Multiple imputation

Vincent Audigier

CNAM, Paris

JES 2021

Imputation

- Imputation consists in replacing each missing value by (a) plausible value(s)
- Objective: applying a statistical analysis despite missing values
- Two kinds of methods
  - single imputation: replacement by a unique one value
  - multiple imputation: replacement by several values

Missing values

Two kinds of missing values

- values which are not observed, but exist Ex: temperature is missing for one patient
- values that are not observed because they would have no sense! Ex: date of death is missing for a living patient

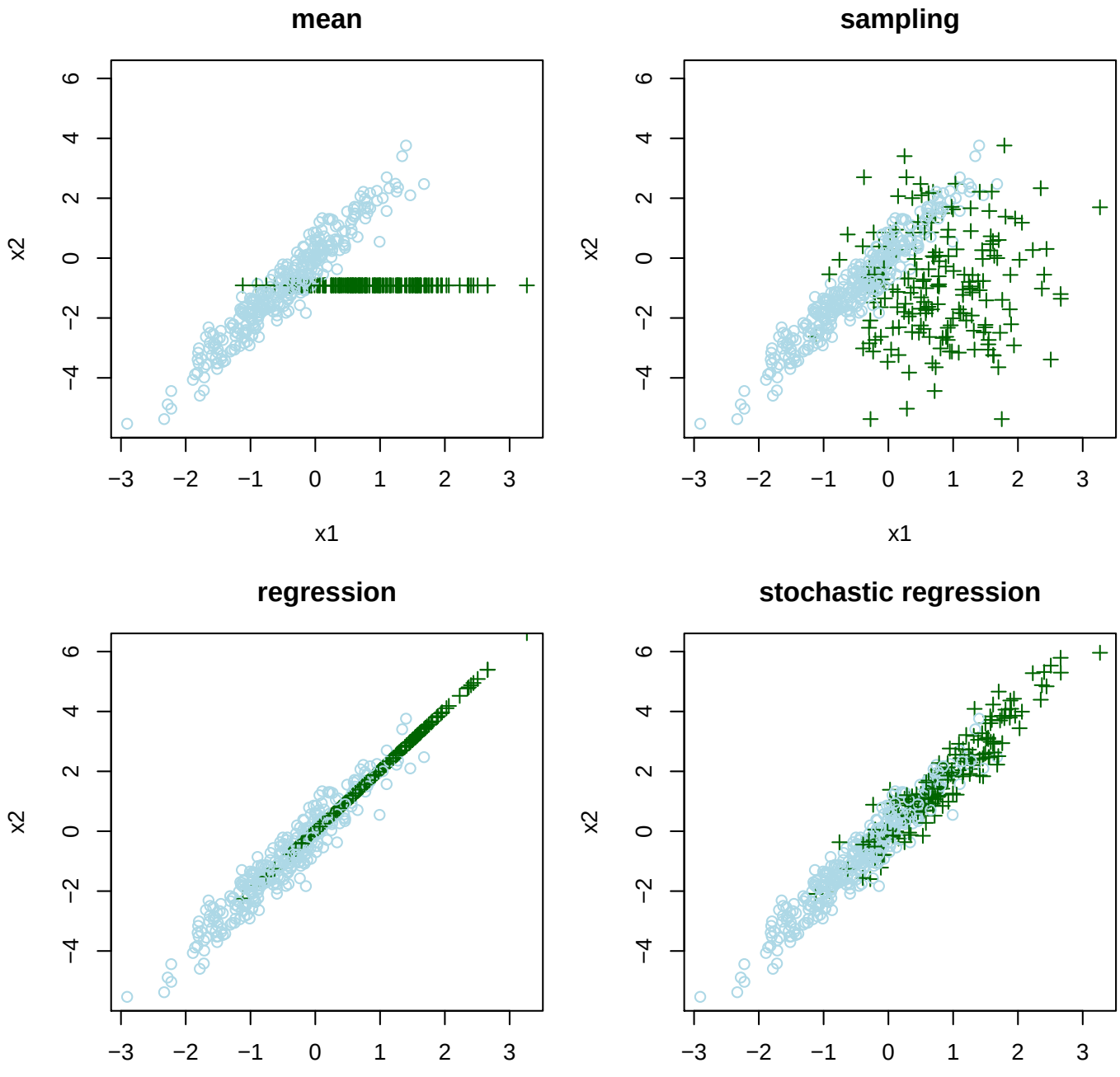
Here, we focus on values which are not observed, but exist.

Single imputation

Assuming only one (continuous) variable is incomplete, several classical methods can be used

- mean
- median
- sampling observed data
- regression
- stochastic regression
- PCA

Examples

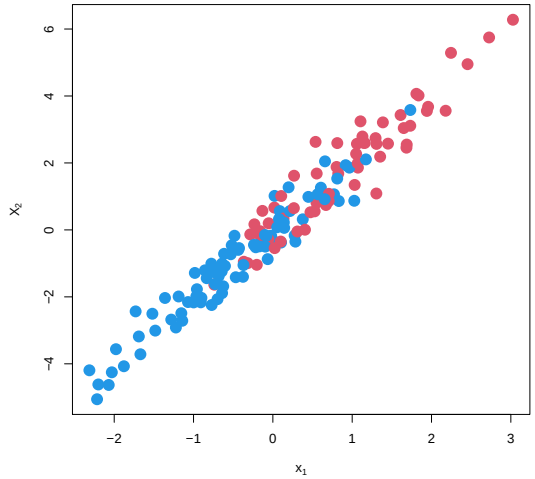

mean
sampling
regression
stochastic regression
x1
x2
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
6

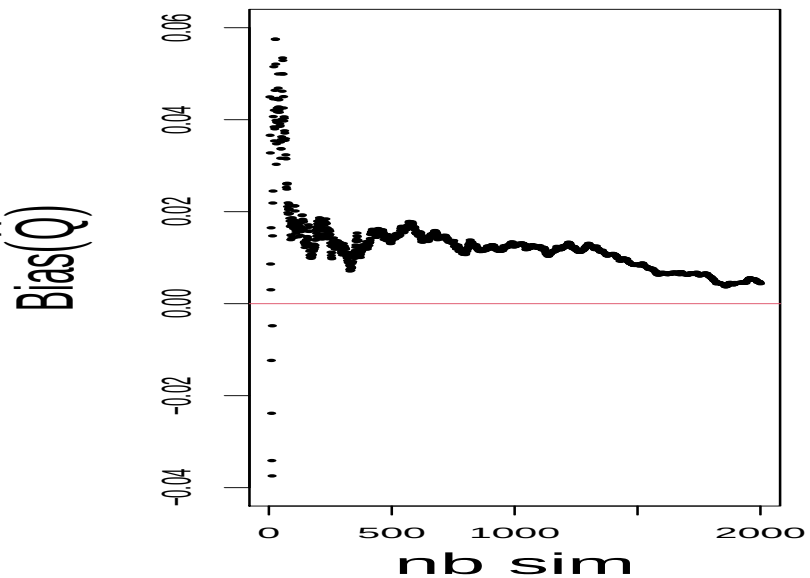
Typology of single imputation methods

- parametric (ex: stochastic regression)
  - advantages: performs well on small datasets
  - drawbacks: sensitive to the model specification
- non-parametric (ex: knn, random forest)
  - advantages: preserves the nature of the variables
  - drawbacks: requires a large number of individuals
- semi-parametric (ex: predictive mean matching)
  - advantages: preserves the nature of the variables, more robust to model misspecification
  - drawbacks: requires a moderate number of individuals

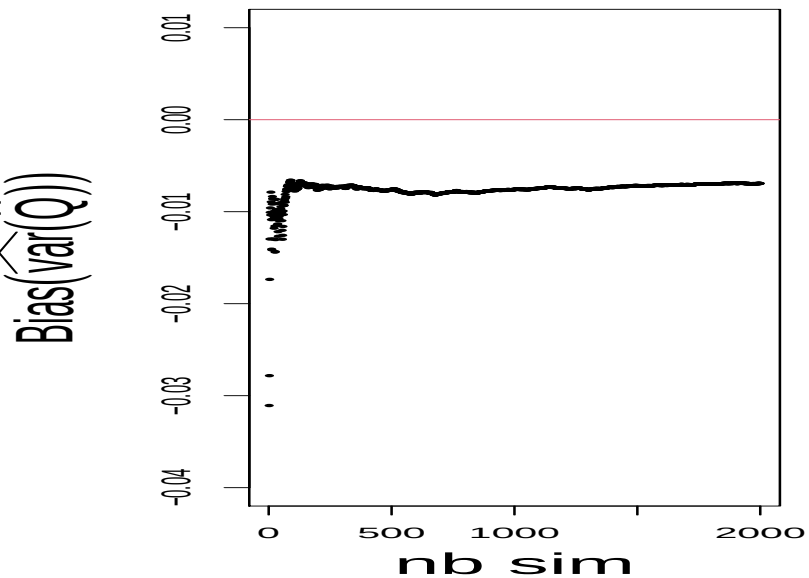
Single imputation is a limited approach

- $n = 150$, $p = 2$
- missing values on X_2 (MAR)
- parameter: $Q = \mathbb{E}[X_2]$





- The estimator of Q is unbiased



- The estimator of $Var\left(\widehat{Q}\right)$ is downwardly biased

Outline

Notations and vocabulary (1)

- n: number of individuals
- p: number of variables
- $\mathbf{X}_{n\times p}$: the full data matrix (partially unknown)
- $\mathbf{R}_{n\times p}$: the missing data pattern $\mathbf{R} = (r_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ with $r_{ij} = 0$ if x_{ij} is missing and 1 otherwise
- x_i^{obs} observed profile of the individual i et x_i^{miss} the unobserved profile

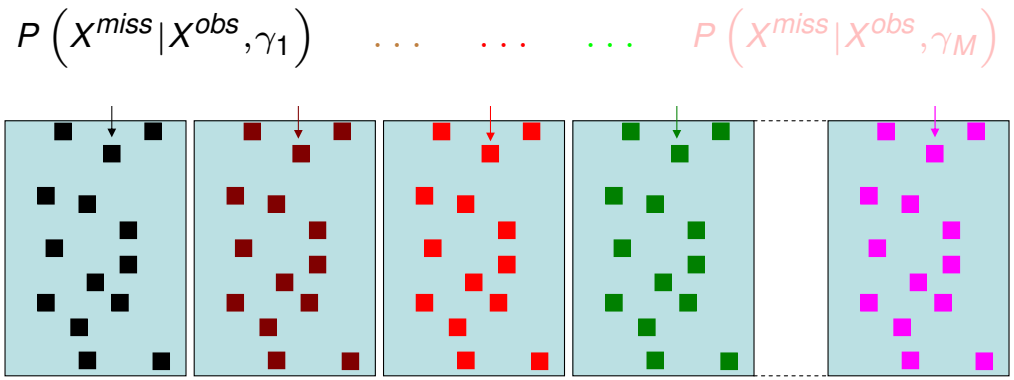
Notations and vocabulary (2)

$\mathbf{X}_{n\times p}$, $\mathbf{R}_{n\times p}$, x_i^{obs} et x_i^{miss} can be seen as realisations of random variables

- $X = (X_1, \ldots, X_p)$: random variables associated to $\mathbf{X}_{n\times p}$
- $R = (R_1, \ldots, R_p)$ random variables associated to $\mathbf{R}_{n\times p}$
- X^{obs} and X^{miss} : random variables associated to observed and unobserved parts of X so that $X = \left(X^{obs}, X^{miss}\right)$
- $f(X;\gamma)$ complete data distribution

Multiple imputation (Rubin, 1987)

1. Generate a set of M parameters $(\gamma_m)_{1 \leq m \leq M}$ of an imputation model to generate M plausible imputed data sets



2. Fit the analysis model on each imputed data set

3. Combine the results using Rubin's rules

$\Rightarrow$ Provide estimation of the parameters and of their variability

Theoretical foundations (1)

- Q a quantity of interest (mean, correlation coefficient, regression coefficient, ...)
- Adopting a bayesian point of view, we aim to infer $f(Q|\mathbf{X}^{obs})$ ($f(Q|\mathbf{X}^{obs},R)$ in the non-ignorable case)
- However, $f(Q|\mathbf{X}^{obs})$ is generally intractable...

$$f\left(Q|\mathbf{X}^{obs}\right) = \int \underbrace{f\left(Q|\mathbf{X}^{obs},\mathbf{X}^{miss}\right)}_{posterior} \underbrace{f\left(\mathbf{X}^{miss}|\mathbf{X}^{obs}\right)}_{predictive\ distribution} d\mathbf{X}^{miss}$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} f\left(Q|\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right)$$

Theoretical foundations (2)

We are generally interested in the expectation and variance of the posterior distribution only

- $E\left(Q|\mathbf{X}^{obs}\right) = E\left(E\left(Q|\mathbf{X}^{obs},\mathbf{X}^{miss}\right)|\mathbf{X}^{obs}\right)$

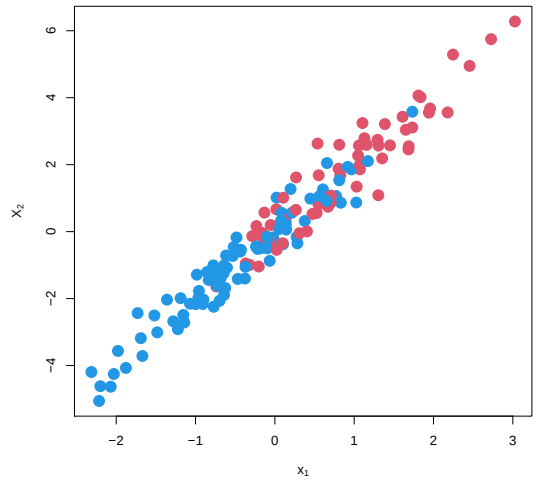
  $Q^{MI} \approx \frac{1}{M}\sum_{m=1}^{M}\widehat{Q}_m$

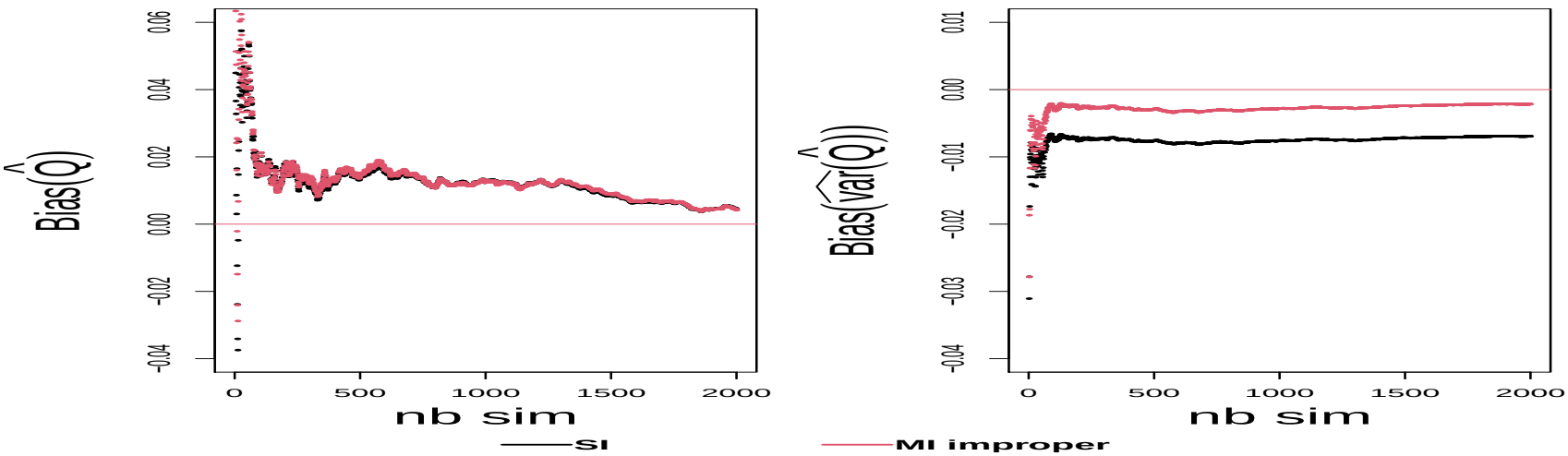
- $V\left(Q|\mathbf{X}^{obs}\right) = E\left(V\left(Q|\mathbf{X}^{obs},\mathbf{X}^{miss}\right)|\mathbf{X}^{obs}\right) + V\left(E\left(Q|\mathbf{X}^{obs},\mathbf{X}^{miss}\right)|\mathbf{X}^{obs}\right)$

  $T \approx \frac{1}{M}\sum_{m=1}^{M}U_m + \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{Q}_m - Q^{MI}\right)^2$

MI $= M \times$ SI?

- $n = 150$, $p = 2$
- missing values on X_2 (MAR)
- parameter: $\mathbb{E}[X_2]$





Theory assumes $f\left(\mathbf{X}^{miss}|\mathbf{X}^{obs}\right)$ is known

Predictive distribution

$$f\left(\mathbf{X}^{miss}|\mathbf{X}^{obs}\right) = \int f\left(\mathbf{X}^{miss}|\mathbf{X}^{obs},\gamma\right) f\left(\gamma|\mathbf{X}^{obs}\right) d\gamma$$

Imputation step consists in

1. generating M values of γ from $f\left(\gamma|\mathbf{X}^{obs}\right)$
2. imputing $\mathbf{X}^{miss}$ from $f\left(\mathbf{X}^{miss}|\mathbf{X}^{obs};\gamma_m\right)$

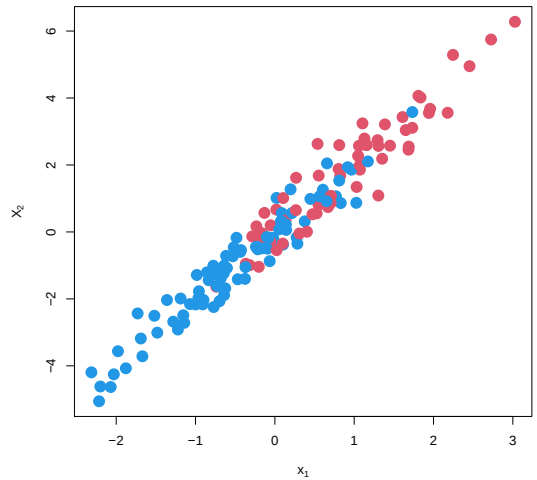
By following these two steps, the MI procedure is said *proper*

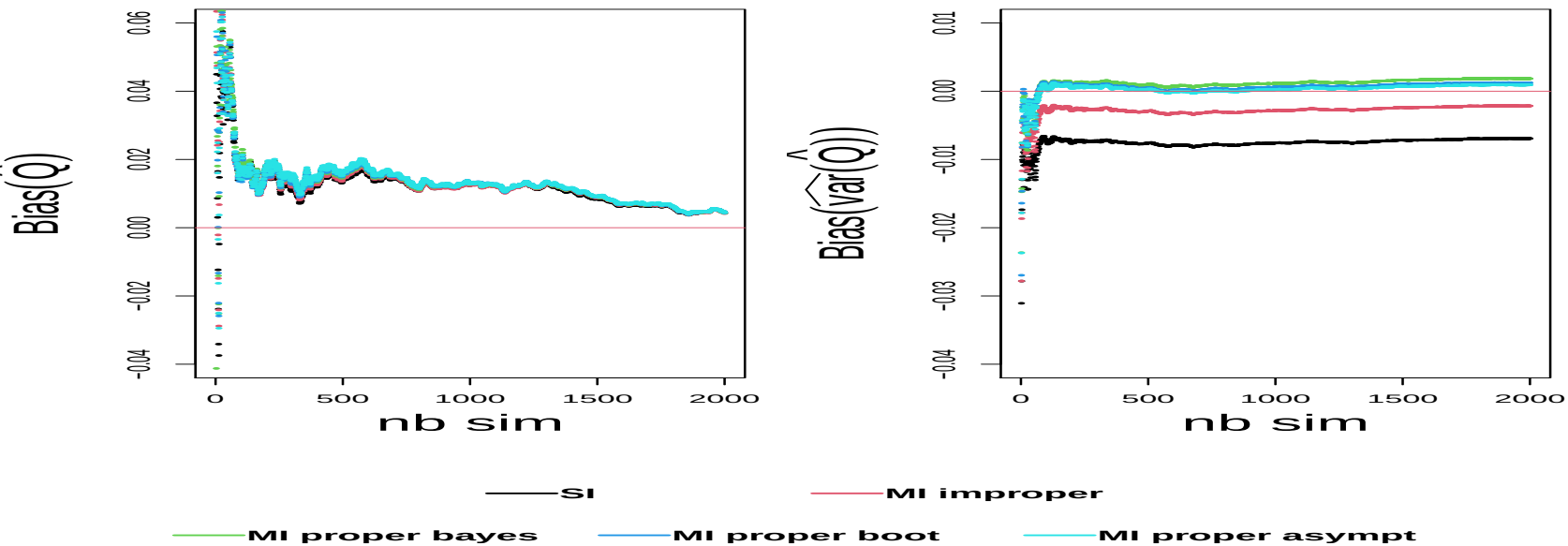
Generation of $(\gamma_m)_{1 \leq m \leq M}$

- Bayesian
  - Prior distribution $p(\gamma)$
  - Derive the posterior distribution $p(\gamma|Z^{obs})$ (Data-Augmentation)
  - Draw from $p(\gamma|Z^{obs})$ M times
- Non-parametric Bootstrap
  - Sampling observations with replacement M times
  - Estimate γ_m from each bootstrap sample (EM)
- Sampling from the asymptotic distribution
  - For asymptotically Gaussian estimator, estimate mean and variance
  - Draw M values from $\mathcal{N}\left(\widehat{\gamma}, \widehat{\text{Var}}\left(\widehat{\gamma}\right)\right)$

Example

- $n = 150$, $p = 2$
- missing values on X_2 (MAR)
- parameter: $\mathbb{E}[X_2]$





Outline

The gold standard (Schafer, 1997)

Some imputation models for continuous data

Many imputation models based on the multivariate normal distribution (*Joint modelling*)

	Model	R package
Schafer (1997)	Gaussian model / Bayesian	norm
Honaker et al. (2011)	Gaussian model / Bootstrap	Amelia
Quartagno and Carpenter (2016)	multivariate LMM/ Bayesian	jomo
Schafer (1997)	multivariate LMM/ Bayesian	pan

However

- the Gaussian distribution can be inappropriate with a moderate/large number of variables
- models have many parameters leading to overfitting with a small number of observations

Fully conditional specification

Instead of specifying one joint distribution $P(X;\gamma)$, a conditional distribution is specified for each (incomplete) variable $P\left(X_j|X_{-j};\gamma_j\right)$

Ex : $P\left(X_j|X_{-j};\gamma_j\right) = \mathcal{N}(X_{-j}\beta,\sigma^2) \quad \gamma_j = (\beta,\sigma)$

To impute the mth data set

- initialize x_i^{miss} for all i
- for j in 1 ... p
  a. generate γ_j based on observed individuals on X_j
  b. impute X_j^{miss} according to $P(X_j|X_{-j};\gamma_j)$
- repeat until convergence

Relationship between JM and FCS

FCS is close to a Gibbs sampler, but

- In a Gibbs sampler, a joint distribution $P(X; \gamma)$ is specified and conditional models $P\left(X_j | X_{-j}; \gamma_j\right)$ $(1 \leq j \leq p)$ are accordingly chosen
- In FCS, conditional models are specified and convergence to an unknown joint distribution is expected

$\Rightarrow$ When conditional imputation models are specified from a joint model, FCS MI is equivalent to JM MI.

Ex: JM by multivariate gaussian model $=$ FCS by linear regression (Hughes et al., 2014)

FCS pros and cons

Pros

- sparsity
- accounting for interactions effects
- addressing outliers
- semi or non-parametric models

Cons

- time consuming
- no theoretical guaranties (except in specific cases)
- checking convergence is not possible with a large number of variables

Predictive mean matching

Instead of drawing $X_j^{miss} \sim P(X_j | X_{-j}; \gamma_j)$

1. identify a set of D donors (complete on X_j) according to a *matching type*
2. draw a donor d
3. impute x_{ij}^{miss} according to x_{dj}^{obs}

Matching types

	$\widehat{x}_{ij}$ (receveur)	$\widehat{x}_{dj}$ (donneur)
Type 0 :	$\widehat{\beta}_0 + \sum_{j' \neq j} \widehat{\beta}_{j'} x_{ij'}$	$\widehat{\beta}_0 + \sum_{j' \neq j} \widehat{\beta}_{j'} x_{dj'}$
Type 1 :	$\widehat{\beta}_0 + \sum_{j' \neq j} \widehat{\beta}_{j'} x_{ij'}$	$\dot{\beta}_0 + \sum_{j' \neq j} \dot{\beta}_{j'} x_{dj'}$
Type 2 :	$\dot{\beta}_0 + \sum_{j' \neq j} \dot{\beta}_{j'} x_{ij'}$	$\dot{\beta}_0 + \sum_{j' \neq j} \dot{\beta}_{j'} x_{dj'}$
Type 3 :	$\dot{\beta}_0 + \sum_{j' \neq j} \dot{\beta}_{j'} x_{ij'}$	$\ddot{\beta}_0 + \sum_{j' \neq j} \ddot{\beta}_{j'} x_{dj'}$

Properties

- similar performances when the imputation model is well specified
- better performances otherwise

Outline

Congeniality

- In an ideal world, the imputation model would be based on the true data model. Then, any analysis model could be applied
- In a real world, the true distribution for $P(X;\gamma)$ is unknown... and the imputation model is usually misspecified
- To avoid a bias due to the imputation step, the imputation model should be in lines with the assumptions related to the analysis model

$\Rightarrow$ Both models should be *congenial*

Example (Schafer, 1997)

- model 1: $Z_3 = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$
- model 2: $Z_3 = \beta_0 + \beta_1 Z_1 + \varepsilon$

If the true model is (2)

- (2) + (2) $\rightarrow$ congenial
- (1) + (2) $\rightarrow$ uncongenial, but conservative

If the true model is (1)

- (1) + (1) $\rightarrow$ congenial
- (2) + (1) $\rightarrow$ uncongenial, biased

$\Rightarrow$ The imputation model should not impose restrictions on unknown parameters

Model fitting

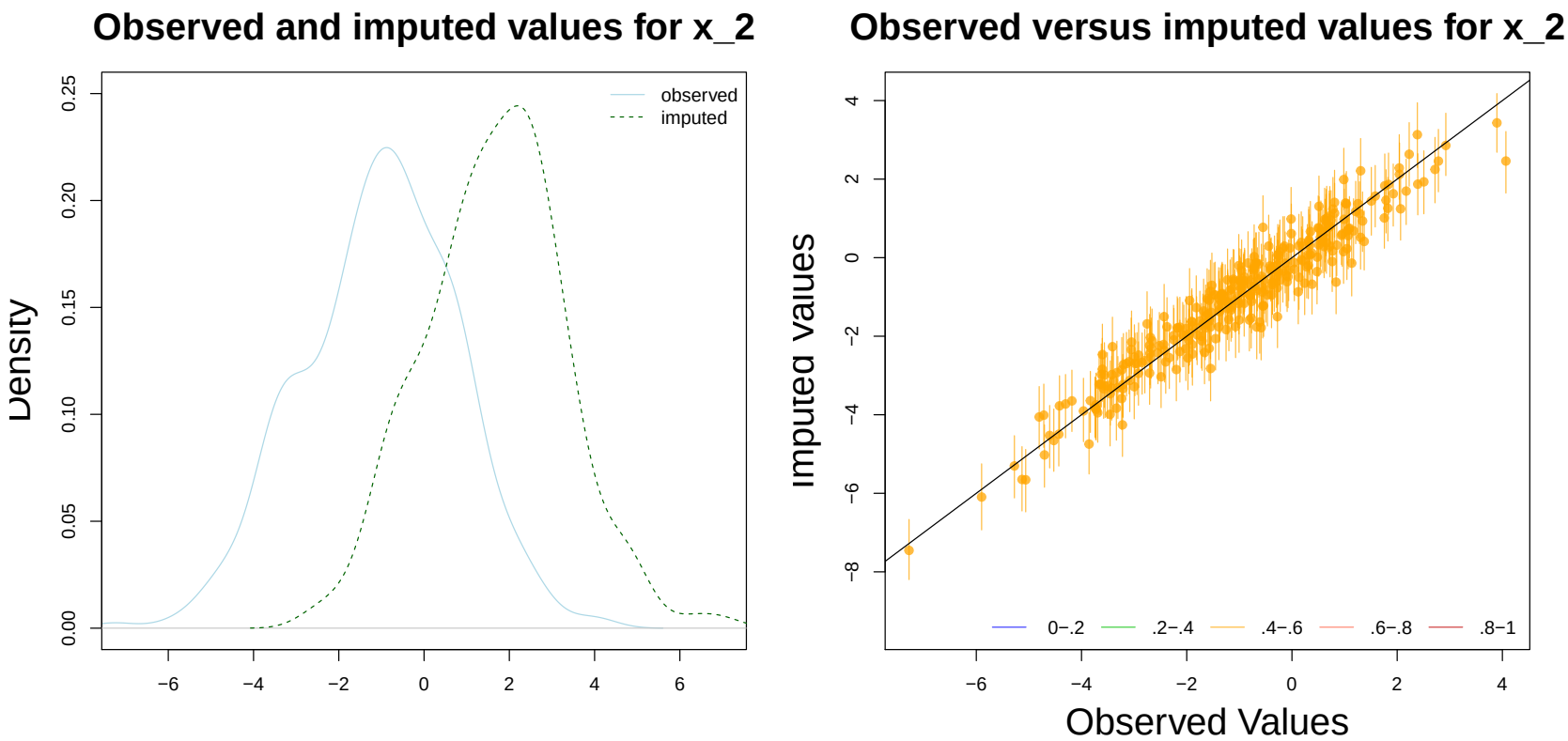


FIG.: Comparative distributions and overimputation

Outline

JM

Categorical data

- log-linear model (Schafer, 1997)
- latent class model (Vidotto et al., 2015)
- multiple correspondences analysis (Audigier et al., 2017)

Mixed data

- general location model (Schafer, 1997)
- multivariate probit model (Boscardin et al., 2008; He, 2012)
- nonparametric bayesian joint models (Murray and Reiter, 2016)

FCS

- random forest
- logistic regression
- linear discriminant analysis
- cart
- Polytomous logistic regression

R packages

JM	FCS
norm	mice, micemd, miceMNAR
Amelia	mi
cat	Baboon
mix	VIM
missMDA	
jomo	
DPImputeCont (github)	
NPBayesImputeCat	
MixedDataImpute (not available since R 4.0.0)	

Outline

In summary

- MI allows inference missing values (no bias on SE)
- MI separates the missing data issue and the analysis step

  imputation model$\neq$analysis model!

  - advantages: relevant for problems where direct inference is difficult
  - drawbacks: uncongeniality
  - if both models are identical, no benefit compared to direct methods
- A large number of imputation models are available in R packages (Josse et al., 2021)

Complements

- Most of methods assume a MAR mechanism
- The robustness to its violation is performed a posteriori using a sensitivity analysis (Leacy et al., 2017)
- MI allows building confidence intervals and statistical tests for scalar quantities of interest or vectors
- Rubin's rules are essentially tailored for (generalized) linear models (less clear for variable selection, clustering, etc)

References I

D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.

James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. URL `http://www.jstatsoft.org/v45/i07/`.

M. Quartagno and J. Carpenter. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17): 2938–2954, 2016. ISSN 1097-0258.

J. Schafer. Imputation of missing covariates under a multivariate linear mixed model. Technical report, Dept. of Statistics, The Pennsylvania State University, 1997.

R. A. Hughes, I. R. White, S. Seaman, J. Carpenter, K. Tilling, and J. Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14 (1):28, 2014.

D. Vidotto, M. Kaptein, and J. Vermunt. Multiple imputation of missing categorical data using latent class models: State of art. *Psychological test and assessment modeling*, 57:542–576, 2015.

Vincent Audigier, François Husson, and Julie Josse. Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2):501–518, 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9635-4.

References II

W.J. Boscardin, X. Zhang, and T.R. Belin. Modeling a mixture of ordinal and continuous repeated measures. *Journal of Statistical Computation and Simulation*, 78(10): 873–886, 2008.

R. He. *Multiple Imputation of High-dimensional Mixed Incomplete Data*. PhD thesis, University of California, 2012.

Jared S. Murray and Jerome P. Reiter. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016. doi: 10.1080/01621459.2016.1174132. URL `https://doi.org/10.1080/01621459.2016.1174132`.

J. Josse, N. Tierney, and N. Vialaneix. CRAN Task View: Missing Data. `https://cran.r-project.org/web/views/MissingData.html#multiple`, 2021. [En ligne; consulté le 14 Juin 2021].

Finbarr P Leacy, Sian Floyd, Tom A Yates, and Ian R White. Analyses of Sensitivity to the Missing-at-Random Assumption Using Multiple Imputation With Delta Adjustment: Application to a Tuberculosis/HIV Prevalence Survey With Incomplete HIV-Status Data. *American Journal of Epidemiology*, 185(4):304–315, 2017. ISSN 0002-9262. doi: 10.1093/aje/kww107. https://core.ac.uk/download/pdf/111033547.pdf.

`factominer.free.fr/missMDA/appendix_These_Audigier.pdf`

`https://stefvanbuuren.name/fimd/`