Imputation multiple pour données mixtes par analyse factorielle

V. Audigier, F. Husson, M. Resche-Rigon, J. Josse

Journées de la statistique 2019

Frog-ICU

Les données[1]

- 2085 patients admis en réanimation
- 500 variables (caractéristiques du patient, profil à l'entrée en réanimation, en sortie, tests psycho, ...)

Objectifs

- expliquer la survie à 1 an à partir des données collectées pendant la réanimation
- expliquer le décès en réanimation
- évaluation thérapeutique
- mise en évidence de biomarqeurs

Problèmes

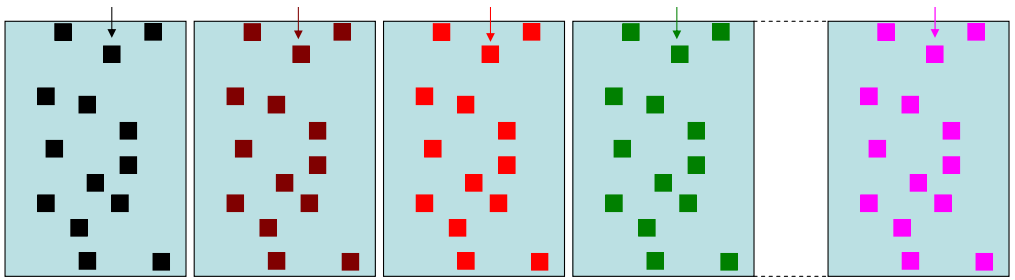
- données incomplètes
- données mixtes
- grande dimension

[1] Remerciements à Etienne Gayat

Imputation Multiple(Rubin, 1987)

1. Générer un jeu de M paramètres $(\theta_m)_{1\leq m\leq M}$ d'un modèle modèle d'imputation pour générer M jeux de données imputés plausibles

$$P\left(X^{miss}|X^{obs},\theta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(X^{miss}|X^{obs},\theta_M\right)$$



2. Ajuster le modèle d'analyse sur chaque jeu imputé : $\hat{\beta}_m, \widehat{\text{Var}}\left(\hat{\beta}_m\right)$

3. Agréger les résultats : $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$

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

$\Rightarrow$ **Fournit une estimation des paramètres du modèle d'analyse et de leur variabilité**

IM par modèle conditionnel ou joint

- Définir une distribution jointe $P\left(X^{obs}, X^{miss}; \theta\right)$
  Ex: $X = \left(X^{obs}, X^{miss}\right) \sim \mathcal{N}_p(\mu, \Sigma)$, $\theta = (\mu, \Sigma)$
  - générer θ **Ex: Bootstrap + EM**
  - déduire les distributions conditionnelles $P\left(X^{miss}|X^{obs}; \theta\right)$
  - imputer en tirant dans ces distributions

- Définir les distributions conditionnelles $P\left(X_j|X_{-j}; \theta_j\right)$
  Ex : $P\left(X_j|X_{-j}\right) = \mathcal{N}(X_{-j}\beta, \sigma^2)$
  - initialiser les données manquantes
  - pour chaque j
    - générer θ_j en utilisant les individus observés sur X_j
    - imputer X_j selon la distribution conditionnelle
  - répéter de 5 à 10 fois

Avantages et inconvénients

Modèles d'imputation joints

Avantages	garanties théoriques
Inconvénients	peu nombreux pour des données mixtes
	souvent sur-paramétrés avec données qualitatives

Modèles d'imputation conditionnels

Avantages	ajustent bien les données
Inconvénients	coûteux en temps de calcul
	peu de garanties théoriques
	impossible de vérifier la convergence si p grand

$\Rightarrow$ **Très peu de solutions pour imputer un nombre de variables important en présence de données mixtes**

IM par analyse factorielle

Les méthodes d'analyse factorielle

- offrent de grandes perspectives en termes de modèle d'imputation
- permettent de gérer le cas de la grande dimension
- ne sont pas sensibles aux problèmes de colinéarité

$\Rightarrow$ **Proposer une méthode d'imputation multiple reposant sur l'analyse factorielle des données mixtes**

AFDM (1)

L'AFDM peut être vue comme une SVD de $\mathbf{Z}$ avec des poids pour

- les variables continues et les catégories : $(\mathbf{D}_\Sigma)^{-1}$
- les individus : $\frac{1}{n}\mathbb{1}_n$

$$\longrightarrow SVD\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right)$$

$$\mathbf{Z} = \begin{array}{|ccc||ccc|ccc|}
\hline
11.04 & \dots & 2.07 & 1 & 0 & \dots & 1 & 0 & 0 \\
10.76 & \dots & 1.86 & 1 & 0 & \dots & 1 & 0 & 0 \\
11.02 & \dots & 2.04 & 1 & 0 & \dots & 1 & 0 & 0 \\
11.02 & \dots & 1.92 & 0 & 1 & \dots & 0 & 1 & 0 \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
11.06 & & 2.01 & 0 & 1 & & 0 & 0 & 1 \\
10.95 & & 1.67 & 0 & 1 & & 0 & 1 & 0 \\
\hline
\end{array}$$

$$D_\Sigma = \begin{array}{|cc|}
\hline
\sigma_{x_1} \;\; \ddots \;\; \sigma_{x_k} & 0 \\
\hline
0 & n_{k+1} \;\; \ddots \;\; n_K \\
\hline
\end{array}$$

AFDM (2)

$$SVD\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right) \longrightarrow \begin{array}{ll} \mathbf{Z}_{n\times K} & = \mathbf{U}_{n\times K}\Lambda^{1/2}_{K\times K}\mathbf{V}^\top_{K\times K} \\ \text{avec} & \mathbf{U}^\top\left(\frac{1}{n}\mathbb{1}_n\right)\mathbf{U} = \mathbb{1}_K \\ & \mathbf{V}^\top\mathbf{D}_\Sigma^{-1}\mathbf{V} = \mathbb{1}_K \end{array}$$

- composantes principales : $\hat{\mathbf{F}}_{n\times S} = \hat{\mathbf{U}}_{n\times S}\hat{\Lambda}^{1/2}_{S\times S}$
- vecteurs propres : $\hat{\mathbf{V}}^\top_{K\times S}$
- matrice ajustée : $\hat{\mathbf{Z}}_{n\times K} = \hat{\mathbf{U}}_{n\times S}\hat{\Lambda}^{1/2}_{S\times S}\hat{\mathbf{V}}^\top_{K\times S}$

$$\| \hat{\mathbf{Z}} - \mathbf{Z} \|^2_{\mathbf{D}_\Sigma^{-1}\otimes\frac{1}{n}\mathbb{1}} = tr\left(\left(\hat{\mathbf{Z}} - \mathbf{Z}\right)\mathbf{D}_\Sigma^{-1}\left(\hat{\mathbf{Z}} - \mathbf{Z}\right)^\top\frac{1}{n}\mathbb{1}_n\right)$$

minimisé sous la contrainte de rang S

AFDM avec données manquantes

$\Rightarrow$ AFDM : moindres carrés

$$\|\mathbf{Z}_{n\times K} - \mathbf{U}_{n\times S}\mathbf{\Lambda}^{\frac{1}{2}}_{S\times S}\mathbf{V}^{\top}_{K\times S}\|^2$$

$\Rightarrow$ AFDM avec données manquantes : moindres carrés pondérés

$$\|\mathbf{W_{n\times K}} * (\mathbf{Z}_{n\times K} - \mathbf{U}_{n\times S}\mathbf{\Lambda}^{\frac{1}{2}}_{S\times S}\mathbf{V}^{\top}_{K\times S})\|^2$$

avec $w_{ij} = 0$ si x_{ij} est manquant, $w_{ij} = 1$ sinon

Plusieurs algorithmes ont été développés dans le cadre de l'ACP comme NIPALS (Christoffersson, 1970) ou l'ACP itérative (Kiers, 1997)

AFDM avec données manquantes

Algorithme d'AFDM itérative :

1. initialisation: imputation par la moyenne/proportion
2. répéter jusqu'à convergence
   (a) estimation des paramètres de l'AFDM
   $\rightarrow$ SVD de $\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right)$
   (b) imputation des données manquantes avec
   $\hat{\mathbf{Z}}_{n\times K} = \hat{\mathbf{U}}_{n\times s}\hat{\Lambda}^{1/2}_{s\times s}\hat{\mathbf{V}}^{\top}_{K\times s}$
   (c) $\mathbf{D}_\Sigma$ est mis à jour

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

AFDM avec données manquantes

Algorithme d'AFDM itérative :

1. initialisation: imputation par la moyenne/proportion
2. répéter jusqu'à convergence
   (a) estimation des paramètres de l'AFDM
   $\rightarrow$ SVD de $\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right)$
   (b) imputation des données manquantes avec
   $\hat{\mathbf{Z}}_{n\times K} = \hat{\mathbf{U}}_{n\times s}\hat{\Lambda}^{1/2}_{s\times s}\hat{\mathbf{V}}^{\top}_{K\times s}$
   (c) $\mathbf{D}_\Sigma$ est mis à jour

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

$\rightarrow$

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

AFDM avec données manquantes

Algorithme d'AFDM itérative :

1. initialisation: imputation par la moyenne/proportion
2. répéter jusqu'à convergence
   (a) estimation des paramètres de l'AFDM
   $\rightarrow$ SVD de $\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right)$
   (b) imputation des données manquantes avec
   $\hat{\mathbf{Z}}_{n\times K} = \hat{\mathbf{U}}_{n\times s}\hat{\Lambda}^{1/2}_{s\times s}\hat{\mathbf{V}}^{\top}_{K\times s}$
   (c) $\mathbf{D}_\Sigma$ est mis à jour

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

$\rightarrow$

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

AFDM avec données manquantes

Algorithme d'AFDM itérative :

1. initialisation: imputation par la moyenne/proportion
2. répéter jusqu'à convergence
   (a) estimation des paramètres de l'AFDM
   $\rightarrow$ SVD de $\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, \frac{1}{n}\mathbb{1}_n\right)$
   (b) imputation des données manquantes avec
   $\hat{\mathbf{Z}}_{n\times K} = \hat{\mathbf{U}}_{n\times S} f\left(\hat{\Lambda}^{1/2}_{S\times S}\right)\hat{\mathbf{V}}^{\top}_{K\times S}$ $\qquad f\left(\lambda_s^{1/2}\right) = \lambda_s^{1/2} - \frac{\hat{\sigma}^2}{\hat{\lambda}_s^{1/2}}$
   (c) $\mathbf{D}_\Sigma$ est mis à jour

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

$\rightarrow$

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

IM par AFDM

1. Variabilité sur les paramètres de l'AFDM $(\hat{\mathbf{U}}_{n\times S}, \hat{\Lambda}^{1/2}_{S\times S}, \hat{\mathbf{V}}^{\top}_{J\times S})$ refléter via un bootstrap non-paramétrique :

   $\rightarrow$ définit M pondérations $(R_m)_{1\leq m\leq M}$ pour les individus

2. Effectuer l'AFDM itérative via la SVD de $\left(\mathbf{Z}, (\mathbf{D}_\Sigma)^{-1}, R_m\right)$

$\hat{\mathbf{Z}}_1$ $\hat{\mathbf{Z}}_2$ $\hat{\mathbf{Z}}_M$

11.04	...	2.07	1	0		9.23	...	2.07	1	0		10.34	...	2.07	1	0
10.76	...	1.86	1	0		10.76	...	1.86	1	0		10.76	...	1.86	1	0
11.02	...	2.04	1	0		11.02	...	1.14	1	0		11.02	...	3.42	1	0
11.02	...	1.92	0	1		11.02	...	1.92	0	1	...	11.02	...	1.92	0	1
11.06		2.01	0.25	0.75		11.06		2.01	0.59	0.41		11.06		2.01	0.21	0.79
10.95		1.67	0	1		12.34		1.67	0	1		11.05		1.67	0	1

3. Ajout d'un bruit Gaussien sur $\left(\hat{\mathbf{Z}}_m\right)_{1\leq m\leq M}$ de variance $\hat{\sigma}^2\mathbf{D}_\Sigma$

General location model (Schafer, 1997)

- Généralisation de l'analyse discriminante à plusieurs variables qualitatives

$$\begin{aligned} X_{quanti}|X_{quali} = c &\sim \mathcal{N}(\mu_c, \Sigma_c) \\ X_{quali} &\sim \mathcal{M}(1, \pi) \end{aligned}$$

1. Variabilité sur $\theta = (\pi, (\mu_c)_{1\leq c\leq C}, (\Sigma_c)_{1\leq c\leq C})$: Bayésien
2. Imputation en utilisant les M paramètres

- `R` package `mix` (J.L. Schafer)

Propriétés :

- Méthode de référence
- Grand nombre de paramètres $\rightarrow$ échoue en grande dimension

Nonparametric Bayesian Joint Model (Murray and Reiter, 2015)

- X_{quali} suit un modèle à classes latentes
- $X_{quali}|X_{quanti}$ également, avec probabilité d'appartenance fonction de X_{quanti}
- $X_{quanti}|X_{quali}$ suit un modèle de mélange Gaussien

1. Variabilité sur θ : Bayésien
2. Imputation selon le jeu de M paramètres

- `R` package `MixedDataImpute` (Jared S. Murray)

Propriétés :

- Indépendance locale
- Modèle assez souple
- Méthode MCMC assez coûteuse

Imputation séquentielle (van Buuren, 2006)

- **Modèle linéaire généralisé** : un modèle de régression linéaire ou logistique/variable (sans interactions)

  Propriétés :
  - sensible à la grande dimension
  - sensible aux fortes liaisons

- **Forêts aléatoires** (Doove *et al.*, 2014)

  Propriétés :
  - modélisation non-paramétrique
  - limitée sur petits échantillons

Simulations

- Modèle d'analyse :

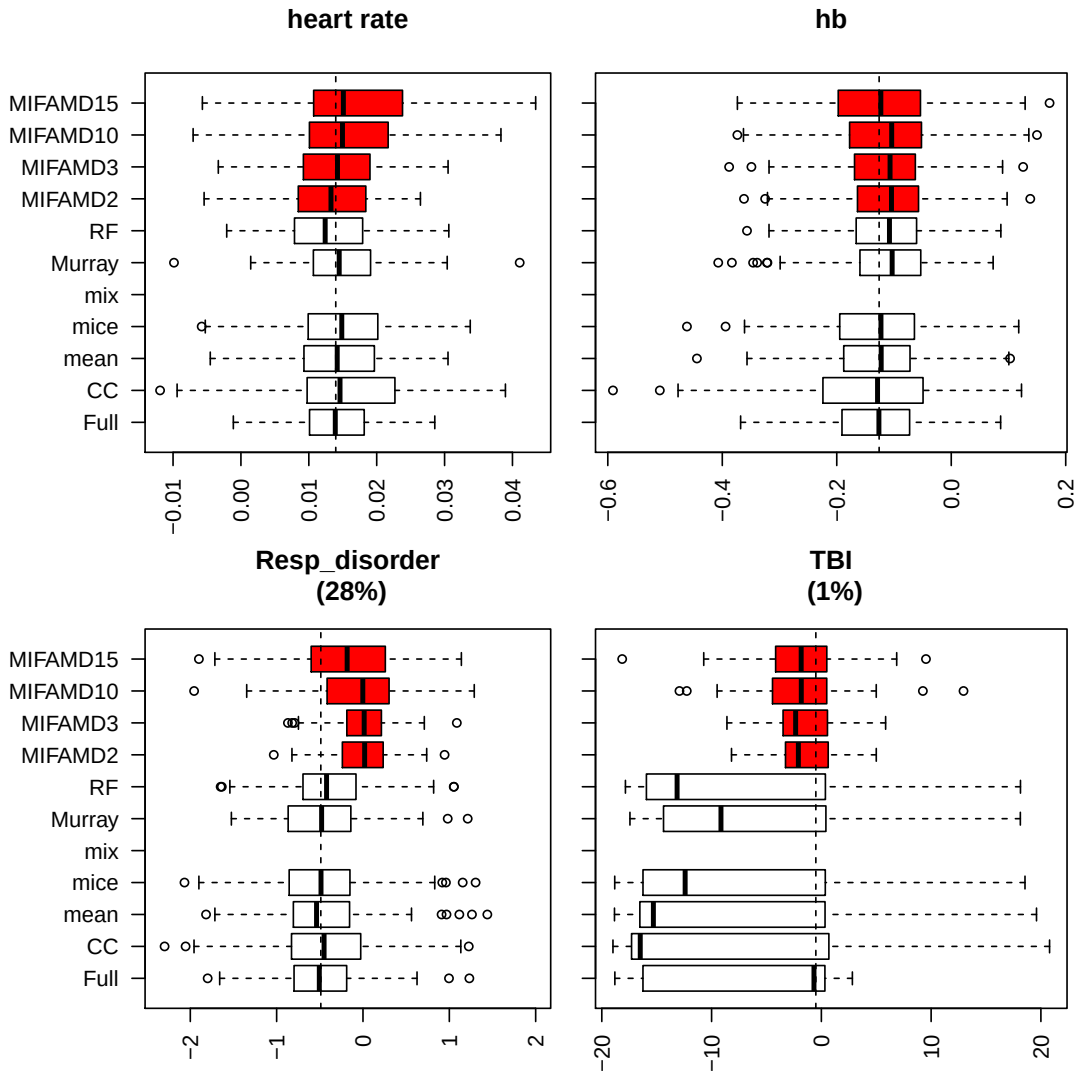
  mort en réa ~ diagnostique + fréq. card. + hémoglob.

  β = coef. régression logistique

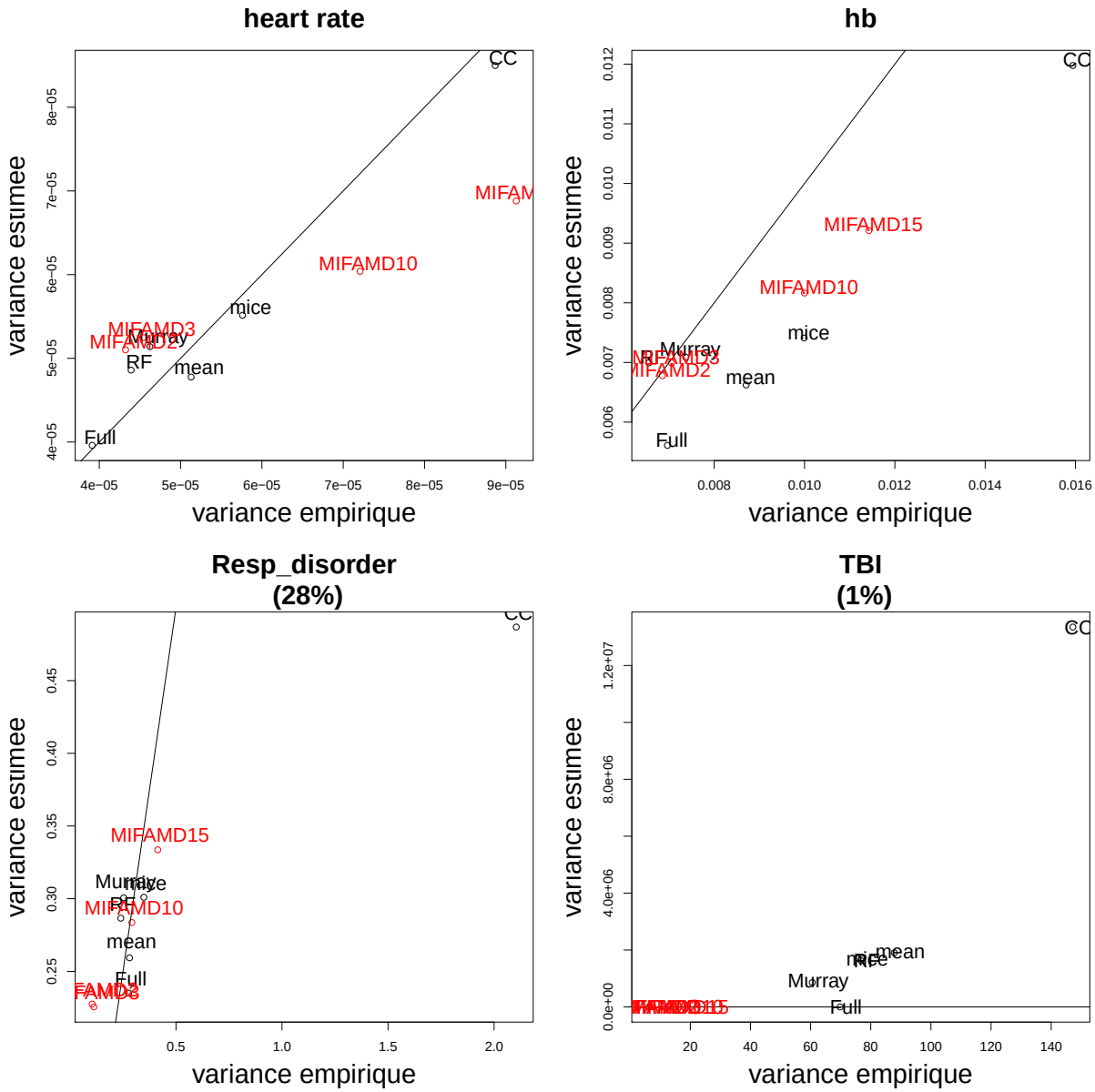
- 200 simulations
  - les individus complets de la base restreinte à 19 variables constituent la population
  - tirage d'un échantillon
  - ajout de 20% de données manquantes
  - IM via $M = 5$ pour différents modèles d'imputation

- Critères
  - Biais sur β
  - Biais sur $Var(\beta)$

Résultats (1)


heart rate
hb
Resp_disorder
(28%)
TBI
(1%)
MIFAMD15
MIFAMD10
MIFAMD3
MIFAMD2
RF
Murray
mix
mice
mean
CC
Full
−0.01
0.00
0.01
0.02
0.03
0.04
−0.6
−0.4
−0.2
0.0
0.2
−2
−1
0
1
2
−20
−10
0
10
20

Résultats (2)


heart rate
variance estimee
variance empirique
CC
MIFAM
MIFAMD10
mice
MIFAMD3
Murray
RF
mean
Full
4e−05
5e−05
6e−05
7e−05
8e−05
9e−05
hb
variance estimee
variance empirique
CC
MIFAMD15
MIFAMD10
mice
Murray
mean
Full
0.006
0.007
0.008
0.009
0.010
0.011
0.012
0.008
0.010
0.012
0.014
0.016
Resp_disorder
(28%)
variance estimee
variance empirique
MIFAMD15
Murray
MIFAMD10
mean
Full
0.25
0.30
0.35
0.40
0.45
0.5
1.0
1.5
2.0
TBI
(1%)
variance estimee
variance empirique
mean
Murray
Full
0.0e+00
4.0e+06
8.0e+06
1.2e+07
20
40
60
80
100
120
140

Résultats (3)

	Minutes
Full	0.00
CC	0.00
mean	0.00
mice	0.54
mix	
Murray	4.19
RF	0.88
MIFAMD2	0.03
MIFAMD3	0.04
MIFAMD10	0.09
MIFAMD15	0.14

Conclusion

Imputation multiple par AFDM

Points forts : méthode de réduction de la dimension

- prise en compte des ressemblances entre individus / liaisons entre variables
- prise en compte des modalités rares
- peu de paramètres

D'un point de vue pratique

- peut être appliquée sur des jeux de dimensions variées
- fournit des inférences valides et rapidement
- un paramètre à régler : le nombre de dimensions

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.

Jared S Murray and Jerome P Reiter. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence, 2015. URL `http://arxiv.org/abs/1410.0438`.

V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1): 5–26, 2016. ISSN 1862-5355. doi: 10.1007/s11634-014-0195-1. URL `http://dx.doi.org/10.1007/s11634-014-0195-1`.

S. van Buuren. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 2012.

L.L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92 – 104, 2014. ISSN 0167-9473.