Multiple Imputation with MCA

Vincent Audigier & Julie Josse & François Husson

Agrocampus Ouest, Rennes

Rencontres doctorales Lebesgue
Nantes, October 29, 2015

General framework

a	c	d	a	b	b	c	e
a	d	b	d	d	c	a	a
e	f	c	b	b	a	b	a
a	a	a	b	e	b	a	c

General framework

NA	c	d	a	b	NA	NA	e
a	d	b	NA	d	c	a	a
e	NA	c	b	b	a	b	NA
a	a	NA	NA	NA	b	a	c

To apply a statistical method:

- Deletion of individuals: listwise deletion
- Expectation-Maximization
- Multiple imputation

Single imputation

Notations

$$X = \left[\begin{array}{cc|c|ccc} 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 0 & 1 & \dots & 0 & 1 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ 0 & 1 & \dots & 0 & 0 & 1 \\ 0 & 1 & \dots & 0 & 1 & 0 \end{array}\right] \qquad D_\Sigma = \begin{bmatrix} I_1 & & 0 \\ & \ddots & \\ 0 & & I_J \end{bmatrix}$$

$$\text{SVD}\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_\Sigma\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right) \longrightarrow \mathbf{X}_{I\times J} = \mathbf{U}_{I\times J}\Lambda^{1/2}_{J\times J}\mathbf{V}^\top_{J\times J}$$

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

Iterative MCA (Josse *et al.*, 2012)

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 (Josse *et al.*, 2012)

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 (Josse *et al.*, 2012)

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 (Josse *et al.*, 2012)

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 (Josse *et al.*, 2012)

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 $\mathbf{D}_\Sigma$ 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 (Josse *et al.*, 2012)

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 $\mathbf{D}_\Sigma$ 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 (Josse *et al.*, 2012)

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 $\mathbf{D}_\Sigma$ 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

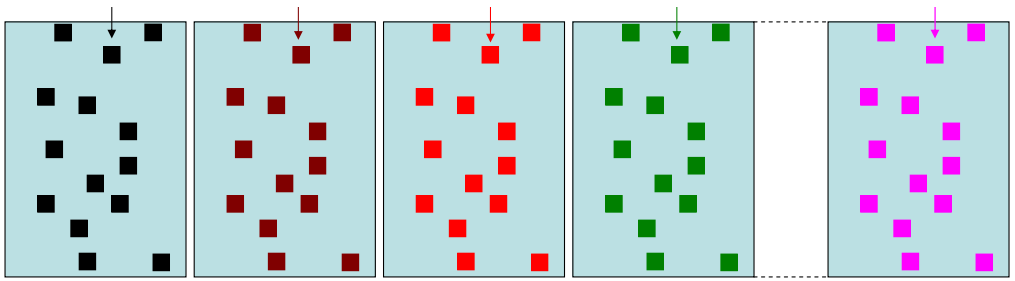
$\pi_{b\|A}$	0.18
$\pi_{a\|A}$	0.82
$\pi_{a\|B}$	0.41
$\pi_{b\|B}$	0.59

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

$\Rightarrow$ 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



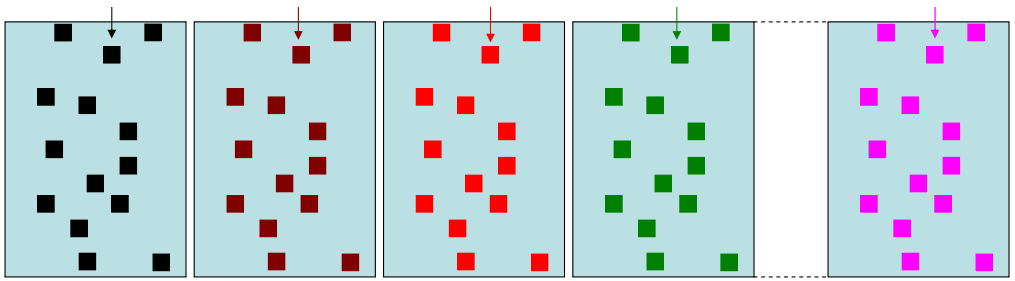
- Perform the analysis on each imputed data set: $\hat{\theta}_m, \widehat{Var}\left(\hat{\theta}_m\right)$
- Combine the results: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$

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

$\Rightarrow$ 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

Multiple imputation with MCA

1. Variability of the parameters of MCA $(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S})$ using a non-parametric bootstrap:

   $\rightarrow$ define M weightings $(R_m)_{1\leq m\leq M}$ for the individuals

Multiple imputation with MCA

1. Variability of the parameters of MCA $(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S})$ using a non-parametric bootstrap:

   $\rightarrow$ define M weightings $(R_m)_{1\leq m\leq M}$ for the individuals

2. Perform iterative MCA using SVD of $\left(\mathbf{X}, \frac{1}{K}(\mathbf{D}_\Sigma)^{-1}, R_m\right)$

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

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

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

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

$\ldots$

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

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

Multiple imputation with MCA

1. Variability of the parameters of MCA $(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S})$ using a non-parametric bootstrap:

   $\rightarrow$ define M weightings $(R_m)_{1\leq m\leq M}$ for the individuals

2. Perform iterative MCA using SVD of $\left(\mathbf{X}, \frac{1}{K}\left(\mathbf{D}_{\Sigma}\right)^{-1}, R_m\right)$

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

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

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

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

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

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

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

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

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

...

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

Properties

Multiple imputation using MCA:

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

- 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)

General principle:

- specify one conditional model per incomplete variable
- incomplete variables are successively imputed
- cycle through variables
- repeat M times

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

Properties:

- More flexible
- Time consuming

Conditional modelling

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

  Properties: captures relationships between pairs of variables

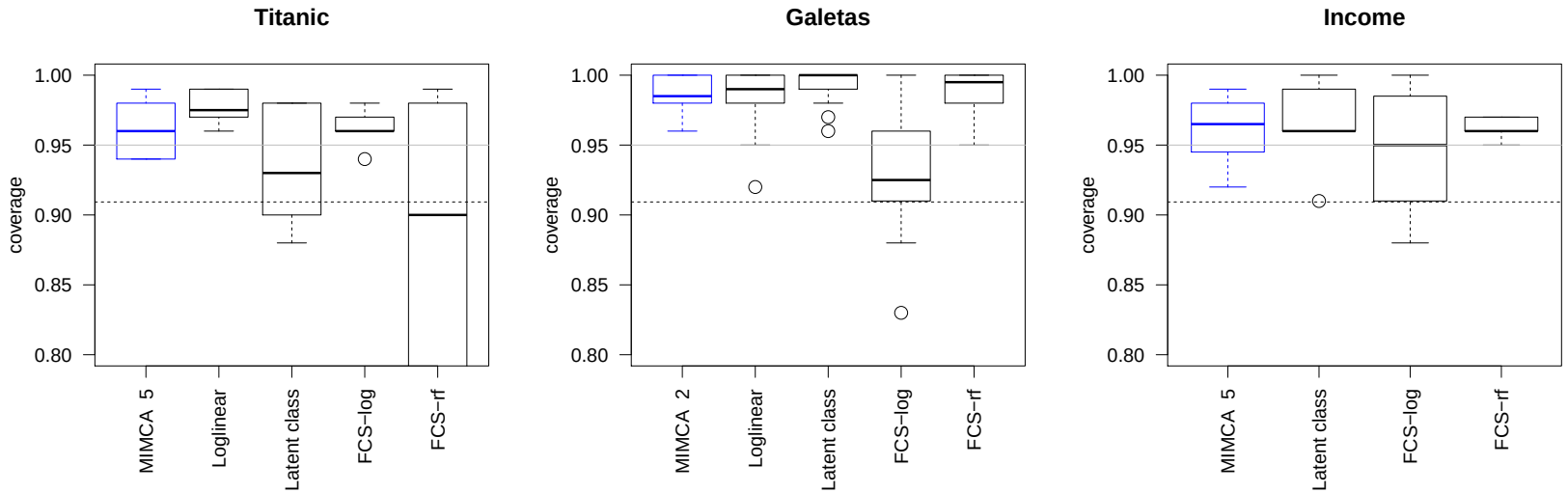
- A recent one: one **random forest**/variable (Doove *et al.*, 2014)

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

Simulations from real data sets

- 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 - 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

A new multiple imputation method based on MCA

Strongest point: dimensionality reduction method

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

From a practical point of view:

- can be applied on data sets of various dimensions (many categories or not / few individuals or not)
- provides correct inferences and performs quickly
- a tuning parameter: the number of dimensions

Perspective:

- mixed data