Multiple Imputation with MCA

Vincent Audigier & Julie Josse & François Husson

Agrocampus Ouest, Rennes

Jds, Lille, June 1, 2015

Missing values

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

- Deletion of individuals: complete case
- Expectation-Maximisation
- Imputation

MCA

- Dimensionality reduction method
- MCA can be seen as the SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_{\Sigma}\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$:

$$\mathbf{X}_{I\times J} = \mathbf{U}_{I\times J}\Lambda^{1/2}_{J\times J}\mathbf{V}^{\top}_{J\times J}$$

$\rightarrow$ retain S dimensions:

$$\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}$$

$$X = \begin{array}{|ccc|cc|cc|c|} \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots & \text{NA NA} \\ \text{NA} & \text{NA} & \text{NA} & 0 & 1 & 0 & 0 & \ldots & 0\ 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 \\ & & & & & & & & \\ & & & & & x_{ik} & & & \\ & & & & & & & & \\ 0 & 0 & 1 & \text{NA} & \text{NA} & 0 & \ldots & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & \ldots & 0\ 1 \\ \hline \end{array} \begin{array}{c} J \\ \\ \\ \\ \\ J \\ \\ \\ \\ \\ J \end{array}$$

$$\begin{array}{ccccc} I_1 & & I_k & & I_K \end{array} \quad IJ$$

$$D_{\Sigma} = \begin{array}{|ccc|} \hline I_1 & & \\ & \ddots & \\ & & \ddots \\ & I_k & \\ & & \ddots \\ & & I_K \\ \hline \end{array} \quad \begin{array}{l} \\ \\ 0 \\ \\ 0 \\ \\ \end{array}$$

Iterative MCA

Iterative MCA algorithm:

	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	**NA**	**NA**	1	0	...
ind 2	**NA**	**NA**	**NA**	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	**NA**	**NA**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)

	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	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	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	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) perform the MCA, *i.e.* SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   $\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^\top_{J\times S}$,

	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	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	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	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) perform the MCA, *i.e.* SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   $\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^\top_{J\times S}$,
   (b) imputation of the missing values with $\hat{\mathbf{X}}_{I\times J} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^\top_{J\times S}$

	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	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	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	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) perform the MCA, *i.e.* SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   $\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^\top_{J\times S}$,
   (b) imputation of the missing values with $\hat{\mathbf{X}}_{I\times J} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^\top_{J\times S}$
   (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	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	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	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) perform the MCA, *i.e.* SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   $\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}$,
   (b) imputation of the missing values with $\hat{\mathbf{X}}_{I\times J} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{J\times S}$
   (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	...

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

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) perform the MCA, *i.e.* SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   $\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}$,
   (b) imputation of the missing values with $\hat{\mathbf{X}}_{I\times J} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^{\top}_{J\times S}$
   (c) column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**e**	g	...	u
ind 2	**c**	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	**g**		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	...

Two ways to obtain categories: majority or draw

Single imputation methods

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
$\vdots$	$\vdots$

Majority

$\pi_{b\|A}$	0.15
$\pi_{a\|A}$	0.85
$\pi_{a\|B}$	0.58
$\pi_{b\|B}$	0.42

$cov_{95\%}(\pi_b) = 0.0$

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

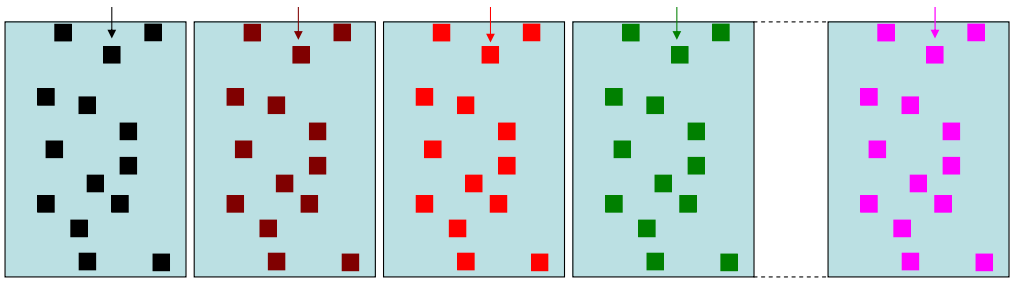
$\pi_{b\|A}$	0.18
$\pi_{a\|A}$	0.82
$\pi_{a\|B}$	0.41
$\pi_{b\|B}$	0.59

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

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

Multiple imputation (Rubin, 1987)

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



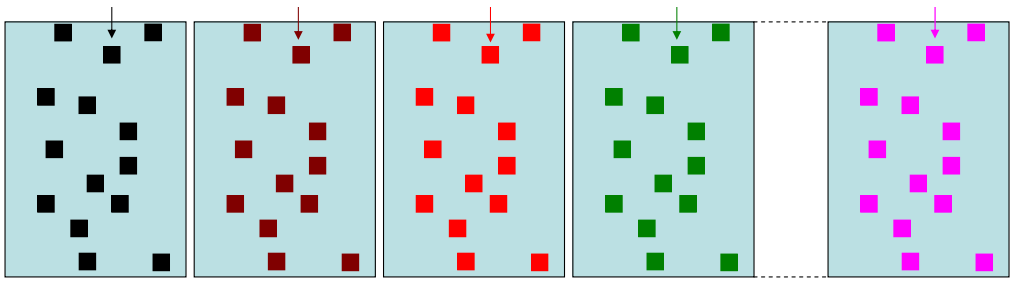
- Perform the analysis on each imputed data set: $\hat{\theta}_m, \widehat{Var}\left(\hat{\theta}_m\right)$
- Combine the results: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$

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

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

Multiple imputation (Rubin, 1987)

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



Bayesian or Bootstrap approach

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

Algorithm MIMCA

1. Variability of the parameters of MCA:
   - non-parametric bootstrap: define M weightings $(r_m)_{1 \leq m \leq M}$ for the individuals

Algorithm MIMCA

1. Variability of the parameters of MCA:
   - non-parametric bootstrap: define M weightings $(r_m)_{1 \leq m \leq M}$ for the individuals
2. Imputation: for $m = 1, \ldots, M$,
   - $\mathbf{X}$ is imputed by iterative MCA according to the weighting r_m
   - return to categorical values: draw one of the categories

Multiple Imputation using the loglinear model

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

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

1. Variability of the parameters
   - prior on θ : $\theta|\theta \in \Theta \sim \mathcal{D}(\alpha)$
   - posterior: $\theta|x, \theta \in \Theta \sim \mathcal{D}(\alpha')$
   - Data Augmentation (M.A. Tanner, W.H. Wong, 1987)
2. Imputation according to the loglinear model using the set of M parameters

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

Conditional modelling using logistic regressions

- Hypothesis: one logistic regression model/variable
- Algorithm:

  One variable with missing values:

  1. Bayesian approach: $\beta|X \sim \mathcal{N}(\hat{\beta}, \hat{V}) \Rightarrow \beta^m$
  2. Imputation: stochastic regression x_{ik}^m drawn from $\mathcal{M}(\theta_k, 1)$
     $\theta_{k\ell} = \mathbb{P}(X_k = \ell | X_{-k}, \beta^m) = \frac{exp(\mathbf{Z}_k \boldsymbol{\beta^m}_{k\ell})}{1+\sum_{\ell=\mathbf{1}}^{L-\mathbf{1}} exp(\mathbf{Z}_k \boldsymbol{\beta^m}_{k\ell})}$

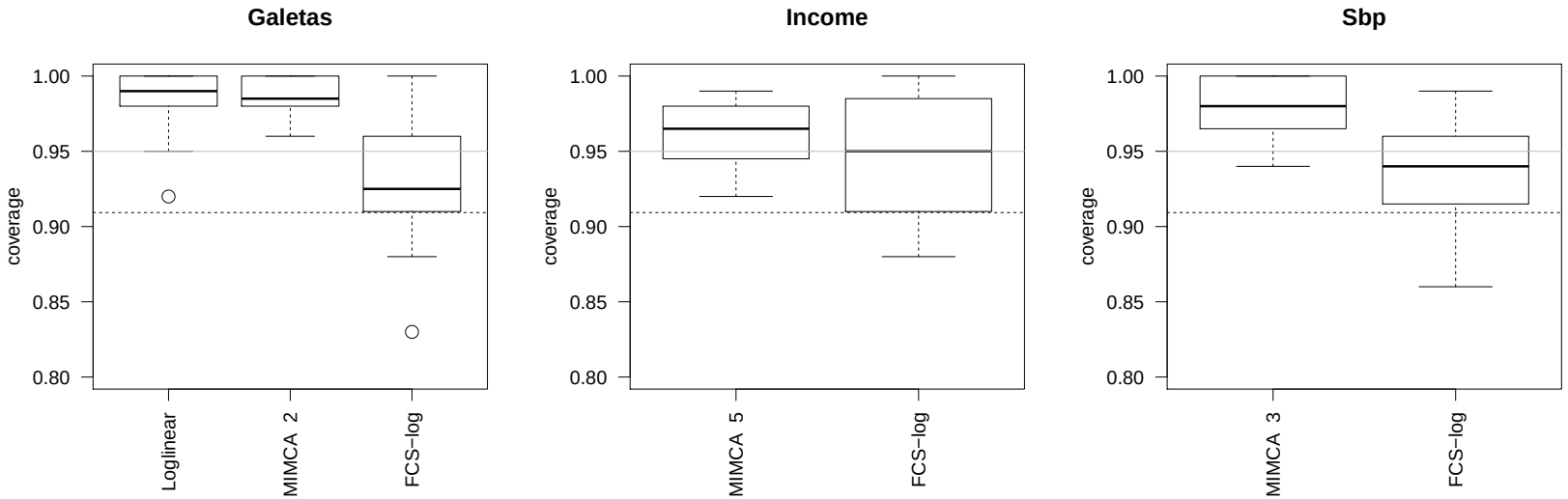
  Many variables with missing values: cycles through variables

- Implemented: `R` package `MICE` (Stef van Buuren)

Simulations

- Quantities of interest: $\theta =$ parameters of a logistic model
- 200 simulations from real data sets
  - 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

Results



	Galetas	Income	Sbp
Number of variables	4	14	18
Number of categories	≤ 11	≤ 9	≤ 4

Results

	Galetas	Income	Sbp
Loglinear	4.597	NA	NA
MIMCA	8.972	58.729	7.181
FCS log	38.016	881.188	53.109
FCS forests	112.987	6329.514	193.156
DPMPM	17.414	143.652	56.302

Table: Time consumed in second

	Galetas	Income	Sbp
Number of individuals	1192	6876	500
Number of variables	4	14	18

Conclusion

A new multiple imputation method based on MCA

- strong points: dimensionality reduction method
  - could be applied on data sets of various dimensions
  - performs efficiently for usual statistical methods
- a tuning parameter: the number of dimensions

$\Rightarrow$ mixed data