Multiple imputation using principal component methods

Vincent Audigier & Julie Josse & François Husson

Agrocampus Ouest, Rennes

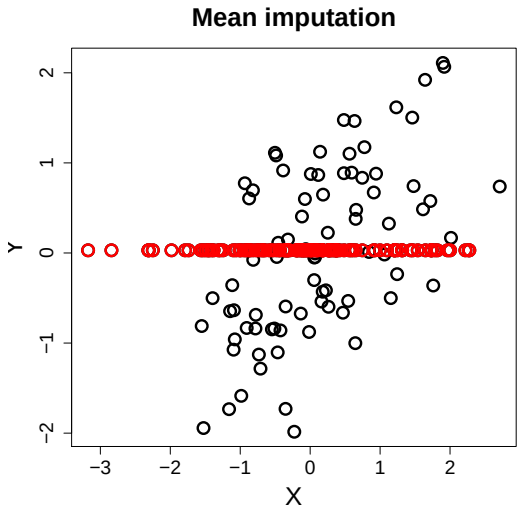
Hôpital Saint Louis, July 15, 2015

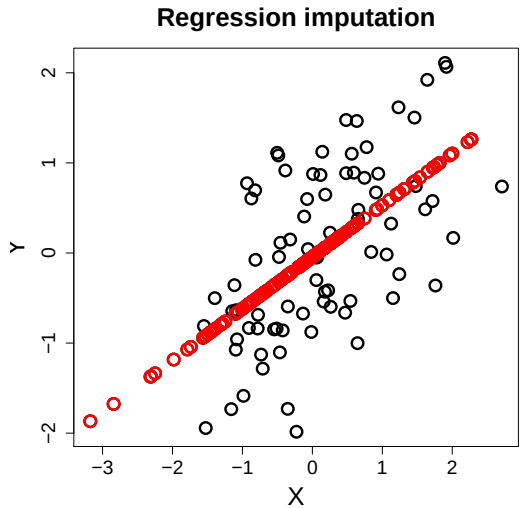
Missing values

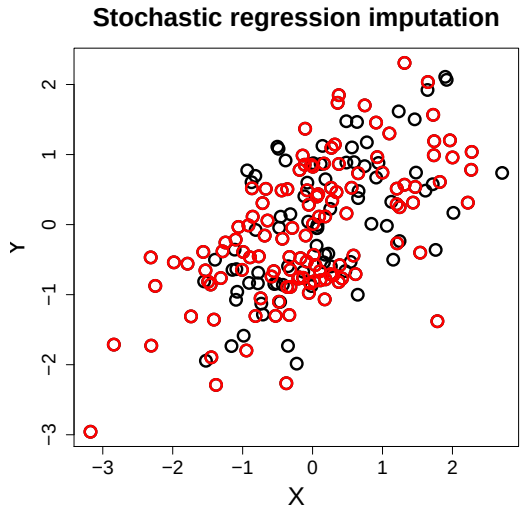
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

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

Single imputation methods





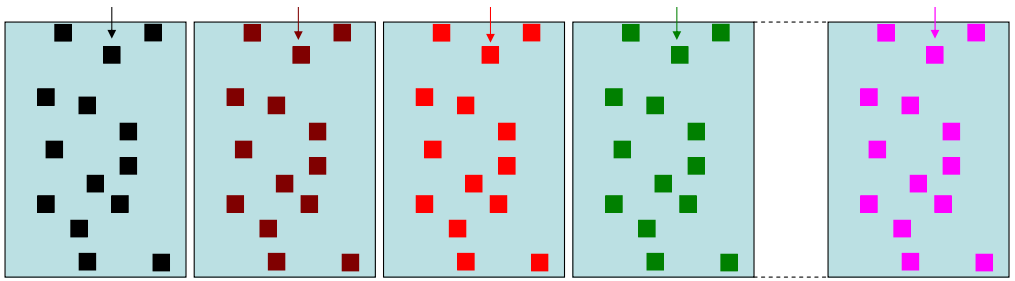


$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\,\mu_y 95\%$	39.4	61.6	70.8

$\Rightarrow$ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) 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



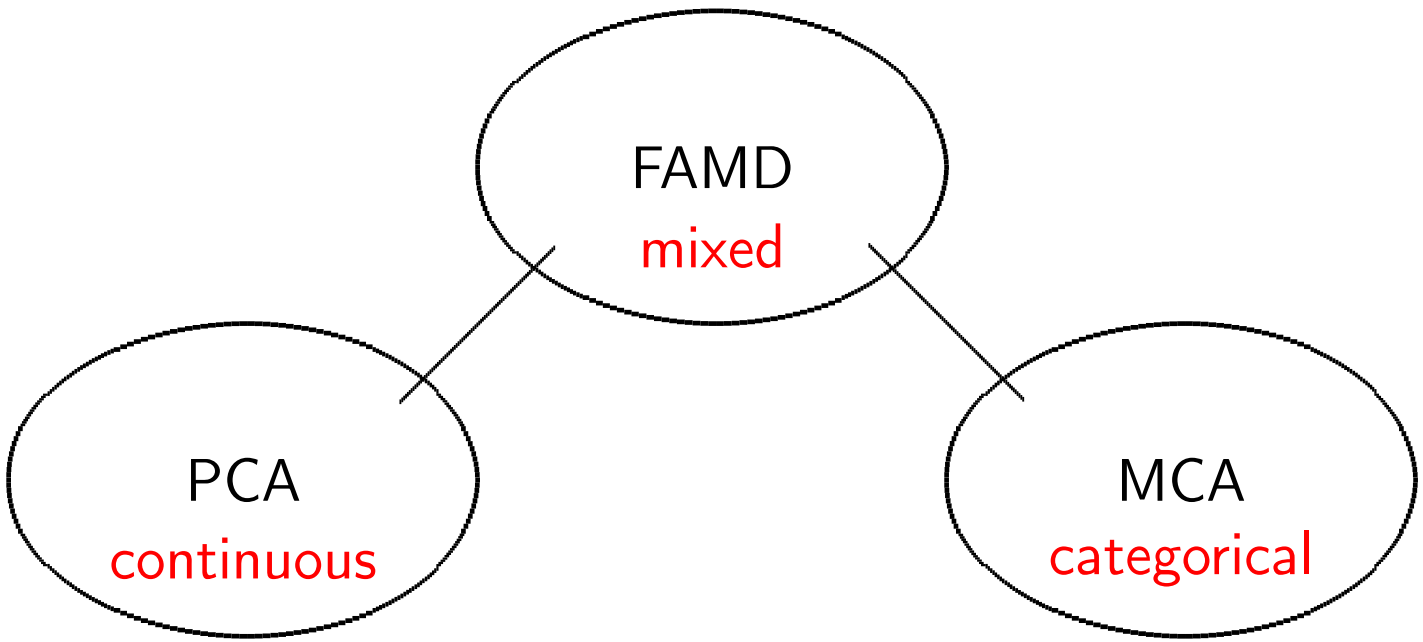
$\rightarrow$ Bootstrap or Bayesian

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

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

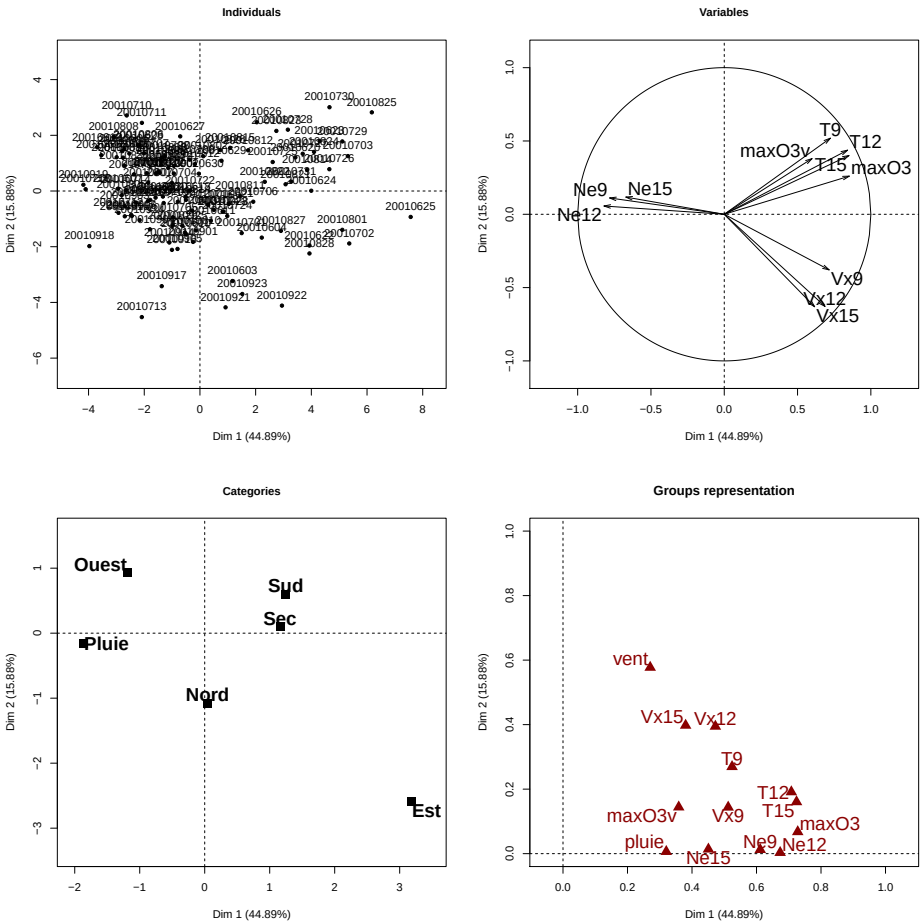
$\Rightarrow$ Aim: provide estimation of the parameters and of their variability

Principal component methods



FAMD

	vent	pluie	maxO3	T9	T12	...
20010602	Nord	Sec	82	17.0	18.4	...
20010603	Est	Sec	92	15.3	17.6	...
20010604	Nord	Sec	114	16.2	19.7	...
20010605	Ouest	Sec	94	17.4	20.5	...
20010606	Ouest	Pluie	80	17.7	19.8	...
...	...	...	...	...	...	...



Properties

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

How to perform FAMD?

FAMD can be seen as the SVD of $\mathbf{X}$ with weights for

- the continuous variables and categories: $(\mathbf{D}_\Sigma)^{-1}$
- the individuals: $\frac{1}{I}\mathbb{1}_I$

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

$$X = \begin{array}{|cc|c||cc|c|ccc|}
\hline
11.04 & \ldots & 2.07 & 1 & 0 & \ldots & 1 & 0 & 0 \\
10.76 & \ldots & 1.86 & 1 & 0 & \ldots & 1 & 0 & 0 \\
11.02 & \ldots & 2.04 & 1 & 0 & \ldots & 1 & 0 & 0 \\
11.02 & \ldots & 1.92 & 0 & 1 & \ldots & 0 & 1 & 0 \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
11.06 & & 2.01 & 0 & 1 & & 0 & 0 & 1 \\
10.95 & & 1.67 & 0 & 1 & & 0 & 1 & 0 \\
\hline
\end{array}$$

$$D_\Sigma = \begin{array}{|c|c|}
\hline
\sigma_{x_1} \;\; \ddots \;\; \sigma_{x_k} & 0 \\
\hline
0 & I_{k+1} \;\; \ddots \;\; I_K \\
\hline
\end{array}$$

How to perform FAMD?

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

- 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: $\hat{\mathbf{X}}_{I\times J} = \hat{\mathbf{U}}_{I\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^\top_{J\times S}$
  $\| \hat{\mathbf{X}} - \mathbf{X} \|^2$ minimized under the constraint of rank S

$X =$

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	2.04	1	0	...	1	0	0
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0	1		0	0	1
10.95		1.67	0	1		0	1	0

$D_\Sigma =$

σ_{x_1} ... σ_{x_k}	0
0	I_{k+1} ... I_K

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (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) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

$\rightarrow$

NA	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	NA	1	0	...	NA	NA	NA
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	NA	NA		0	0	1
NA		1.67	0	1		0	1	0

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (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) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

11.01	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	1.89	1	0	...	0.61	0.19	0.20
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.32	0.68		0	0	1
11.01		1.67	0	1		0	1	0

FAMD with missing values

Iterative FAMD algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (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) $\mathbf{D}_\Sigma$ is updated

NA	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	NA	A	...	NA
11.02	...	1.92	B	...	B
11.06		2.01	NA	...	C
NA		1.67	B	...	B

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	2.04	1	0	...	0.81	0.05	0.14
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.25	0.75		0	0	1
10.95		1.67	0	1		0	1	0

Single imputation with FAMD

Iterative FAMD algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation of the parameters of FAMD
   $\rightarrow$ SVD of $\left(\mathbf{X}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$
   (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) $\mathbf{D}_\Sigma$ is updated

11.04	...	2.07	A	...	A
10.76	...	1.86	A	...	A
11.02	...	2.04	A	...	A
11.02	...	1.92	B	...	B
11.06		2.01	B	...	C
10.95		1.67	B	...	B

$\leftarrow$

11.04	...	2.07	1	0	...	1	0	0
10.76	...	1.86	1	0	...	1	0	0
11.02	...	2.04	1	0	...	0.81	0.05	0.14
11.02	...	1.92	0	1	...	0	1	0
11.06		2.01	0.25	0.75		0	0	1
10.95		1.67	0	1		0	1	0

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

Imputation with Random Forests

A random forest:

- Each tree is built from a subset of observations
- Fitted values are pooled

Imputation with RF:

1. Initial imputation: mean imputation - frequent category
2. Fit a RF X_j^{obs} on the other variables X_{-j}^{obs}
   Predict X_j^{miss} using the trained R on X_{-j}^{miss}
3. Cycle through variables
4. Repeat step 2 and 3 until convergence

Implemented: `R` package `missForest` (Stekhoven)

Simulation study

Several data sets

- Relationships between variables
- Number of categories
- percentage of missing values (10%,20%,30%)

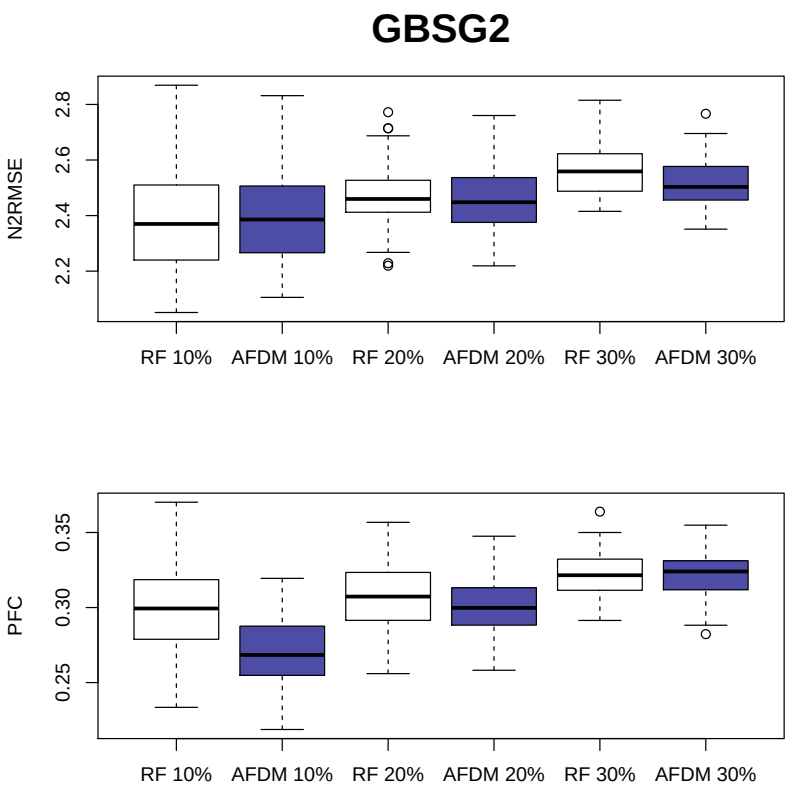
Criteria:

- for continuous data:

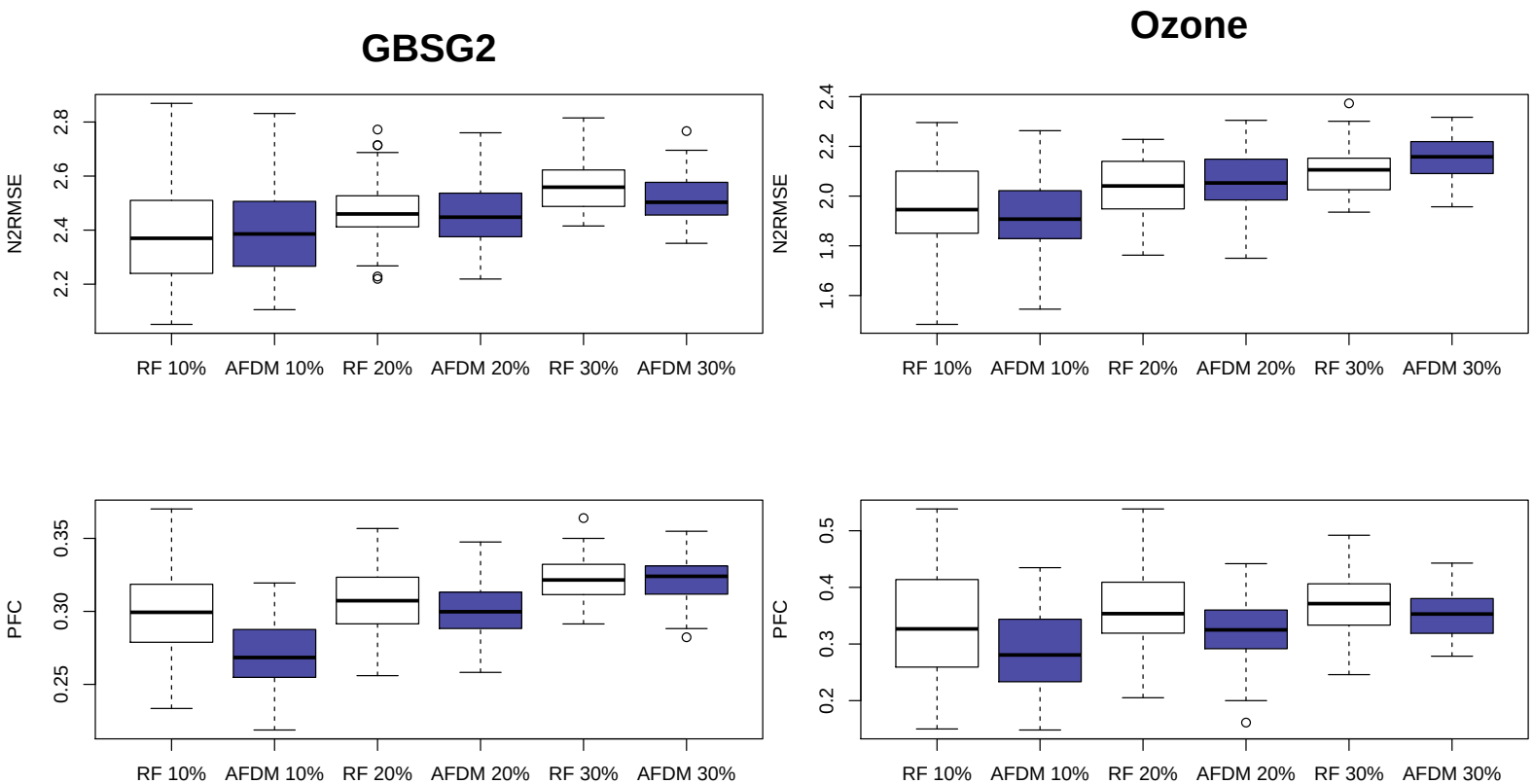
$$NRMSE = \sqrt{\sum_{i \in missing} \frac{mean\left(\left(X_i^{true} - X_i^{imp}\right)^2\right)}{var\left(X_i^{true}\right)}}$$

- for categorical data: proportion of falsely classified entries

Comparisons from real data sets



Comparisons from real data sets



Summary

A single imputation procedure using dimensionality reduction

- using relationships between pairs of variables
- using similarities between individuals
- showing good performances
  - on real data
  - when the relationships between continuous variables are linear
  - to impute rare categories
- imputing mixed, continuous or categorical data

Summary

A single imputation procedure using dimensionality reduction

- using relationships between pairs of variables
- using similarities between individuals
- showing good performances
  - on real data
  - when the relationships between continuous variables are linear
  - to impute rare categories
- imputing mixed, continuous or categorical data

But a single imputation method only

From single imputation to multiple imputation

1. Reflect the variability on the parameters of the imputation model

   $\rightarrow \left(\left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_1, \ldots, \left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_M\right)$

   Bayesian or Bootstrap

2. Add a disturbance on the prediction by $\hat{X}_m = \hat{\mathbf{U}}_m \hat{\Lambda}^{1/2}_m \hat{\mathbf{V}}^{\top}_m$

   $\rightarrow$ need to distinguish continuous and categorical data

Multiple imputation for continuous data using PCA

A set of parameters:

$$\left(\left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_1, \ldots, \left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_M\right)$$

$$\Updownarrow$$

$$\left(\left(\hat{\mathbf{X}}\right)_1, \ldots, \left(\hat{\mathbf{X}}\right)_M\right)$$

obtained using a Bayesian approach:

- A model for PCA (complete case)
- A Bayesian formulation of the model (complete case)
- A way to draw in the posterior with missing values

PCA model (Caussinus, 1986)

Model

$$\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p}$$

with $\tilde{\mathbf{X}}$ a low rank matrix, $\varepsilon \sim \mathcal{N}\left(0, \mathbb{1}_p \sigma^2\right)$

Maximum Likelihood:

$$\hat{\mathbf{X}}^S = \widehat{\mathbf{U}}_{n\times S} \widehat{\mathbf{\Lambda}}^{\frac{1}{2}}_{S\times S} \widehat{\mathbf{V}}^{\top}_{p\times S} \quad \rightarrow \quad \widehat{\sigma^2} = \| \mathbf{X} - \widehat{\mathbf{X}}^S \|^2 \text{ /degrees of f.}$$

Bayesian PCA (Verbanck *et al.*, 2013)

$$\begin{aligned}
\text{Model: } \mathbf{X}_{n\times p} &= \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p} \\
x_{ij} &= \tilde{x}_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\
&= \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij} \\
&= \sum_{s=1}^{S} \tilde{x}_{ij}^{(s)} + \varepsilon_{ij}
\end{aligned}$$

Prior: $\tilde{x}_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$

Posterior: $\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) \sim \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$ with $\Phi_s = \frac{\tau_s^2}{\tau_s^2 + \sigma^2}$

Empirical Bayes for τ_s^2: $\hat{\tau}_s^2 = \left(\hat{\lambda}_s - \hat{\sigma}^2\right)$ (max likelihood)

$$\hat{\Phi}_s = \frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s} = \frac{\text{signal variance}}{\text{total variance}}$$

Multiple imputation Bayes-PCA

1. Variability of the parameters, M plausible $(\hat{x}_{ij})^1, \ldots, (\hat{x}_{ij})^M$
   - Posterior distribution: Bayesian PCA

$$\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$$

   - Data Augmentation (Tanner and Wong, 1987)
     (0) Max likelihood estimate
     (I) impute using the PCA model
     (P) draw new parameters from the posterior

2. Imputation according to the PCA model using the set of M parameters

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: $(\mu^1, \Sigma^1), \ldots, (\mu^M, \Sigma^M)$
  2. Imputation: $x_{i.}^m$ drawn from $\mathcal{N}(\mu^m, \Sigma^m)$

- Implemented: `R` package `Amelia` (J. Honaker, G. King, M. Blackwell)

Fully conditional modelling using Bayesian regressions

- Hypothesis: one model/variable
  all variables continuous and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

- Algorithm:

  One variable with missing values:

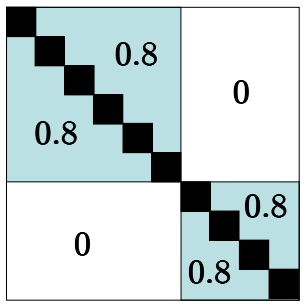
  1. Bayesian approach: (β^m, σ^m)
  2. Imputation: stochastic regression x_{ij}^m drawn from $\mathcal{N}(X_{-j}\beta^m, \sigma^m)$

  Many variables with missing values: cycles through variables

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

Simulations

- Quantities of interest: $\theta_1 = \mathbb{E}[Y], \theta_2 = \beta_1, \theta_3 = \rho$
- 1000 simulations
  - data set drawn from $\mathcal{N}_p(\mu, \Sigma)$ with a two-block structure, varying n (30 or 200), p (6 or 60) and ρ (0.3 or 0.9)



  - 10% or 30% of missing values using a MCAR mechanism
  - multiple imputation using $M = 20$ imputed arrays
- Criteria
  - bias
  - CI width, coverage

Results for the expectation

	parameters				confidence interval width			coverage		
	n	p	ρ	%	*Amelia*	*MICE*	*BayesMIPCA*	*Amelia*	*MICE*	*BayesMIPCA*
1	30	6	0.3	0.1	0.803	0.805	0.781	0.955	0.953	0.950
2	30	6	0.3	0.3		1.010	0.898		0.971	0.949
3	30	6	0.9	0.1	0.763	0.759	0.756	0.952	0.95	0.949
4	30	6	0.9	0.3		0.818	0.783		0.965	0.953
5	30	60	0.3	0.1			0.775			0.955
6	30	60	0.3	0.3			0.864			0.952
7	30	60	0.9	0.1			0.742			0.953
8	30	60	0.9	0.3			0.759			0.954
9	200	6	0.3	0.1	0.291	0.294	0.292	0.947	0.947	0.946
10	200	6	0.3	0.3	0.328	0.334	0.325	0.954	0.959	0.952
11	200	6	0.9	0.1	0.281	0.281	0.281	0.953	0.95	0.952
12	200	6	0.9	0.3	0.288	0.289	0.288	0.948	0.951	0.951
13	200	60	0.3	0.1		0.304	0.289		0.957	0.945
14	200	60	0.3	0.3		0.384	0.313		0.981	0.958
15	200	60	0.9	0.1		0.282	0.279		0.951	0.948
16	200	60	0.9	0.3		0.296	0.283		0.958	0.952

Properties for multiple imputation with Bayes-PCA

- good inferences for regression coefficient, correlation, mean
- works in a large range of configurations:
  - $n < p$ or $n > p$
  - strong or weak relationships
  - low or high percentage of missing values

Multiple imputation for categorical data using MCA

A set of parameters:

$$\left(\left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_1, \ldots, \left(\hat{\mathbf{U}}_{I\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S}\right)_M\right)$$

obtained using a non-parametric Bootstrap approach:

1. Generate M bootstrap replicates
2. Estimate the parameters on each incomplete replicate
3. Add uncertainty on the prediction

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

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. Estimate MCA parameters 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

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. Estimate MCA parameters 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

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

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

...

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

majority $\Rightarrow$ lack of 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

2. Estimate MCA parameters 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

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

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

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

$\ldots$

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

MI 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)

MI using a DPMPM model (Si and Reiter, 2013)

- Hypothesis: $\mathbb{P}(X = (x_1, \ldots, x_K); \theta) = \sum_{\ell=1}^{L} \left(\theta_\ell \prod_{k=1}^{K} \theta_{x_k}^{(\ell)} \right)$

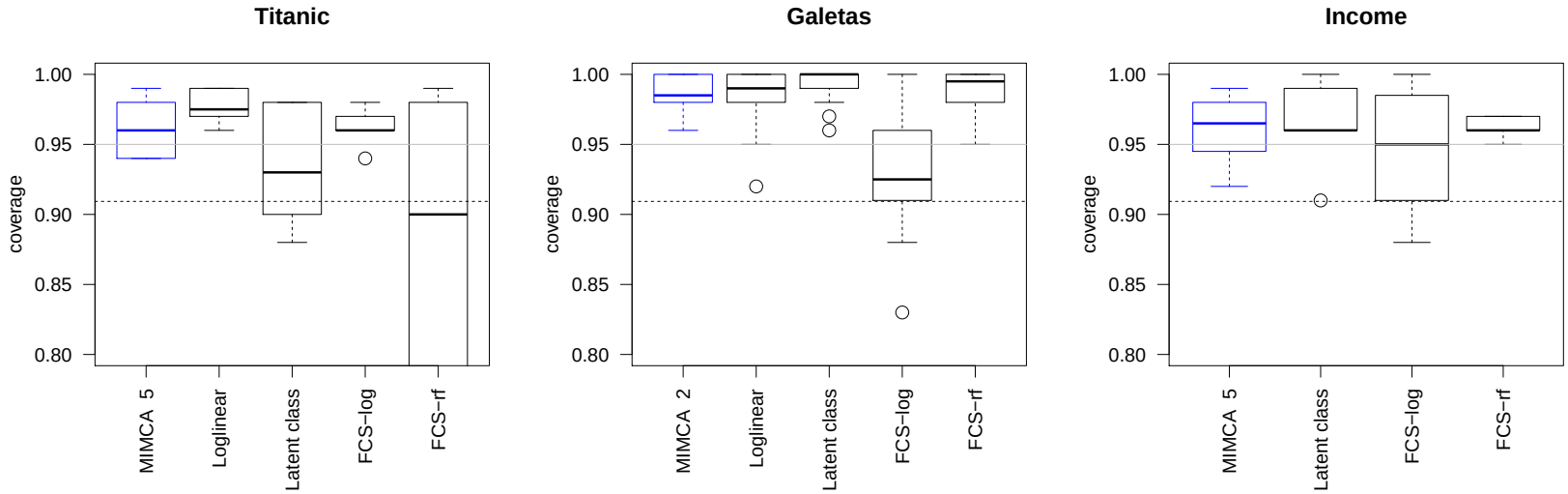
1. Variability of the parameters:
   - a hierarchic prior on θ:
     $\alpha \sim \mathcal{G}(.25, .25) \quad \zeta_\ell \sim \mathcal{B}(1, \alpha) \quad \theta_\ell = \zeta_\ell \prod_{g<\ell} (1 - \zeta_g)$ for ℓ in $1, \ldots, \infty$
   - posterior on θ: untractable
     $\rightarrow$ Gibbs sampler and Data Augmentation
2. Imputation according to the mixture model using the set of M parameters

- Implemented: `R` package `mi` (Gelman *et al.*)

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 - 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

Multiple imputation methods for continuous and categorical data using dimensionality reduction method

Properties:

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

From a practical point of view:

- can be applied on data sets of various dimensions
- provides correct inferences for analysis model based on relationships between pairs of variables
- requires to choose the number of dimensions S

Perspective:

- mixed data