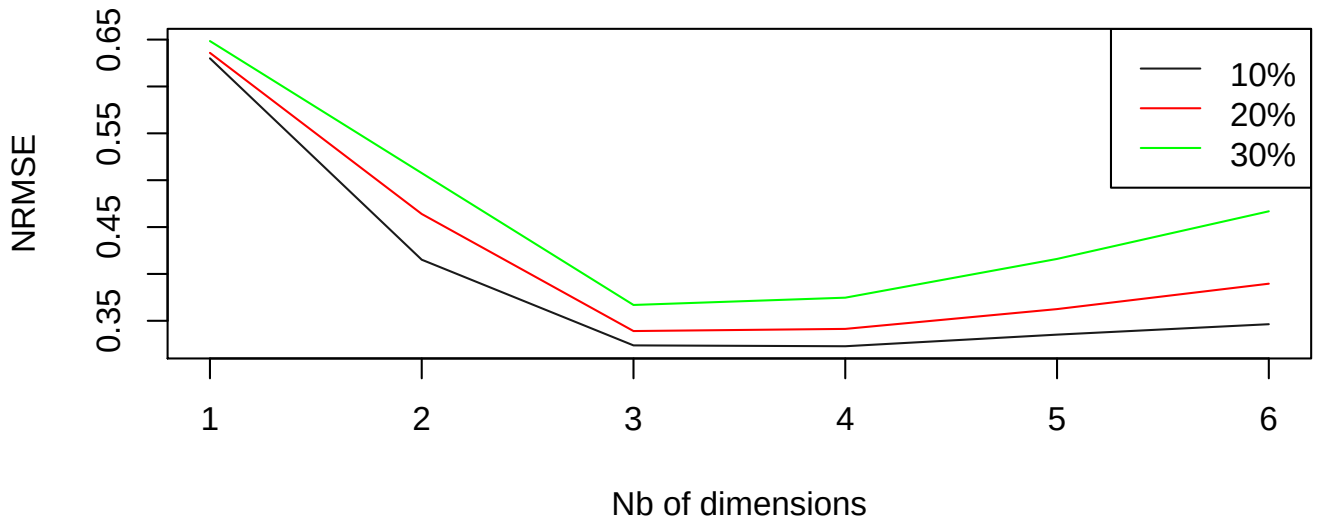
- adding missing values on the incomplete data set
- predicting each of them using FAMD for several number of dimensions
- calculating the prediction error

Several ways:

- Leave-one-out (Bro et al., 2008)
- Repeated cross-validation

Misspecification of the number of dimensions





Single imputation MAR

- A mixed data set is simulated by splitting normal data
- Missing values are added on one variable Y according to a MAR mechanism:$\mathbb{P}(Y = NA) = \frac{exp(\beta_0+\beta_1 X_1)}{1+exp(\beta_0+\beta_1 X_1)}$
- Data are imputed using FAMD and RF