Handling missing values with regularized iterative multiple correspondence analysis

J. Josse[1], M. Chavent[2], B. Liquet[3] & F. Husson[1]

[1] Agrocampus Rennes
[2] University of Bordeaux 2
[3] ISPED Bordeaux

ICC conference, St Andrew, 11 July 2011

Handling missing values in PCA

$\Rightarrow$ Minimization of:

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

$\Rightarrow$ 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

1. initialization $\ell = 0$: $\mathbf{X}^0$ (mean imputation)
2. step ℓ:
   (a) PCA is performed on the completed data set $\longrightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$; S dimensions are kept
   (b) missing values are imputed with the model matrix $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell\prime}$; the new imputed dataset is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$
3. steps (a) and (b) are repeated until convergence.

$\Rightarrow$ The number of dimensions S has to be chosen *a priori*
$\Rightarrow$ EM algorithm of $x_{ij} = \sum_{s=1}^{S} f_{is} u_{js} + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

$\Rightarrow$ Nora-Chouteau in CA (1974)

MCA as a particular PCA

MCA can be seen as the PCA of (data, metric, row masses)

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

with $\mathbf{X}$ the indicator matrix and $\mathbf{D}_\Sigma$ the diagonal matrix of the column margins of $\mathbf{X}$,

$$X = \begin{array}{|ccc|ccc|cc|c|}
\hline
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 & J \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots & \text{NA NA} & \\
\text{NA} & \text{NA} & \text{NA} & 0 & 1 & 0 & 0 & \ldots & 0\ 1 & J \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 & \\
\hline
 & & & & & & & & & \\
 & & & & & xik & & & & J \\
 & & & & & & & & & \\
\hline
0 & 0 & 1 & \text{NA} & \text{NA} & 0 & \ldots & 0 & 1 & \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 & J \\
\hline
\multicolumn{3}{c}{I_1} & \multicolumn{3}{c}{I_k} & \multicolumn{2}{c}{I_K} & IJ &
\end{array}$$

$$D_\Sigma = \begin{pmatrix} I_1 & & & 0 \\ & \ddots & & \\ & & I_k & \\ 0 & & & \ddots \\ & & & & I_K \end{pmatrix}$$

Iterative MCA

1. initialization $\ell = 0$: $\mathbf{X}^0$ missing values are imputed with the proportion of the category $\frac{I_k}{I}$ (the sum must equal one);
2. step ℓ:
   (a) MCA on $\mathbf{X}^{\ell-1}$: PCA on $\left(I\mathbf{X}^{\ell-1}(\mathbf{D}_\Sigma^{\ell-1})^{-1}, \frac{1}{IJ}\mathbf{D}_\Sigma^{\ell-1}, \frac{1}{I}\mathbb{I}_I\right) \rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$; S dimensions are kept
   (b) impute the indicator matrix using the model matrix:

$$\hat{x}_{ik}^\ell = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \hat{f}_{is}^\ell \hat{u}_{ks}^\ell\right)\mathbf{D}_\Sigma^{\ell-1}$$

   The new imputed dataset is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$
   (c) $\mathbf{D}_\Sigma^\ell$ is updated with the new column margins I_k^ℓ of $\mathbf{X}^\ell$;
3. steps (a), (b) and (c) are repeated until convergence

Iterative MCA

1. initialization $\ell = 0$: $\mathbf{X}^0$ missing values are imputed with the proportion of the category $\frac{I_k}{I}$ (the sum must equal one);
2. step ℓ:
   (a) MCA on $\mathbf{X}^{\ell-1}$: PCA on $\left(I\mathbf{X}^{\ell-1}(\mathbf{D}_\Sigma^{\ell-1})^{-1}, \frac{1}{IJ}\mathbf{D}_\Sigma^{\ell-1}, \frac{1}{I}\mathbb{I}_I\right) \rightarrow (\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell)$; S dimensions are kept
   (b) impute the indicator matrix using the model matrix:

$$\hat{x}_{ik}^\ell = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \hat{f}_{is}^\ell \hat{u}_{ks}^\ell\right)\mathbf{D}_\Sigma^{\ell-1}$$

   The new imputed dataset is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
   (c) $\mathbf{D}_\Sigma^\ell$ is updated with the new column margins I_k^ℓ of $\mathbf{X}^\ell$
3. steps (a), (b) and (c) are repeated until convergence

Iterative MCA

- Step 0: missing fuzzy average = reconstruction of order 0

$$\hat{x}_{ik}^{0} = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \hat{f}_{is}^{\ell}\hat{u}_{ks}^{\ell}\right)\mathbf{D}_{\Sigma}^{0} = \mathbf{D}_{\Sigma}^{0}/I$$

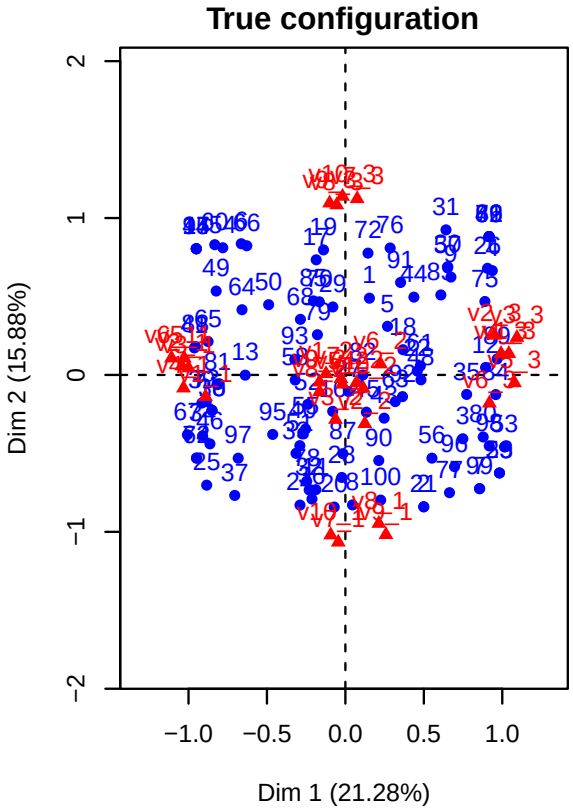
- The algorithm returns a completed indicator matrix

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

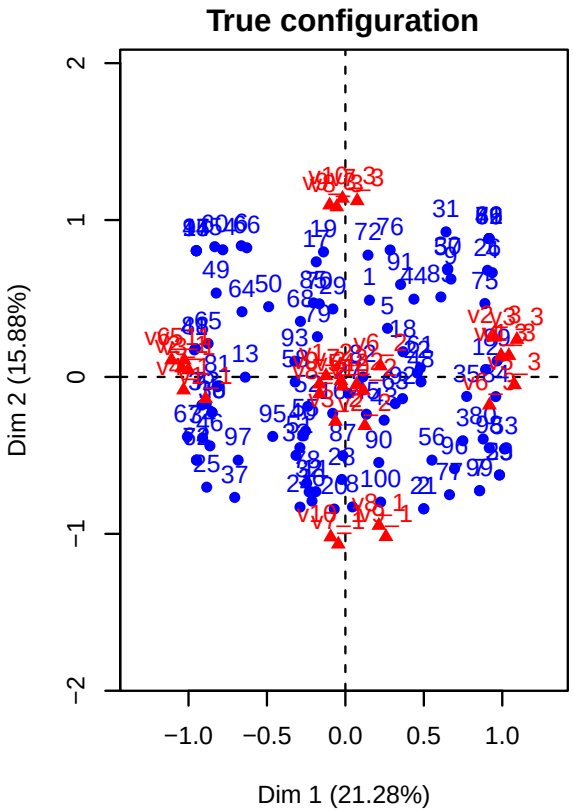
	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h
ind 1	1	0	0	**0.71**	**0.29**	1	0
ind 2	**0.13**	**0.29**	**0.59**	0	1	1	0
ind 3	1	0	0	1	0	0	1
ind 4	1	0	0	1	0	0	1
ind 5	0	1	0	0	1	0	1
ind 6	0	0	1	0	1	0	1
ind 7	0	0	1	0	1	**0.37**	**0.63**

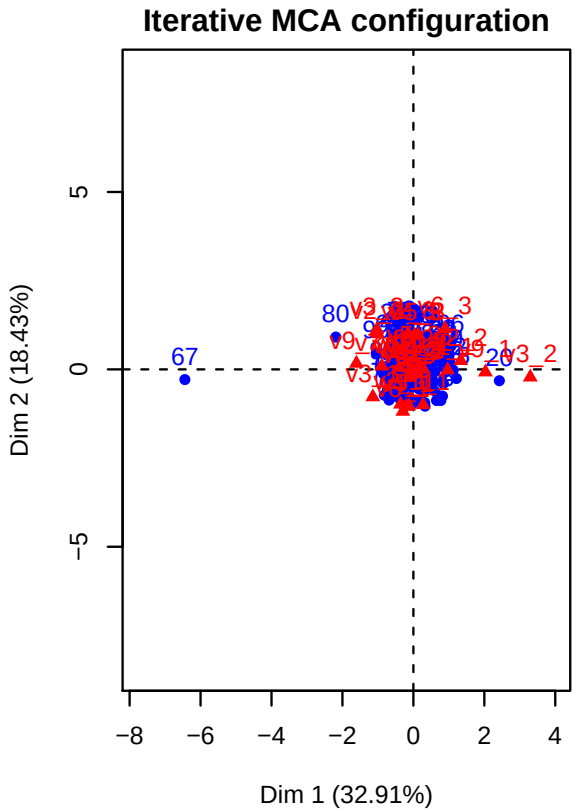
$\Rightarrow$ imputed values can be seen as degree of membership

Overfitting

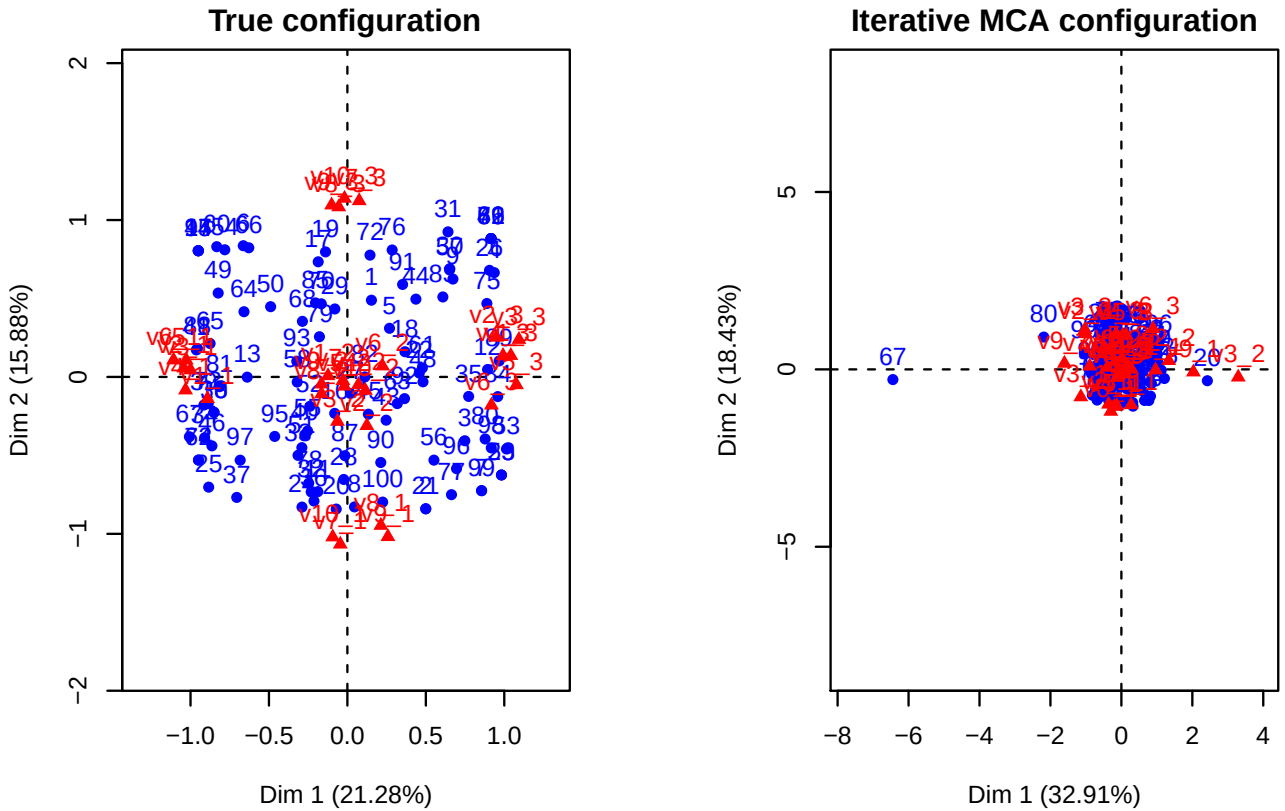


Overfitting





Overfitting



$\underset{i,k \in \text{obs}}{\text{mean}}(x_{ik} - \hat{x}_{ik})^2 = 0.026$ whereas $\underset{i,k \notin \text{obs}}{\text{mean}}(x_{ik} - \hat{x}_{ik})^2 = 0.12$

$\Rightarrow$ observed values are well-fitted but the prediction of the missing values and the estimation of the axes and components are very poor

$\Rightarrow$ convergence problems of EM $\longrightarrow$ overfitting problems

Overfitting

$\Rightarrow$ Fitting is good, prediction is bad

- Many parameters are estimated with respect to the number of observed values: the number of dimensions S and the number of missing values are important
- The relationship between variables are not strong

1. Reduce S
2. Early stopping

3. Shrinkage methods

Regularized Iterative MCA

$\Rightarrow$ Initialization, estimation step - imputation step.

The imputation step:

$$\hat{x}_{ik}^{\ell} = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \hat{f}_{is}^{\ell}\hat{u}_{ks}^{\ell}\right)\mathbf{D}_{\Sigma}^{\ell-1} = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\|\hat{\mathbf{f}}_{s}^{\ell}\|}(\sqrt{\lambda_s})\hat{u}_{ks}^{\ell}\right)\mathbf{D}_{\Sigma}^{\ell-1},$$

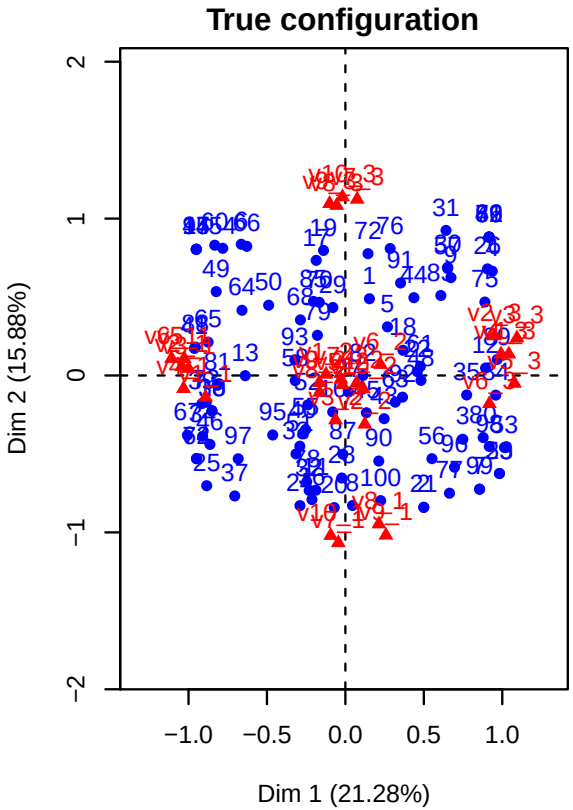
is replaced by a "shrunk" imputation step:

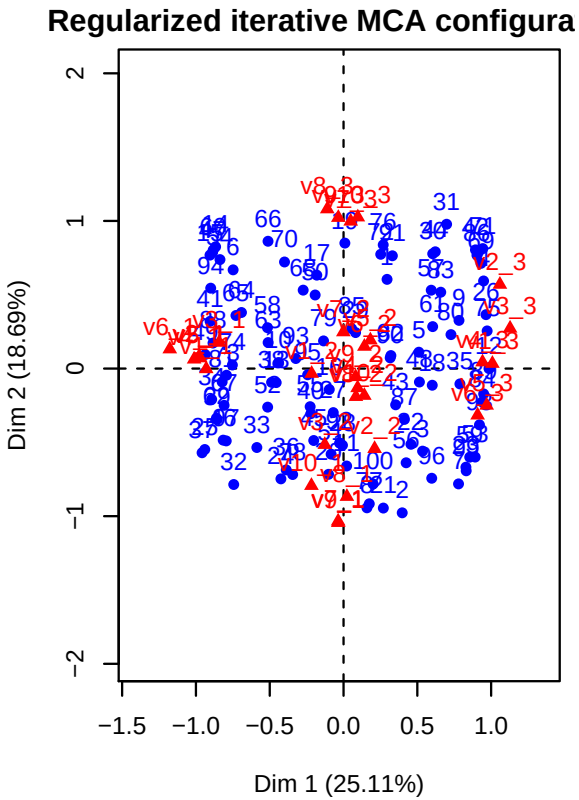
$$\hat{x}_{ik}^{\ell} = \frac{1}{I}\left(1 + \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\|\hat{\mathbf{f}}_{s}^{\ell}\|}\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)\hat{u}_{ks}^{\ell}\right)\mathbf{D}_{\Sigma}^{\ell-1}$$

with $\hat{\sigma}^2 = \frac{1}{K-J-S}\sum_{s=S+1}^{K-J}\lambda_s$.

$\Rightarrow$ Remove the noise to avoid instability on the predictions

Regularized iterative MCA





$$\underset{i,k \notin \text{obs}}{\text{mean}}(x_{ik} - \hat{x}_{ik})^2 = 0.056$$

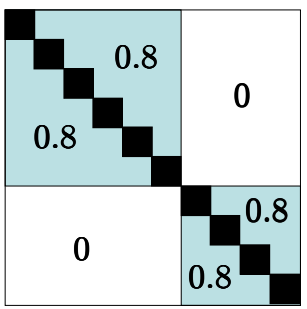
Simulations

Many scenarios are considered:

- percentage of missing values: small, medium
- missing values mechanism: MCAR, MAR
- pattern of missing values: random or not random
- relationship between variables: low or strong
- 1000 simulations

The simulated data:

- 100 individuals
- 10 variables from a normal distribution
- each variable is cut in 3 equal-count categories
  $\Rightarrow$ By construction, 4 underlying dimensions



Missing single

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

	V1_a	V1_b	V1_c	**V1_NA**	V2_e	V2_f	**V2_NA**	V3_g	V3_h	**V3_NA**
ind 1	1	0	0	0	**0**	**0**	**1**	1	0	0
ind 2	**0**	**0**	**0**	**1**	0	1	0	1	0	0
ind 3	1	0	0	0	1	0	0	0	1	0
ind 4	1	0	0	0	1	0	0	0	1	0
ind 5	0	1	0	0	0	1	0	0	1	0
ind 6	0	0	1	0	0	1	0	0	1	0
ind 7	0	0	1	0	0	1	0	**0**	**0**	**1**

- A new category is added for missing values
  $\Rightarrow$ well-adapted for "not really missing" or MNAR

Missing passive modified margin

- Missing passive (Meulman, 1982)

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

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h
ind 1	1	0	0	**0**	**0**	1	0
ind 2	**0**	**0**	**0**	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**	**0**

Row margins are not equal $\Rightarrow$ many properties of MCA are lost

- Missing passive modified margin (Escofier, 1987)
  $\Rightarrow$ row margins are fixed to J
  $\Rightarrow$ Equivalence with subset MCA (Greenacre & Pardo, 2006)

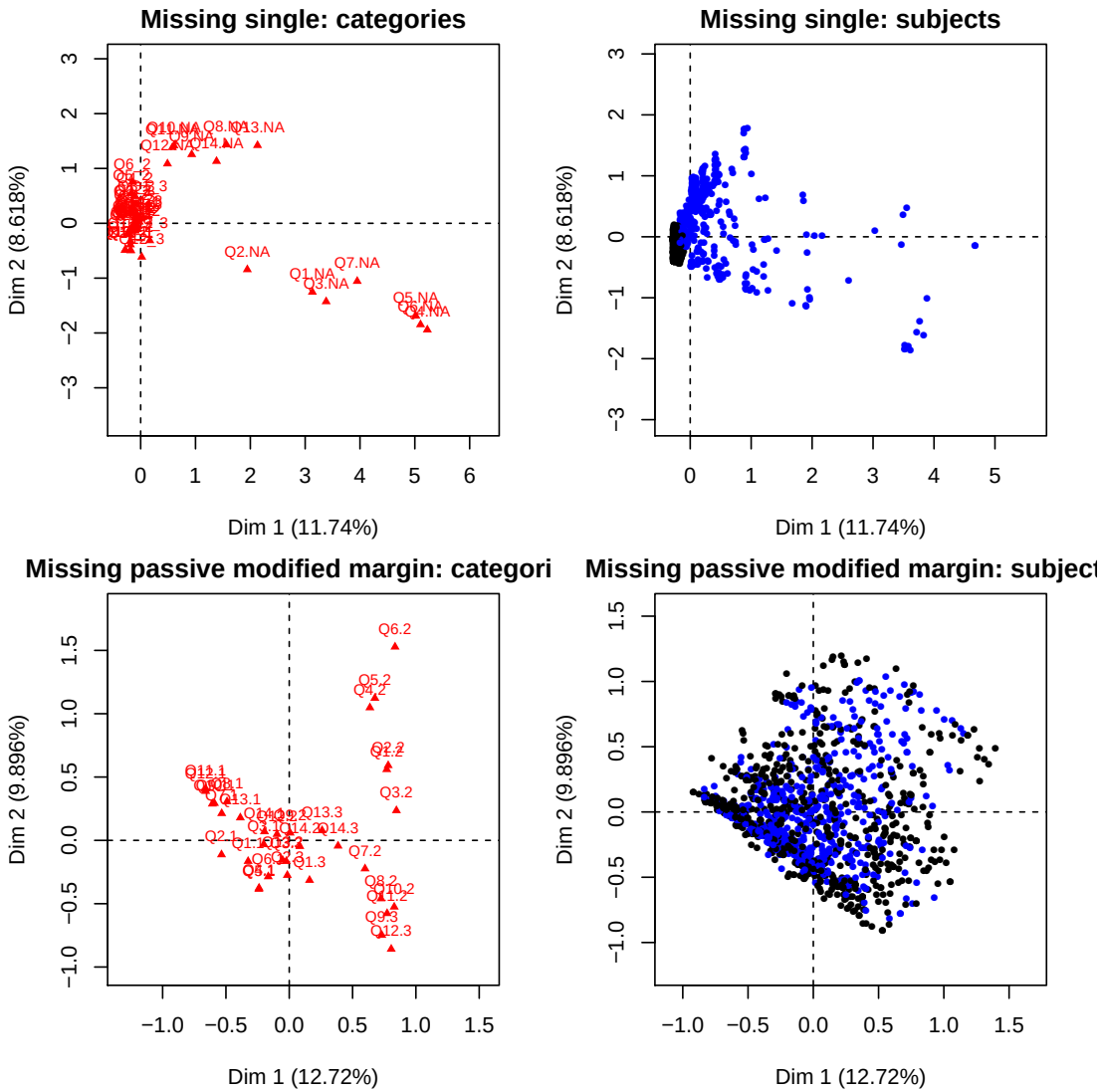
Simulations

The criterion used is the RV coefficient between the configuration without missing values and the one obtained from the algorithm

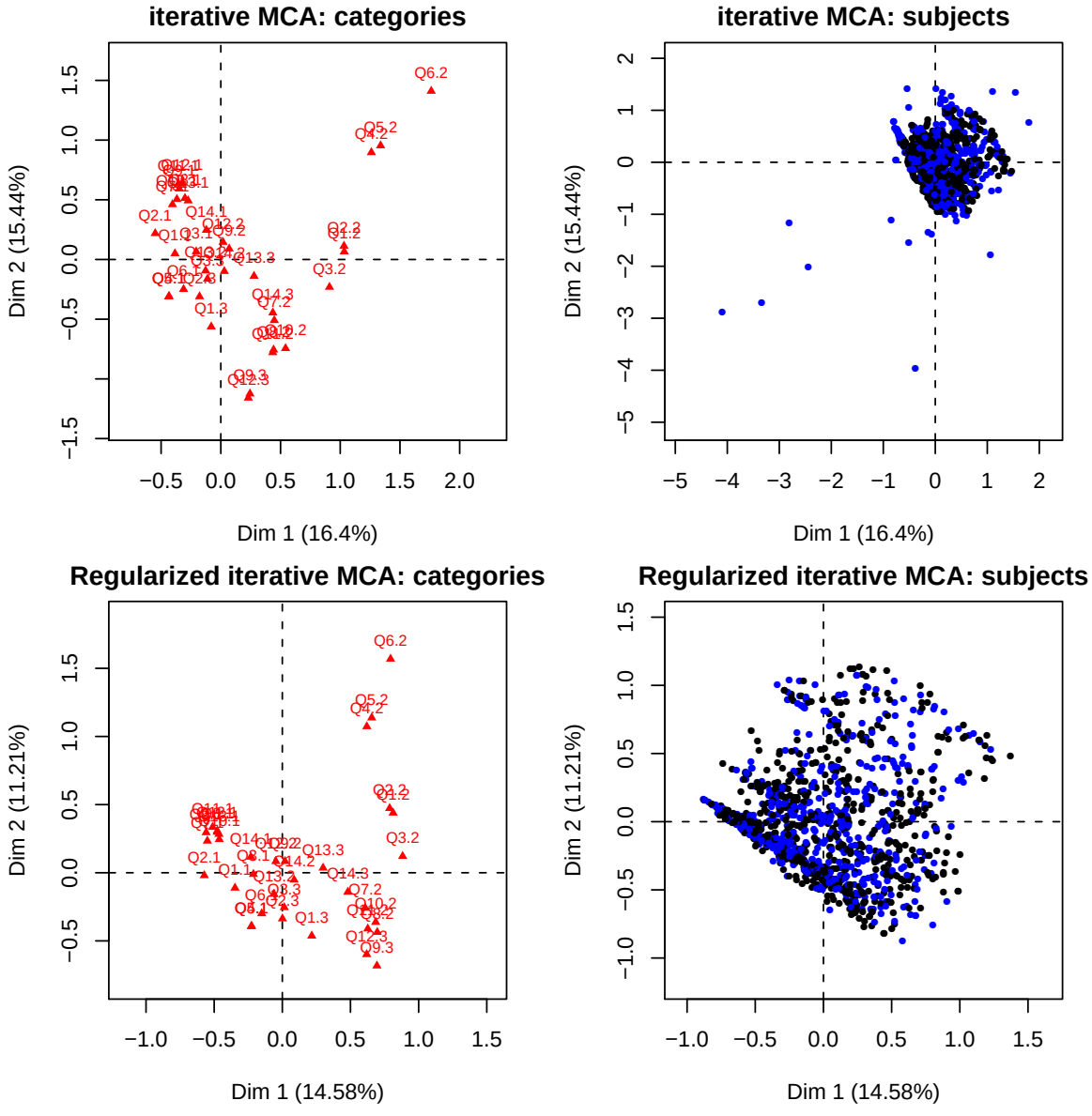
Missing	Link	Missing Passive Modified Margin R - NR	Missing Fuzzy Average R - NR	Missing single R - NR	RiMCA R - NR
10% MCAR	low	**0.94** - 0.91	**0.94** - 0.92	0.87 - 0.47	**0.94** - **0.93**
10% MCAR	strong	0.97 - 0.94	0.97 - 0.95	0.96 - 0.68	**0.98** - **0.97**
30% MCAR	low	**0.77** - 0.44	0.77 - 0.77	0.67 - 0.32	0.76 - **0.78**
30% MCAR	strong	0.88 - 0.71	0.88 - **0.91**	0.86 - 0.46	**0.91** - 0.90
8% MAR	low	0.94 - 0.91	0.94 - 0.91	0.72 - 0.28	**0.95** - **0.92**
8% MAR	strong	0.96 - 0.91	0.96 - 0.90	0.96 - 0.54	**0.98** - **0.96**
16% MAR	low	0.86 - 0.80	0.83 - 0.79	0.50 - 0.29	**0.88** - **0.83**
16% MAR	strong	0.89 - 0.80	0.84 - 0.78	0.88 - 0.55	**0.95** - **0.90**

A real example

- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents



A real example



Conclusion

Regularized iterative MCA

- gives "good" results
- is efficient when strong relationships between variables (you learn from the other variables) ...
- ... but needs tuning parameters
- can be used as an imputation method?
- can be used to perform a clustering on categorical variables with missing values
- is available in the `missMDA` package that imputes the indicator matrix and the `FactoMineR` package that performs the MCA from an indicator matrix