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

CARME conference, Rennes, 9 February 2011

Type of missing values

- **“Really missing”** and “not really missing”
- **MCAR, MAR**, MNAR (Rubin, 1976)

⇒ In MCA, van der Heijden & Escofier (1987) discussed which method is well suited for which kind of missing data

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**

- Missing single: a new category is added for missing values
  $\Rightarrow$ well-adapted for “not really missing” or MNAR

Missing passive modified margin

- Missing passive (Benzécri, 1973; 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**

Missing values are skipped
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
  $\Rightarrow$ Good properties: $\mathbf{f}_s$ maximises $\sum_{j=1}^{J} \hat{\eta}^2_{\mathbf{f}_s|\mathbf{v}_j}$
  $\Rightarrow$ Equivalence with subset MCA (Greenacre & Pardo, 2006)

Handling missing values in exploratory multivariate analysis

The method consists to find the components **F** and the axes **U** that minimize the reconstruction error:

$$\mathcal{C} \quad = \quad \|\mathbf{X} - \mathbf{F}\mathbf{U}'\|^2_{\mathbf{M},\mathbf{D}}$$

With missing values, a matrix of weights **W** is introduced:

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

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

$\Rightarrow$ Use of iterative algorithms

Iterative algorithms

- Initialization: missing values in $\mathbf{X}$ are imputed with initial values (such as the mean of each variable)
- Estimation step: the analysis is performed on the completed data set
- Imputation step: missing values are imputed with the reconstruction formulae with S dimensions

$$\mathbf{X} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * (\hat{\mathbf{F}}\hat{\mathbf{U}}')$$

- Steps E and M are repeated until convergence

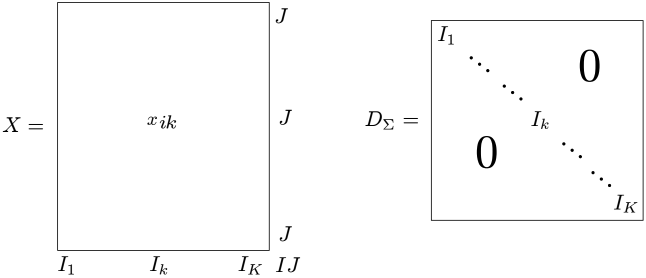
⇒ EM type algorithms
⇒ The number of dimensions S has to be chosen *a priori*
⇒ Nora-Chouteau in CA (1974); Kiers in PCA (1997)

Iterative MCA

MCA can be seen as the SVD 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 $\mathbf{X}$,



Iterative MCA

1. initialization $\ell = 0$: $\mathbf{X}^0$ missing values are imputed with the proportion of the category (the sum must equal one) $\Rightarrow \mathbf{D}_\Sigma^0$;
2. step ℓ:
   a) MCA on $\mathbf{X}^{\ell-1}$: $\hat{\mathbf{F}}$ and $\hat{\mathbf{U}}$ are obtained from a PCA on

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

   b) Impute the indicator matrix using the reconstruction formulae:

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

   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 (2.a), (2.b) and (2.c) are repeated until convergence

Iterative MCA

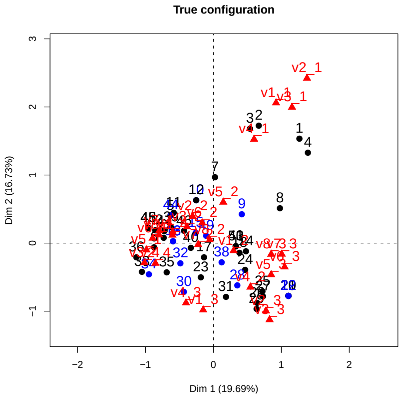
- Step 0: missing fuzzy average = reconstruction of order 0
- The algorithm can return a completed indicator matrix

	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**

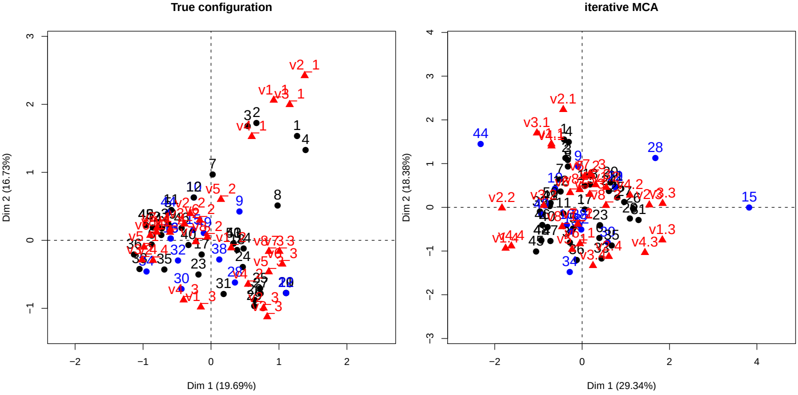
	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h
ind 1	1	0	0	**0.71**	**0.29**	1	0
ind 2	**0.13**	**0.29**	**0.59**	0	1	1	0
ind 3	1	0	0	1	0	0	1
ind 4	1	0	0	1	0	0	1
ind 5	0	1	0	0	1	0	1
ind 6	0	0	1	0	1	0	1
ind 7	0	0	1	0	1	**0.37**	**0.63**

- Imputed values can be seen as degree of membership

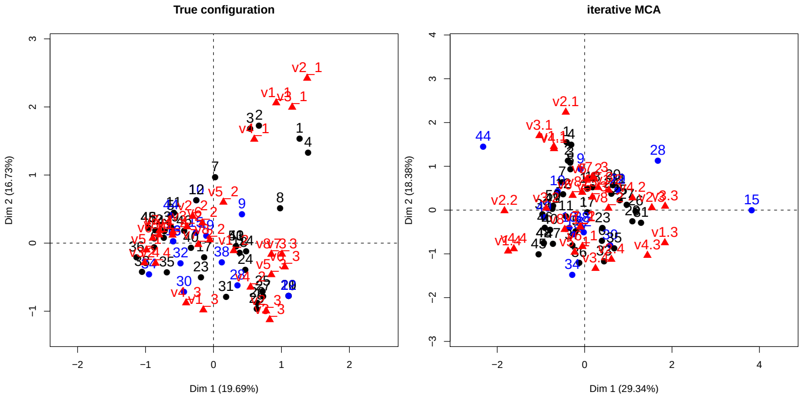
Overfitting



Overfitting



Overfitting



$$\underset{i,k\in\text{obs}}{\text{mean}}(x_{ik}-\hat{x}_{ik})^2 = 0.03 \text{ whereas } \underset{i,k\notin\text{obs}}{\text{mean}}(x_{ik}-\hat{x}_{ik})^2 = 0.34$$

Observed values are well-fitted but missing ones are badly predicted

... and consequently axes and components are badly predicted

$\Rightarrow$ Regularization methods

Regularized Iterative MCA

$$\sum_{s=1}^{S} \hat{f}_{is}^{\ell} \hat{u}_{ks}^{\ell} = \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\|\hat{\mathbf{f}}_{s}^{\ell}\|_{\mathbf{D}}} (\sqrt{\lambda_s}) \hat{u}_{ks}^{\ell}$$

The eigenvalues can be shrunk in the reconstruction step:

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

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

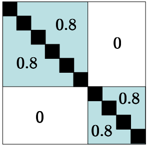
Simulations

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



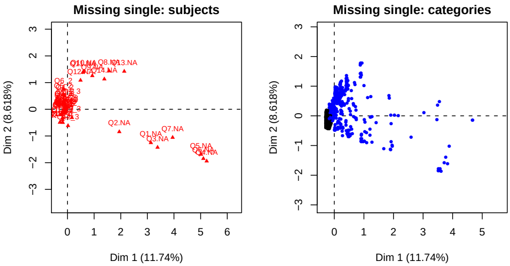
Simulations

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**

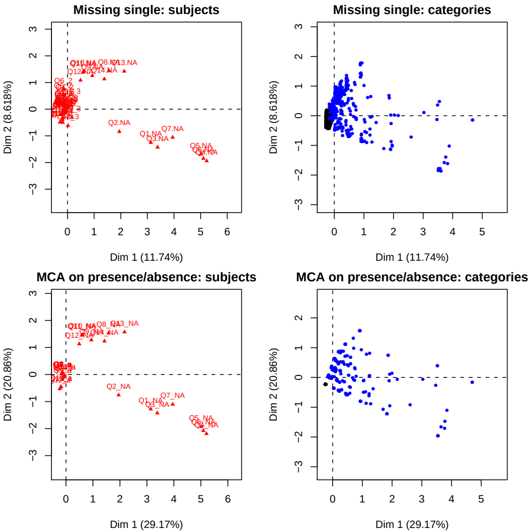
A real example

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

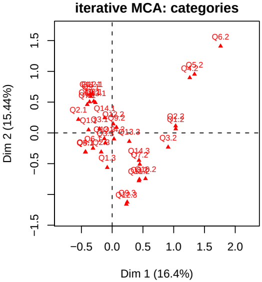


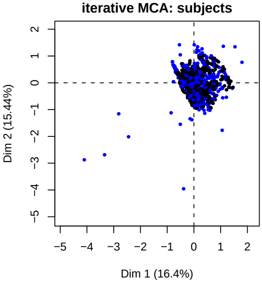
A real example

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

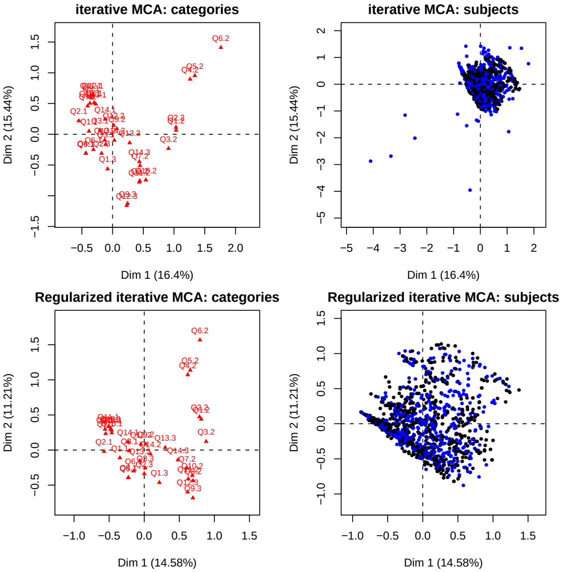


A real example





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