Multiple Imputation with MCA

Vincent Audigier & Julie Josse & François Husson

Agrocampus Ouest, Rennes

Jstar, Rennes, 23 octobre 2014

Missing values

			X				Y
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

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

Imputation with MCA for categorical variables

MCA can be seen as the PCA on an indicator matrix **X** with specific weights depending on $\mathbf{D}_\Sigma$

$$X = \begin{array}{|ccc|cc|cc|cc|} \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & \text{NA} & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ & & & & & & & & \\ & & & & & xik & & & \\ & & & & & & & & \\ & & & & & & & & \\ 0 & 0 & 1 & \text{NA} & \text{NA} & 0 & 0 & 1 & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ \hline \end{array} \begin{array}{l} J \\ \\ \\ \\ \\ J \\ \\ \\ \\ J \end{array}$$

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

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

Iterative MCA

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) PCA on the completed indicator matrix divided by column margins $\Rightarrow \hat{\mathbf{F}}$, $\hat{\mathbf{U}}$ (Estimation)
   (b) imputation of the missing values with the model matrix
   (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	**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)
2. iterate until convergence
   (a) PCA on the completed indicator matrix divided by column margins $\Rightarrow \hat{\mathbf{F}}$, $\hat{\mathbf{U}}$ (Estimation)
   (b) imputation of the missing values with the model matrix
   (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	...

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

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
$\vdots$	$\vdots$

Proportion

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

$cov_{95\%}(\pi_b) = 0.00$

Regression

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

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

Stochastic regression

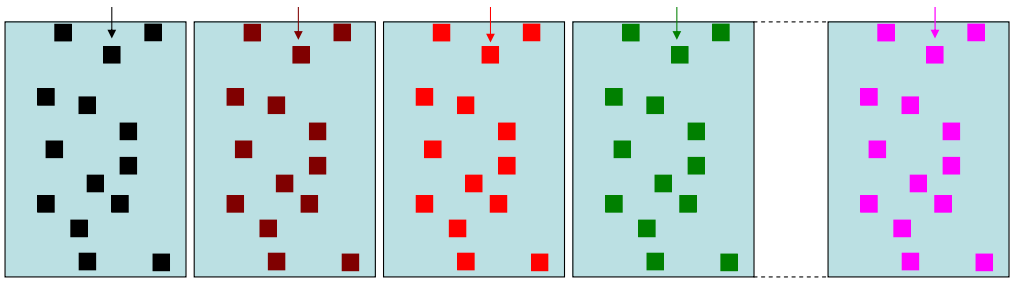
$\pi_{b\|A}$	0.18
$\pi_{a\|A}$	0.82
$\pi_{a\|B}$	0.41
$\pi_{b\|B}$	0.59

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

$\Rightarrow$ 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



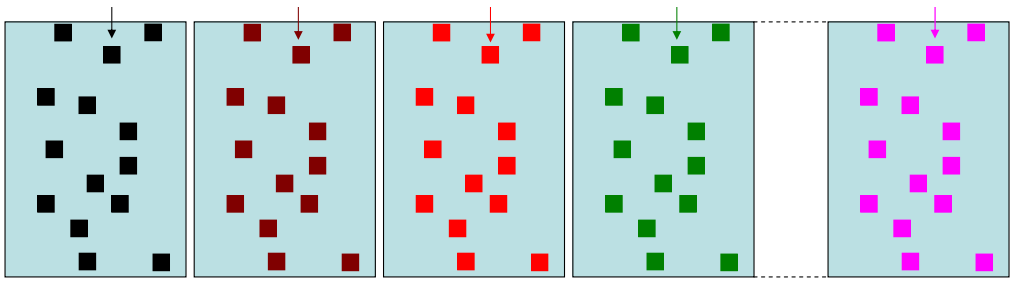
- Perform the analysis on each imputed data set: $\hat{\theta}_m, \widehat{Var}\left(\hat{\theta}_m\right)$
- Combine the results: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$

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

$\Rightarrow$ 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



Bootstrap or Bayesian 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:
   - non parametric bootstrap
   - estimate F_m, U_m, μ_m, σ_m^2 using iterative MCA

Algorithm MIMCA

1. Variability of the parameters:
   - non parametric bootstrap
   - estimate F_m, U_m, μ_m, σ_m^2 using iterative MCA
   - but the estimates are defined for others individuals than those of the original data set

Algorithm MIMCA

1. Variability of the parameters:
   - non parametric bootstrap
   - estimate F_m, U_m, μ_m, σ_m^2 using iterative MCA
   - but the estimates are defined for others individuals than those of the original data set
     - deduce the M covariance matrices $(\widehat{\Sigma}_m = \widehat{U}\widehat{\Lambda}\widehat{U}^\top)_{1 \leq m \leq M}$

Algorithm MIMCA

1. Variability of the parameters:
   - non parametric bootstrap
   - estimate F_m, U_m, μ_m, σ_m^2 using iterative MCA
   - but the estimates are defined for others individuals than those of the original data set
     - deduce the M covariance matrices $(\widehat{\Sigma}_m = \widehat{U}\widehat{\Lambda}\widehat{U}^\top)_{1 \leq m \leq M}$
2. Imputation: for $m = 1, \ldots, M$, missing values x_{ij} are imputed by drawing from the distribution $\mathcal{N}(\widehat{\mu}_m, \widehat{\Sigma}_m + \widehat{\sigma}_m^2)$

Algorithm MIMCA

1. Variability of the parameters:
   - non parametric bootstrap
   - estimate F_m, U_m, μ_m, σ_m^2 using iterative MCA
   - but the estimates are defined for others individuals than those of the original data set
     - deduce the M covariance matrices $(\widehat{\Sigma}_m = \widehat{U}\widehat{\Lambda}\widehat{U}^\top)_{1 \leq m \leq M}$
2. Imputation: for $m = 1, \ldots, M$, missing values x_{ij} are imputed by drawing from the distribution $\mathcal{N}(\widehat{\mu}_m, \widehat{\Sigma}_m + \widehat{\sigma}_m^2)$
3. Return to categorical values
   - scale the continuously values into probabilities
   - draw one of the categories

Expectation Maximization Bootstrap Algorithm

- Hypothesis $X_{n\times p} : x_{i.} \sim \mathcal{N}(\mu, \Sigma)$
- Algorithm:
  1. Bootstrap rows: $X^1, \ldots, X^M$
     EM algorithm: $(\mu^1, \Sigma^1), \ldots, (\mu^M, \Sigma^M)$
     $\rightarrow$ difference with MIMCA
  2. Imputation: $x_{i.}^m$ drawn from $\mathcal{N}(\mu^m, \Sigma^m)$
  3. Return to categorical values
- Implemented: `R` package `Amelia` (J. Honaker, G. King, M. Blackwell)

Data augmentation Bayesian Iterative proportional fitting

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

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

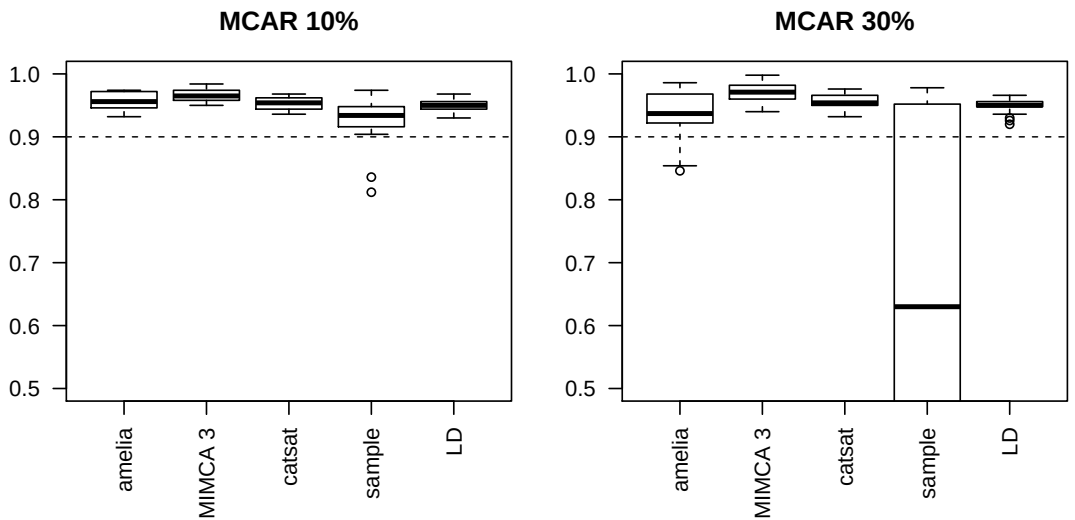
1. Variability of the parameters
   - prior on θ : $\theta|\theta \in \Theta \sim \mathcal{D}(\alpha)$
   - posterior: $\theta|x, \theta \in \Theta \sim \mathcal{D}(\alpha')$
   - Data Augmentation:
     (0) Max likelihood estimate for θ (no closed form $\rightarrow$ IPF)
     (I) impute using the loglinear model
     (P) draw new parameters from the posterior (IPF)
2. Imputation according to the loglinear model using the set of M parameters

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

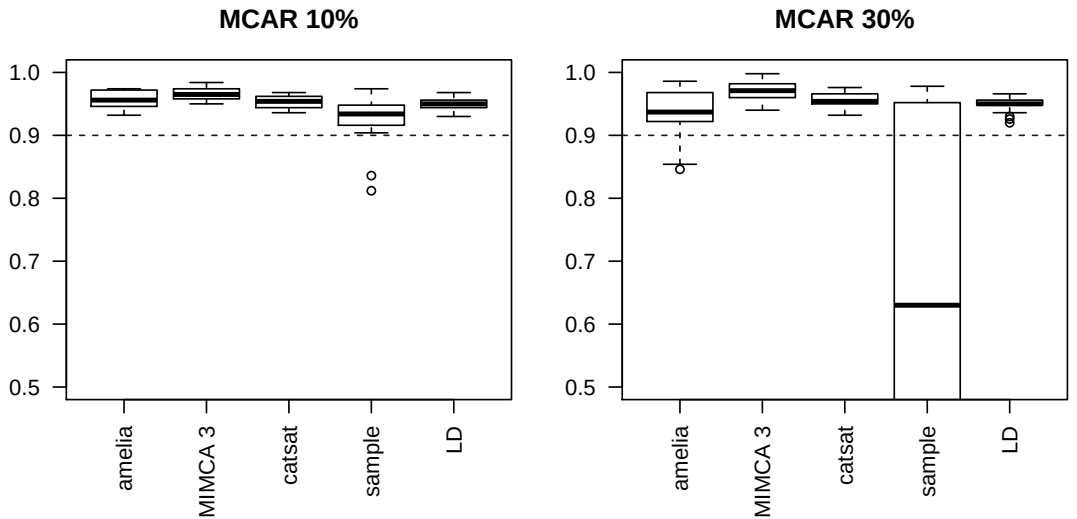
Simulations

- Quantities of interest: θ = parameters of a loglinear model
- 500 simulations
  - data set with 2000 individuals and 5 variables with 3 categories
  - drawn from a loglinear model with two-way associations (1+10+16 parameters)
  - 10% or 30% of missing values using a MCAR or MAR mechanism
  - multiple imputation using $M = 20$ imputed arrays
- Criteria
  - bias, rmse
  - CI width, coverage

Results


MCAR 10%
MCAR 30%
1.0
0.9
0.8
0.7
0.6
0.5
amelia
MIMCA 3
catsat
sample
LD

Results



amelia	*MIMCA3*	*catsat*	*sample*	*LD*
0.6907	0.6888	0.6974	0.6935	0.8255

Conclusion - Perspectives

A new multiple imputation method based on MCA

- strong points:
  - could be applied on data sets of various dimensions
  - robustness to the missing data mechanism
  - perform well in case of low dimensionnal structure
- weak points:
  - a tuning parameter
  - scaling
  - independent draws

- mixed data