Gestion des données manquantes en/par ACM et analyse de données mixtes

François Husson

UP de mathématiques appliquées - l'institut Agro

Journées d'études en statistique – SFdS 2021

Plan

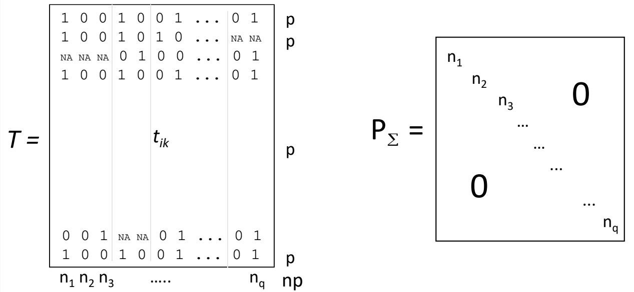
Les méthodes d'analyse factorielle

- Analyse exploratoire de tableaux de données
- Dépend de la structure et de la natre des variables :
  - ACP : variables quantitatives
  - ACM : variables qualitatives
  - AFDM : variables quantitatives et qualitatives
  - AFM : structure avec des groupes de variables
  - ...

Toutes les méthodes d'analyse factorielle peuvent être vues comme une ACP sur une matrice particulière avec des poids spécifiques pour les lignes et les colonnes

« Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize » (Benzécri)

- Analyse exploratoire d'un tableau de variables qualitatives
- Analyse de questionnaires



L'ACM comme une ACP pondérée

ACM vue comme l'ACP du triplet

$$\left(n\mathbf{T}\mathbf{P}_\Sigma^{-1}, \frac{1}{np}\mathbf{P}_\Sigma, \frac{1}{n}I_n\right)$$

$$\left(n\mathbf{TP}_{\Sigma}^{-1}, \frac{1}{np}\mathbf{P}_{\Sigma}, \frac{1}{n}I_n\right)$$

ACP d'un triplet (A, M, P)

L'ACP d'un triplet ($\mathbf{A}$, $\mathbf{M}$, $\mathbf{P}$) est la SVD suivante :

$$A = UDV'$$

avec $\mathbf{U}$ les vecteurs propres de $\mathbf{AMA'P}$ et tels que $\mathbf{U'PU} = Id$
et $\mathbf{V}$ les vecteurs propres de $\mathbf{A'PAM}$ et tels que $\mathbf{V'MV} = Id$

$$\left(n\mathbf{T}\mathbf{P}_{\Sigma}^{-1}, \frac{1}{np}\mathbf{P}_{\Sigma}, \frac{1}{n}I_n\right)$$

ACP d'un triplet (A, M, P)

L'ACP d'un triplet $(\mathbf{A}, \mathbf{M}, \mathbf{P})$ est la SVD suivante :

$$A = UDV'$$

avec $\mathbf{U}$ les vecteurs propres de $\mathbf{AMA'P}$ et tels que $\mathbf{U'PU} = Id$
et $\mathbf{V}$ les vecteurs propres de $\mathbf{A'PAM}$ et tels que $\mathbf{V'MV} = Id$

$\mathbf{U}; \mathbf{D}; \mathbf{V}$ minimisent le critère d'erreur de reconstitution :

$$\mathcal{C} = \|\mathbf{A} - \mathbf{UDV'}\|^2_{\mathbf{M},\mathbf{P}}$$

Plan

Exemple : traitement d'un questionnaire

Les données

1232 répondants, 14 questions, 35 modalités, 9% de NA pour 42% des répondants

Les données

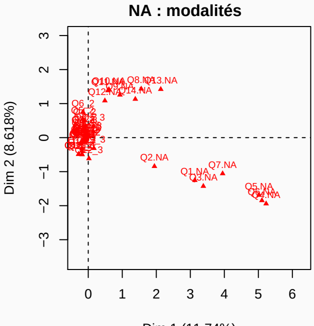
1232 répondants, 14 questions, 35 modalités, 9% de NA pour 42% des répondants

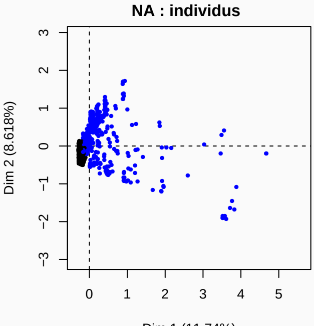
Création de nouvelles modalités

Création d'une modalité NA pour chaque variable ayant au moins une valeur manquante

	V1	V2	V3
ind 1	a	**NA**	g
ind 2	**NA**	f	g
ind 3	a	e	h
ind 4	a	e	h
ind 5	b	f	h
ind 6	c	f	h
ind 7	c	f	**h**

	V1_a	V1_b	V1_c	**V1_NA**	V2_e	V2_f	**V2_NA**	V3_g	V3_h
ind 1	1	0	0	0	**0**	**0**	**1**	1	0
ind 2	**0**	**0**	**0**	**1**	0	1	0	1	0
ind 3	1	0	0	0	1	0	0	0	1
ind 4	1	0	0	0	1	0	0	0	1
ind 5	0	1	0	0	0	1	0	0	1
ind 6	0	0	1	0	0	1	0	0	1
ind 7	0	0	1	0	0	1	0	0	1





ACM spécifique – missing passive modified margin

Utilisation de l'ACM spécifique

L'ACM spécifique permet de construire les axes en mettant en supplémentaire des modalités. L'idée est ici de mettre les modalités NA en supplémentaire

Remarque

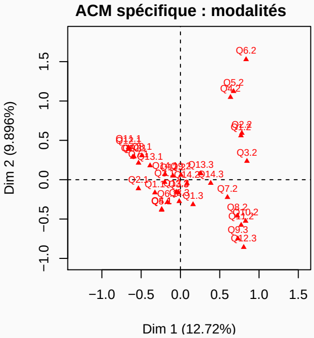
Cette méthode donne des résultats équivalents à la méthode *missing passive modified margin*

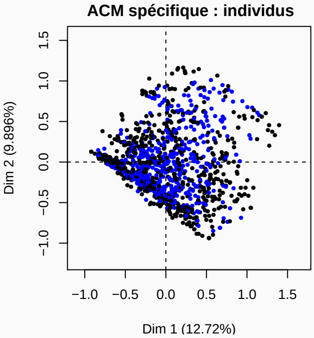
Utilisation de l'ACM spécifique

L'ACM spécifique permet de construire les axes en mettant en supplémentaire des modalités. L'idée est ici de mettre les modalités NA en supplémentaire

Remarque

Cette méthode donne des résultats équivalents à la méthode *missing passive modified margin*





Plan

ACM itérative régularisée (Josse *et al.*, 2012)

1. Initialisation : imputation de la matrice indicatrice (proportion)
2. Itération jusqu'à convergence
   (a) Estimation de $\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell$: ACM sur le tableau complété, i.e. l'ACP du triplet

$$\left(n\mathbf{T}^{\ell-1}(\mathbf{P}_\Sigma^{\ell-1})^{-1}, \frac{1}{np}\mathbf{P}_\Sigma^{\ell-1}, \frac{1}{n}\mathbb{I}_n\right)$$

   (b) Utiliser la formule de reconstitution (prendre les valeurs singulières régularisées) :

$$(\hat{a}_{ik}^\ell - 1)\sqrt{\frac{n_k^{\ell-1}}{np}} = \left(\sum_{s=2}^{S} \hat{u}_{is}\left(\hat{d}_{ss} - \frac{\hat{\sigma}^2}{\hat{d}ss}\right)\hat{v}_{ks}^\ell\right)$$

   Calculer les valeurs reconstituées en utilisant les marges de l'étape $\ell - 1$: $\hat{\mathbf{T}}^\ell = \frac{1}{n}\hat{\mathbf{A}}^\ell \mathbf{P}_\Sigma^{\ell-1}$ et le nouveau tableau disjonctif complété est $\mathbf{T}^\ell = \mathbf{R} * \mathbf{T} + (1 - \mathbf{R}) * \hat{\mathbf{T}}^\ell$

   (c) Mise à jour des marges : les marges colonnes n_k^ℓ du nouveau tableau complété $\mathbf{T}^\ell$ sont calculées et enregistrées dans $\mathbf{P}_\Sigma^\ell$;
3. les étapes (2.a), (2.b) et (2.c) sont répétées jusqu'à convergence.

ACM itérative régularisée (Josse *et al.*, 2012)

	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

ACM itérative régularisée (Josse *et al.*, 2012)

	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	...

ACM itérative régularisée (Josse *et al.*, 2012)

	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	…

	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	…

Les valeurs imputées peuvent être vues comme des degrés d'appartenance

ACM itérative régularisée (Josse *et al.*, 2012)

	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	…

	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	…

	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	**h**		v
…	…	…	…		…
ind 1232	c	f	h		v

Les valeurs imputées peuvent être vues comme des degrés d'appartenance

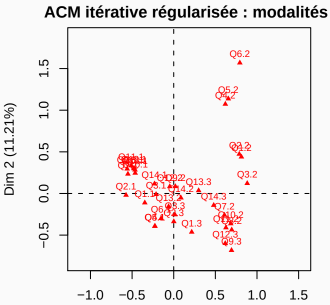
Mise en œuvre

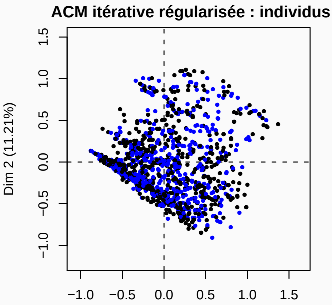
Imputation du tableau disjonctif

```
> library(missMDA)
> data(vnf)
> ncp <- estim_ncpMCA(vnf)
> res.impute <- imputeMCA(vnf, ncp=4)
```

ACM sur le tableau complété (utilisation de l'argument `tab.disj`)

```
> res.mca <- MCA(vnf, tab.disj = res.impute$tab.disj)
```





Plan

Analyse Factorielle de Données Mixtes (cas complet)

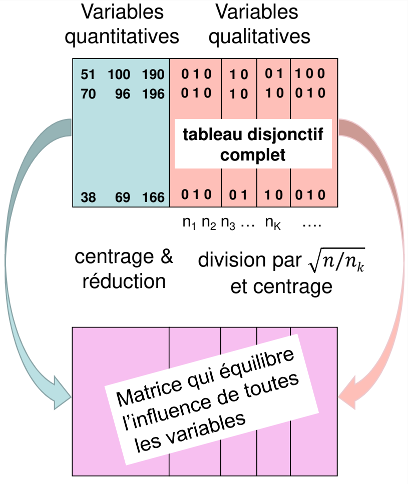
AFDM (Escofier, 1979), PCAMIX (Kiers, 1991)

- ACP sur une matrice pondérée
- La distance entre individus s'écrit :

$$d^2(i,l) = \sum_{j=1}^{p_1}(t_{ik} - t_{lk})^2 + \sum_{j=1}^{p_2}\sum_{k=1}^{K_j}\frac{1}{n_{k_j}}(t_{ij} - t_{lj})^2$$

- Les composantes principales $\mathbf{F}_s$ maximisent :

$$\sum_{j=1}^{p_1} r^2(\mathbf{F}_s, v_j) + \sum_{j=1}^{p_2}\eta^2(\mathbf{F}_s, v_j)$$



Algorithme d'AFDM itératif

1. Initialisation : imputation par la moyenne (quanti) et la proportion (quali)
2. Itérer jusqu'à convergence
   (a) estimation : AFDM sur le jeu complété $\Rightarrow \mathbf{U}, \mathbf{D}, \mathbf{V}$
   (b) imputation des valeurs manquantes avec le modèle de reconstitution
   (c) moyennes, écarts-types et marges sont mis à jour

age	weight	size	alcohol	sex	snore	tobacco
NA	100	190	NA	M	yes	no
70	96	186	1-2 gl/d	M	NA	<=1
NA	104	194	No	W	no	NA
62	68	165	1-2 gl/d	M	no	<=1

NA	100	190	NA	NA	NA	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	NA	NA	0	1	0
NA	104	194	1	0	0	0	1	1	0	NA	NA	NA
62	68	165	0	1	0	1	0	1	0	0	1	0

imputeAFDM

51	100	190	0.2	0.7	0.1	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	0.8	0.2	0	1	0
48	104	194	1	0	0	0	1	1	0	0.1	0.8	0.1
62	68	165	0	1	0	1	0	1	0	0	1	0

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1-2 gl/d	M	yes	no
70	96	186	1-2 gl/d	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1-2 gl/d	M	no	<=1

Les valeurs imputées peuvent être vues comme des degrés d'appartenance

Simulations

- Dispositif de simulations
  - 2 variables indépendantes provenant d'une distribution normale
  - 1 variable répétée 4 fois, l'autre 8 $\Rightarrow$ 2 dimensions
  - Bruit ajouté
  - La moitié des variables sur chaque dimension sont découpées en 3 classes
  - 10%, 20% or 30% de données manquantes au hasard

$\Rightarrow$ Données sont construites pour être en 4 dimensions

- Critère
  - pour données quantitatives :

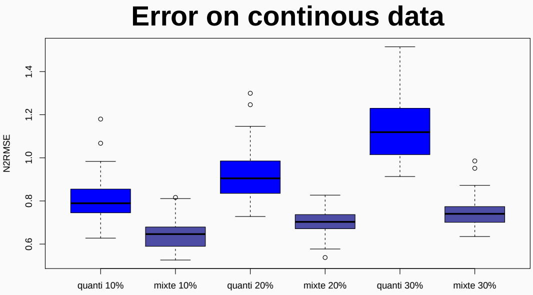
$$N2RMSE = \sqrt{\sum_{i \in \text{manquant}} \frac{moyenne\left(\left(X_i^{vrai} - X_i^{imp}\right)^2\right)}{var\left(X_i^{true}\right)}}$$

  - pour données qualitatives : proportion de modalités mal prédites

Simulations

Imputation avec var. quanti uniquement

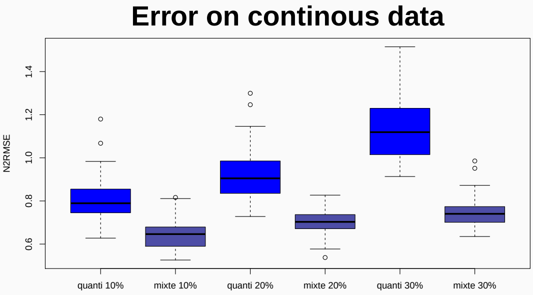
Imputation avec variables quanti et quali



Simulations

Imputation avec var. quanti uniquement

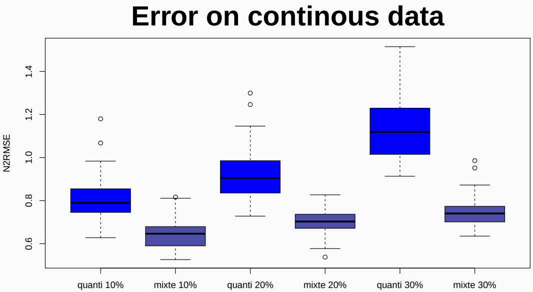
Imputation avec variables quanti et quali

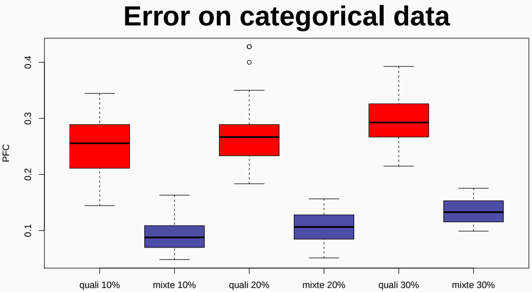


Variables quali améliorent
l'imputation sur variables quanti ...

Simulations

Imputation avec var. quanti uniquement Imputation avec var. quali uniquement
Imputation avec variables quanti et quali

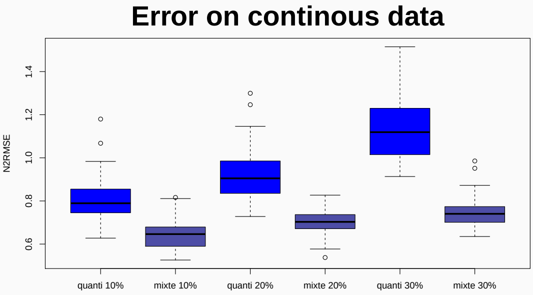


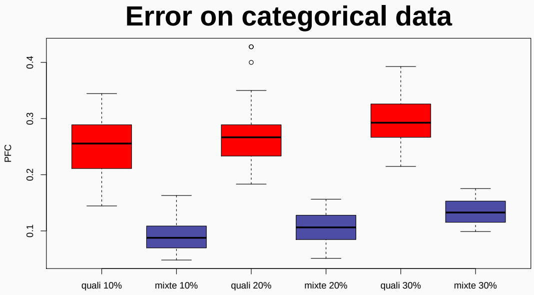


Variables quali améliorent
l'imputation sur variables quanti ...

Simulations

Imputation avec var. quanti uniquement Imputation avec var. quali uniquement
Imputation avec variables quanti et quali

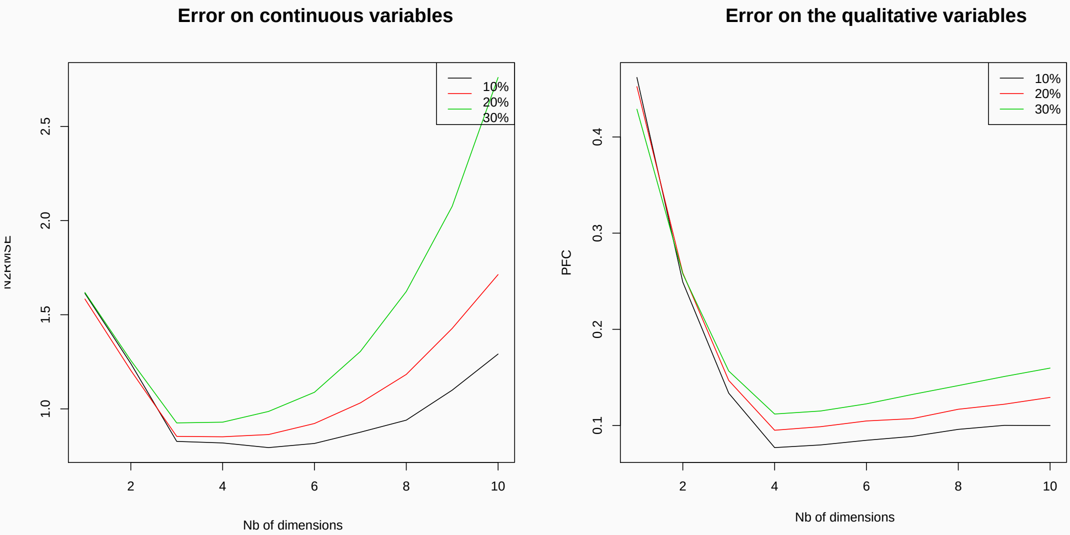




Variables quali améliorent l'imputation sur variables quanti ...

... et variables quanti améliorent l'imputation des variables quali

Simulations

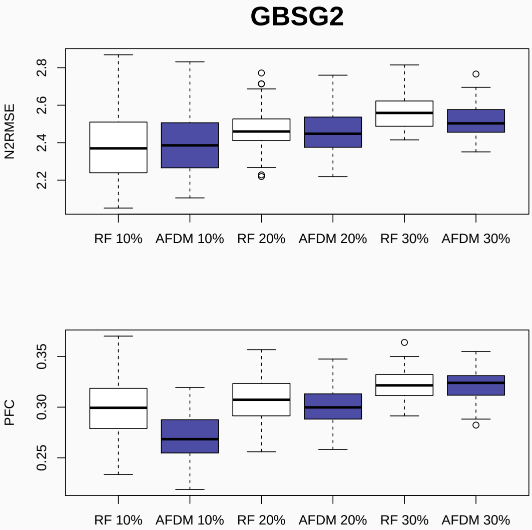


⇒ L'erreur sur le choix du nombre de dimensions a un impact faible sur l'erreur d'imputation

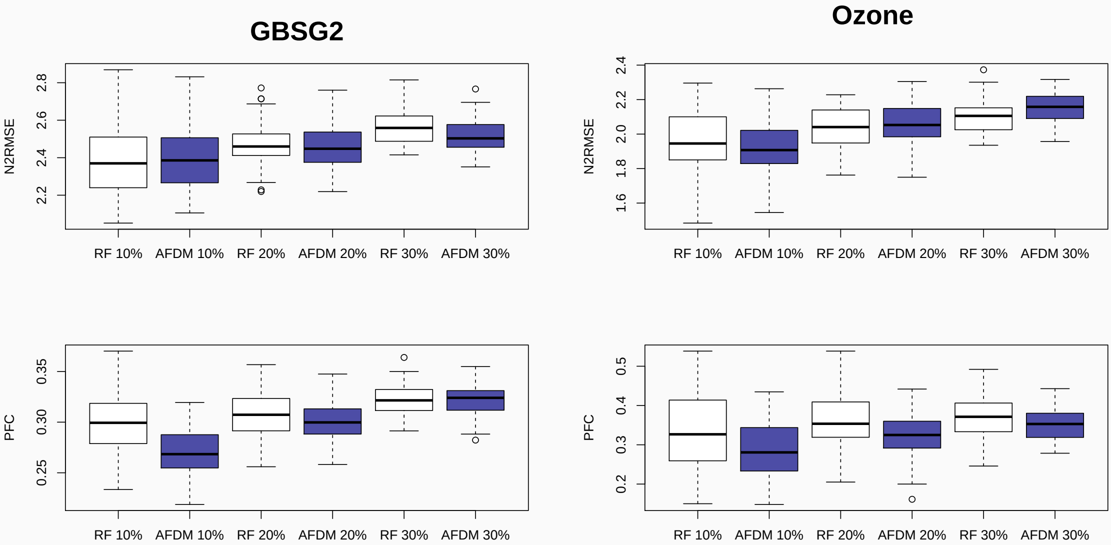
... si l'estimation n'est pas trop mauvaise

Comparaison avec forêts aléatoires

Imputations obtenues par forêts aléatoires & ACP itérative



Imputations obtenues par forêts aléatoires & ACP itérative



Imputation de données mixtes en pratique

```
> library(missMDA)
> nb <- estim_ncpFAMD(mydata)   ## tps de calcul long
> res.imp <- imputeFAMD(mydata, ncp = nb$ncp)
> res.famd <- FAMD(mydata, ,tab.disj = res.imp$tab.disj)

> library(missForest)
> missForest(mydata)

> library(mice)
> mice(mydata)
> mice(mydata, defaultMethod = "rf") ## mice avec forêts aléatoires
```

Même principe avec mise à jour des premières valeurs propres de chaque groupe en plus

Cas de groupes quantitatifs uniquement : le tableau est complété et l'AFM est lancée sur le tableau complété :

```
> data(orange)
> res.comp <- imputeMFA(orange, group=c(5,3), type=rep("s",2), ncp=2)
> res.mfa <- MFA(res.comp$completeObs, group=c(5,3), type=rep("s",2))
```

Cas où au moins un groupe qualitatif : le "tableau disjonctif" complété est fournit à l'AFM avec l'argument `tab.comp` :

```
> data(vnf)
> res.comp <- imputeMFA(vnf,group=c(6,5,3),type=c("n","n","n"),ncp=2)
> res.mfa <- MFA(vnf,group=c(6,5,3),type=c("n","n","n"), tab.comp=res.comp)
```

Bilan sur l'imputation simple

⇒ Données manquantes en analyse factorielle

- tableau simple : ACP, ACM, analyse fact. de données mixtes
- tableaux multiples (AFM)

⇒ Pré-traitement avant classification (avec données manquantes)

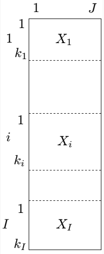
⇒ package R missMDA (complémentaire de FactoMineR)

⇒ Imputation des données quantitatives, qualitatives, mixtes

- basée sur la reconstitution de l'ACP (axes et composantes)
- prise en compte des liaisons entre var. quantitatives et qualitatives
- bonne alternative aux méthodes d'imputation (forêts aléatoires, etc.) si liaisons linéaires, pour les variables qualitatives (notamment les modalités rares)

Plan

Analyse en composantes multi-niveaux



Ex : patients hiérarchisés dans hôpitaux $X \in \mathbb{R}^{K \times J}$

- similarités entre hôpitaux ? niveau 1
- similarités entre patients dans un même hôpital ? niveau 2
- relations entre variables à chaque niveau

$$\begin{aligned} x_{ijk_i} = x_{.j.} + (x_{ij.} - x_{.j.}) \quad &+ \quad (x_{ijk_i} - x_{ij.}) \\ \text{Between} \quad &+ \quad \text{Within} \end{aligned}$$

Analysis de variance : décomposer la somme des carrés pour chaque variable j

$$\sum_{i=1}^{I} \sum_{k=1}^{k_i} (x_{ijk_i})^2 = \sum_{i=1}^{I} k_i (x_{.j.})^2 + \sum_{i=1}^{I} k_i (x_{ij.} - x_{.j.})^2 + \sum_{i=1}^{I} \sum_{k=1}^{k_i} (x_{ijk_i} - x_{ij.})^2$$

ACP ou ACM multi-niveaux : MLPCA

⇒ Modèle pour la partie between et within $i = 1, ..., I$ groupes, J var

$$X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} U_i^b D^b V^{b'} + F_i^w D^w V^{w'} + E_i$$

- F_i^b $(Q_b \times 1)$ between component scores of group i
- V^b $(J \times Q_b)$ between loading matrix
- F_i^w $(k_i \times Q_w)$ within component scores of group i
- V_w $(J \times Q_w)$ within loading matrix. Constant across groups

Solution obtenue par moindre carrés (Timmerman, 2006)

Possibilité de faire des calculs distribués

Limites de l'imputation

Remarque

*"**The idea of imputation is both seductive and dangerous**. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."* (Dempster & Rubin, 1983)

Limites de l'imputation

Remarque

*"**The idea of imputation is both seductive and dangerous**. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."* (Dempster & Rubin, 1983)

Imputation simple *versus* imputation multiple

On ne peut accorder la même confiance à une valeur imputée et une valeur observée

L'imputation simple retourne 1 seule valeur pour chaque valeur manquante et 1 seule valeur ne permet pas de connaître l'incertitude sur la prédiction de cette valeur

$\Longrightarrow$ Imputation multiple