Gestion des données manquantes en/par analyse factorielle

F. Husson

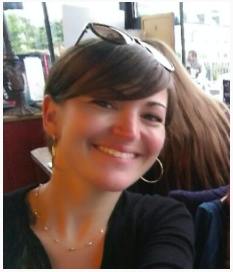
husson@agrocampus-ouest.fr

UP de mathématiques appliquées - l'institut Agro

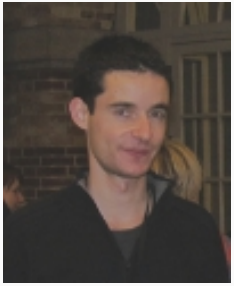
Plan

1. Introduction
2. Imputation simple pour variables quantitatives
3. Imputation simple pour variables qualitatives
4. Imputation simple pour données mixtes
5. Imputation multiple

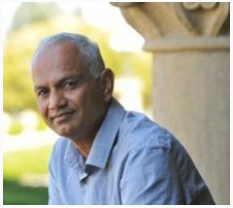
Collaborateurs



Julie Josse
Directrice de recherche
à l'INRIA



Vincent Audigier
Maître de conférences
au CNAM



Balasubramanian Narasimhan
Professeur
Univ. Stanford

Les données manquantes

Gertrude Mary Cox

"*The best thing to do about missing values is not to have any*"

Les données manquantes sont très présentes en pratique : non-réponse à un questionnaire, données perdues, appareils en panne, plantes détruitent (maladie, ravageurs, etc.) ...

Les données manquantes

Gertrude Mary Cox

"*The best thing to do about missing values is not to have any*"

Les données manquantes sont très présentes en pratique : non-réponse à un questionnaire, données perdues, appareils en panne, plantes détruitent (maladie, ravageurs, etc.) ...

Est-ce un problème en big data ?

"One of the ironies of Big Data is that missing data play an ever more significant role" (R. Sameworth, 2019)

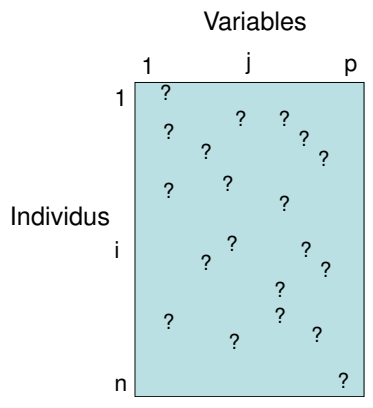
Une matrice $n \times p$, avec chaque cellule ayant une proba 0.01 d'être manquante

$p = 5 \Rightarrow \approx 95\%$ de lignes conservées

$p = 300 \Rightarrow \approx 5\%$ de lignes conservées

Objectifs

- Etude et mise en œuvre des méthodes factorielles en présence de données manquantes : ACP (variables quantitatives), ACM (variables qualitatives), AFDM (données mixtes), AFM (tableaux multiples)
- Imputation de données



Exemple sur des données ozone

Code disponible : http://factominer.free.fr/missMDA/ozone.R

```
> don <- read.table("http://factominer.free.fr/missMDA/ozoneNA.csv",
      header=TRUE,sep=",",row.names=1)
```

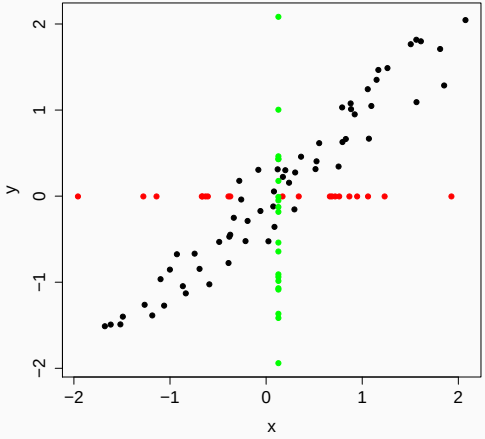
	O3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	O3v
0601	82	15.6	18.5	NA	4	4	8	NA	-1.7101	-0.6946	84
0602	82	NA	NA	NA	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
...	...	...	...	...	...	...	...	...	...	...	...
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

De (mauvaises) solutions faciles à mettre en œuvre

- Suppression des données manquantes : rarement intéressant ... mais souvent utilisée (fonction `lm` de R)

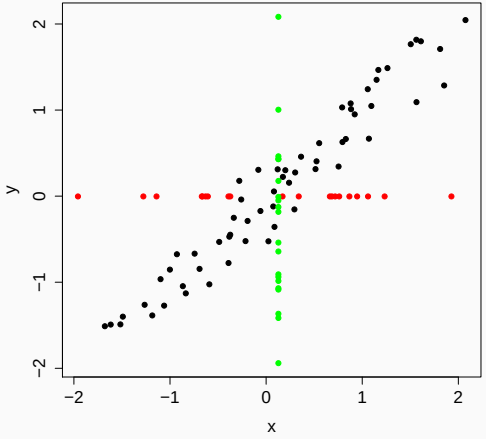
De (mauvaises) solutions faciles à mettre en œuvre

- Suppression des données manquantes : rarement intéressant ... mais souvent utilisée (fonction `lm` de R)
- Imputation par la moyenne (option par défaut dans de nombreux logiciels)



De (mauvaises) solutions faciles à mettre en œuvre

- Suppression des données manquantes : rarement intéressant ... mais souvent utilisée (fonction `lm` de R)
- Imputation par la moyenne (option par défaut dans de nombreux logiciels)



Distorsion très importante des liaisons entre variables

Etude du dispositif de données manquantes

Traitement des données manquantes dépend du :

- dispositif de données manquantes : structuré/non structuré
- mécanisme conduisant à l'apparition de données manquantes (Rubin, 1976)

Etude du dispositif de données manquantes

Traitement des données manquantes dépend du :

- dispositif de données manquantes : structuré/non structuré
- mécanisme conduisant à l'apparition de données manquantes (Rubin, 1976)
  - MCAR : probabilité ne dépend pas de cette valeur ni des autres
  - MAR : probabilité peut dépendre des valeurs d'autres variables
  - MNAR : probabilité dépend de la valeur elle-même

  (Ex : Revenu - âge)

Etude du dispositif de données manquantes

Traitement des données manquantes dépend du :

- dispositif de données manquantes : structuré/non structuré
- mécanisme conduisant à l'apparition de données manquantes (Rubin, 1976)
  - MCAR : probabilité ne dépend pas de cette valeur ni des autres
  - MAR : probabilité peut dépendre des valeurs d'autres variables
  - MNAR : probabilité dépend de la valeur elle-même

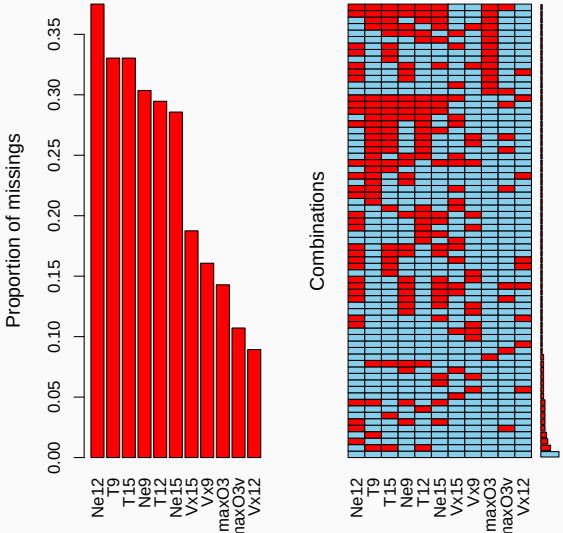
  (Ex : Revenu - âge)

$\Rightarrow$ Visualisation des données manquantes

Décompte des valeurs manquantes

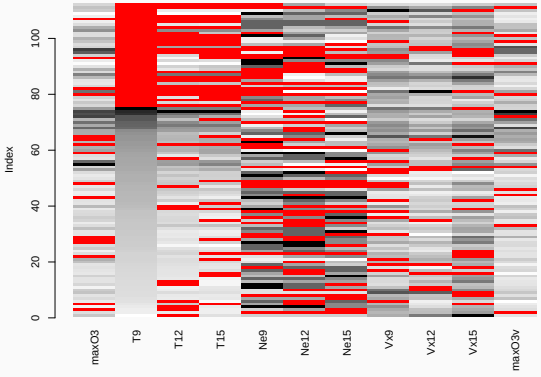
```
> don <- read.table("http://factominer.free.fr/missMDA/ozoneNA.csv",
       header=TRUE,sep=",",row.names=1)
> library(VIM)
> res <- summary(aggr(don,prop=TRUE,combined=TRUE))$combinations
> res[rev(order(res[,2])),]

Variables sorted by
number of missings:                    Combinations Count    Percent
Variable      Count          0:0:0:0:0:0:0:0:0:0:0    13 11.6071429
    Ne12 0.37500000          0:1:1:1:0:0:0:0:0:0:0     7  6.2500000
      T9 0.33035714          0:0:0:0:0:1:0:0:0:0:0     5  4.4642857
     T15 0.33035714          0:1:0:0:0:0:0:0:0:0:0     4  3.5714286
     Ne9 0.30357143          0:1:0:0:1:1:1:0:0:0:0     3  2.6785714
     T12 0.29464286          0:0:1:0:0:0:0:0:0:0:0     3  2.6785714
    Ne15 0.28571429          0:0:0:1:0:0:0:0:0:0:0     3  2.6785714
    Vx15 0.18750000          0:0:0:0:1:1:1:0:0:0:0     3  2.6785714
     Vx9 0.16071429          0:0:0:0:0:1:0:0:0:0:1     3  2.6785714
   maxO3 0.14285714          0:1:1:1:1:0:0:0:0:0:0     2  1.7857143
  maxO3v 0.10714286          0:0:0:0:1:0:0:0:0:1:0     2  1.7857143
    Vx12 0.08928571          0:0:0:0:0:0:1:1:0:0:0     2  1.7857143
                             0:0:0:0:0:0:1:0:0:0:0     2  1.7857143
                             ......................    .  ...
```

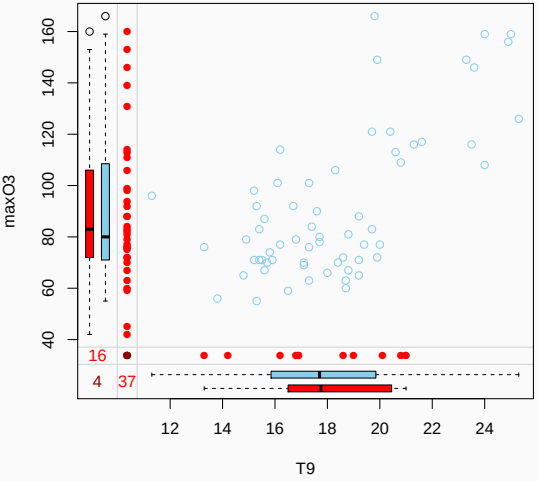
Visualisation du dispositif de données manquantes




```
> library(VIM)
> aggr(don,only.miss=TRUE,sortVar=TRUE)
```

Visualisation du dispositif de données manquantes





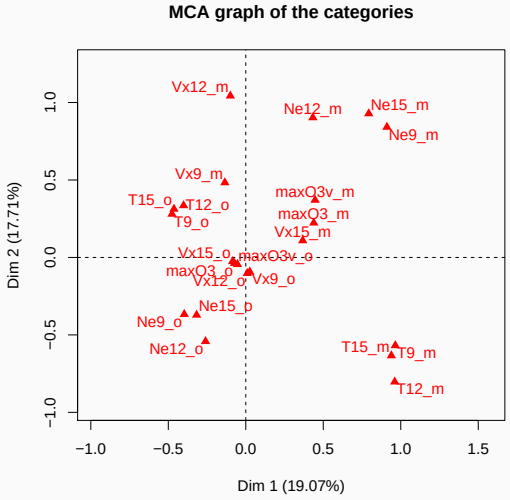

```
> library(VIM)
> matrixplot(don,sortby=2)
> marginplot(don[,c("T9","maxO3")])
```

Visualisation par l'ACM

$\Rightarrow$ Créer une matrice de présence-absence

```
> mis.ind <- matrix("o",nrow=nrow(don),ncol=ncol(don))
> mis.ind[is.na(don)]="m"
> dimnames(mis.ind)=dimnames(don)
> mis.ind

         maxO3 T9  T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
20010601 "o"   "o" "o" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010602 "o"   "m" "m" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010603 "o"   "o" "o" "o" "o" "m"  "m"  "o" "m"  "o"  "o"
20010604 "o"   "o" "o" "m" "o" "o"  "o"  "m" "o"  "o"  "o"
20010605 "o"   "m" "o" "o" "m" "m"  "m"  "o" "o"  "o"  "o"
20010606 "o"   "o" "o" "o" "o" "m"  "o"  "o" "o"  "o"  "o"
20010607 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
20010610 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
```

Visualisation par l'ACM




```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA,invis="ind",title="MCA graph of the categories")
```

Approches recommandées pour gérer les valeurs manquantes

Approche par maximum de vraisemblance

Modifier la méthode, le processus d'estimation pour gérer les données manquantes

Imputation (multiple)

Obtenir un jeu de données complété à partir duquel toute analyse statistique peut être effectuée

Algorithme EM (Dempster, Laird et Rubin, 1977)

Principe de l'algorithme d'espérance-maximisation

- Etape E (Estimation) : remplacer les valeurs manquantes par des valeurs vraisemblables grâce aux données observées et aux paramètres (obtenus à l'étape M)
- Etape M (Maximisation de la vraisemblance) : estimation des paramètres par MV en considérant les données complétées à l'étape E comme de vraies valeurs

Itérer jusqu'à convergence

Besoin de modifier le processus d'estimation (pas toujours facile !)

Hypothèse $\mathbf{x}_{i.} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Estimation ponctuelle avec EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don)) # manipulations préliminaires
> thetahat <- em.norm(pre)           # estimation par MV
> getparam.norm(pre,thetahat)        # résultats
```

Approche du Maximum de vraisemblance

Hypothèse $\mathbf{x}_{i.} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Estimation ponctuelle avec EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don)) # manipulations préliminaires
> thetahat <- em.norm(pre)           # estimation par MV
> getparam.norm(pre,thetahat)        # résultats
```

Variances

- Supplemented EM (Meng, 1991)
- Approche Bootstrap :
  - Bootstrap les lignes : $\mathbf{X}^1$, ... , $\mathbf{X}^B$
  - Algorithme EM : $(\hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\Sigma}}^1)$, ... , $(\hat{\boldsymbol{\mu}}^B, \hat{\boldsymbol{\Sigma}}^B)$

Approche du Maximum de vraisemblance

Hypothèse $\mathbf{x}_{i.} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Estimation ponctuelle avec EM

```
> library(norm)
> pre <- prelim.norm(as.matrix(don)) # manipulations préliminaires
> thetahat <- em.norm(pre)           # estimation par MV
> getparam.norm(pre,thetahat)        # résultats
```

Variances

- Supplemented EM (Meng, 1991)
- Approche Bootstrap :
  - Bootstrap les lignes : $\mathbf{X}^1$, ... , $\mathbf{X}^B$
  - Algorithme EM : $(\hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\Sigma}}^1)$, ... , $(\hat{\boldsymbol{\mu}}^B, \hat{\boldsymbol{\Sigma}}^B)$

Problème : développer une méthode spécifique pour chaque méthode statistique

Plan

1 Introduction

2 Imputation simple pour variables quantitatives

3 Imputation simple pour variables qualitatives

4 Imputation simple pour données mixtes

5 Imputation multiple

Modèle joint : un modèle global

$\Rightarrow$ Hypothèse $\mathbf{x}_{i.} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Cas bivarié avec données manquantes sur Y (régression aléatoire)

- Estimer β et σ
- Tirer à partir de la distribution prédictive $y_i \sim \mathcal{N}\left(x_i\hat{\beta}, \hat{\sigma}^2\right)$

Extension au cas multivarié

- Estimer $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ à partir d'un jeu incomplet avec EM
- Tirer à partir de $\mathcal{N}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right)$

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> rngseed(123)
> imp <- imp.norm(pre,thetahat,don)
```

Modèle conditionnel : un modèle par variable

Exemple avec régression :

1. Initialisation de l'imputation : imputation par la moyenne
2. Ajuster une régression aléatoire de $\mathbf{X}_j^{obs}$ en fonction des autres variables $\mathbf{X}_{-j}^{obs}$
   Prédire $\mathbf{X}_j^{miss}$ à partir du modèle ajusté
3. Boucler sur les variables

```
> library(mice)
> res.cm <- mice(don, m=1)
```

Modèle conditionnel : un modèle par variable

Exemple avec régression :

1. Initialisation de l'imputation : imputation par la moyenne
2. Ajuster une régression aléatoire de $\mathbf{X}_j^{obs}$ en fonction des autres variables $\mathbf{X}_{-j}^{obs}$
   Prédire $\mathbf{X}_j^{miss}$ à partir du modèle ajusté
3. Boucler sur les variables

```
> library(mice)
> res.cm <- mice(don, m=1)
```

$\Rightarrow$ Flexibilité : différents modèles pour chaque variable

Autres méthodes d'imputation simple

- k-plus proches voisins (`class`, `FNN`)
- forêts aléatoires (`missForest`, Stekhoven & Bühlmann, 2011)
- ...

⇒ R CRAN task View: Missing Data
⇒ R-miss-tastic

⇒ Imputation par ACP

Ajustement du nuage en ACP

L'ACP vise à trouver le sous-espace qui fournit la meilleure représentation des données

Ajustement du nuage en ACP

L'ACP vise à trouver le sous-espace qui fournit la meilleure représentation des données

Figure 1: Chameau ou dromadaire ? source J.P. Fenelon

Ajustement du nuage en ACP

L'ACP vise à trouver le sous-espace qui fournit la meilleure représentation des données

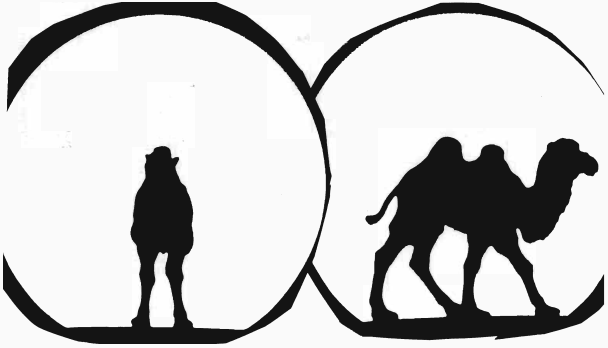


Figure 1: Chameau ou dromadaire ? source J.P. Fenelon

Ajustement du nuage en ACP

L'ACP vise à trouver le sous-espace qui fournit la meilleure représentation des données

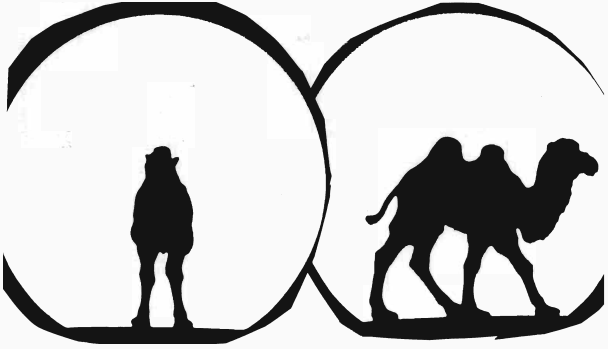


Figure 1: Chameau ou dromadaire ? source J.P. Fenelon

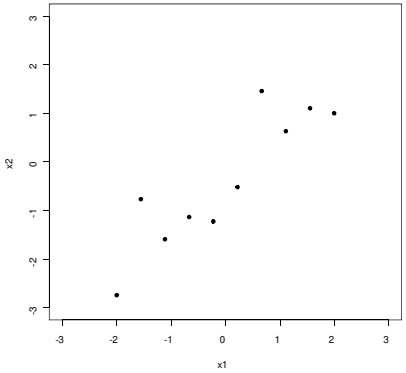
$\Rightarrow$ Meilleure approximation par projection

$\Rightarrow$ Meilleure représentation de la diversité, de la variabilité

Ajustement du nuage en ACP

X

−2.00	−2.36
−1.56	−0.39
−1.11	−1.21
−0.67	−0.75
−0.22	−0.84
0.22	−0.14
0.67	1.84
1.11	1.01
1.56	1.48
2.00	1.38

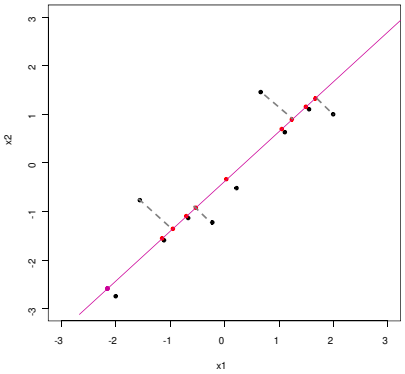


X : données en 2 dimensions

Ajustement du nuage en ACP

X

−2.00	−2.36
−1.56	−0.39
−1.11	−1.21
−0.67	−0.75
−0.22	−0.84
0.22	−0.14
0.67	1.84
1.11	1.01
1.56	1.48
2.00	1.38



−2.16	−2.21
−0.96	−0.98
−1.15	−1.17
−0.70	−0.72
−0.53	−0.54
0.04	0.04
1.25	1.27
1.05	1.07
1.50	1.54
1.67	1.70

$\hat{\mathbf{X}}$

X : données en 2 dimensions

Minimisation de la distance entre les individus et leur projection

Reconstitution en ACP

X

−2.00	−2.36
−1.56	−0.39
−1.11	−1.21
−0.67	−0.75
−0.22	−0.84
0.22	−0.14
0.67	1.84
1.11	1.01
1.56	1.48
2.00	1.38

D

5.60

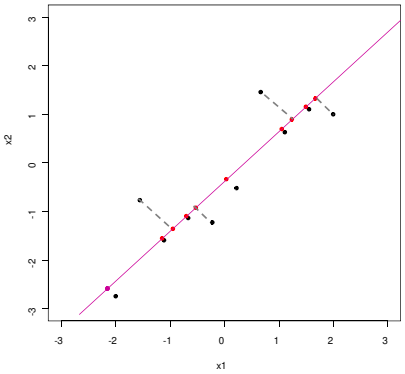
V'

0.699	0.715

U

−0.55
−0.24
−0.29
−0.18
−0.13
0.01
0.32
0.27
0.38
0.43

−2.16	−2.21
−0.96	−0.98
−1.15	−1.17
−0.70	−0.72
−0.53	−0.54
0.04	0.04
1.25	1.27
1.05	1.07
1.50	1.54
1.67	1.70



$\hat{\mathbf{X}} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}'$

$\hat{\mathbf{X}} = \mathbf{M} + \mathbf{UDV}'$ (produit matriciel utilisant les coordonnées des individus et les coordonnées des variables issues de l'ACP)

ACP : cas complet

$\Rightarrow$ Point de vue géométrique : minimiser l'erreur de reconstitution

$\Rightarrow$ Approximation de $\mathbf{X}$ par une matrice de rang $S < p$:

$$\|\mathbf{X}_{n\times p} - \hat{\mathbf{X}}_{n\times p}\|^2 \qquad \text{SVD : } \hat{\mathbf{X}}^{\text{ACP}} = \mathbf{M}_{n\times p} + \mathbf{U}_{n\times S}\mathbf{D}_{S\times S}\mathbf{V}'_{p\times S}$$

$\mathbf{F} = \mathbf{UD}$ composantes principales (scores)

$\mathbf{V}$ axes principaux (loadings)

$\Rightarrow$ Point de vue géométrique : minimiser l'erreur de reconstitution

$\Rightarrow$ Approximation de $\mathbf{X}$ par une matrice de rang $S < p$:

$$\|\mathbf{X}_{n\times p} - \hat{\mathbf{X}}_{n\times p}\|^2 \qquad \text{SVD : } \hat{\mathbf{X}}^{\text{ACP}} = \mathbf{M}_{n\times p} + \mathbf{U}_{n\times S}\mathbf{D}_{S\times S}\mathbf{V}'_{p\times S}$$

$\mathbf{F} = \mathbf{UD}$ composantes principales (scores)
$\mathbf{V}$ axes principaux (loadings)

$\Rightarrow$ Point de vue modèle à effets fixes (Caussinus, 1986)

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

$$x_{ij} = m_j + \sum_{s=1}^{S} d_s u_{is} v_{js} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Estimateurs de maximum de vraisemblance = estimateurs des moindres carrés

$\Rightarrow$ ACP : moindres carrés

$$\left\| \mathbf{X}_{n\times p} - \left(\mathbf{M}_{n\times p} + \mathbf{U}_{n\times S}\mathbf{D}_{S\times S}\mathbf{V}'_{p\times S} \right) \right\|^2$$

$\Rightarrow$ ACP : moindres carrés

$$\left\| \mathbf{X}_{n\times p} - \left(\mathbf{M}_{n\times p} + \mathbf{U}_{n\times S}\mathbf{D}_{S\times S}\mathbf{V}'_{p\times S} \right) \right\|^2$$

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

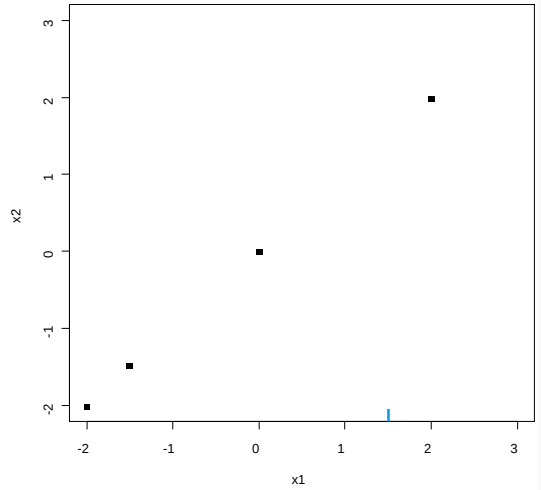
$$\left\| \mathbf{R}_{n\times p} * \left(\mathbf{X}_{n\times p} - \left(\mathbf{M}_{n\times p} + \mathbf{U}_{n\times S}\mathbf{D}_{S\times S}\mathbf{V}'_{p\times S} \right) \right) \right\|^2$$

with $r_{ij} = 0$ si x_{ij} manquant, $r_{ij} = 1$ sinon

Beaucoup d'algorithmes : moindres carrés pondérés alterné (Gabriel & Zamir, 1979) ; ACP iterative (Kiers, 1997)

ACP itérative

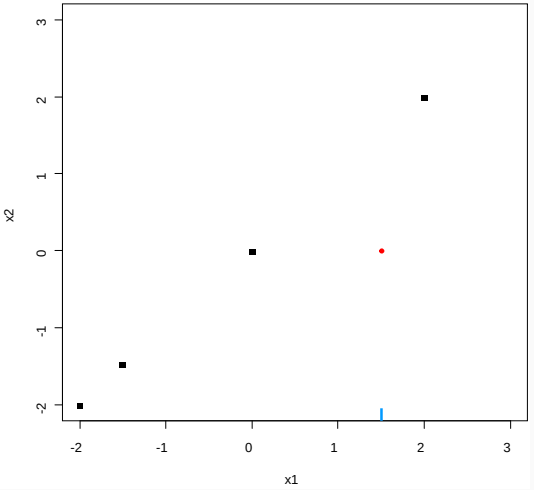
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98
```



ACP itérative

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

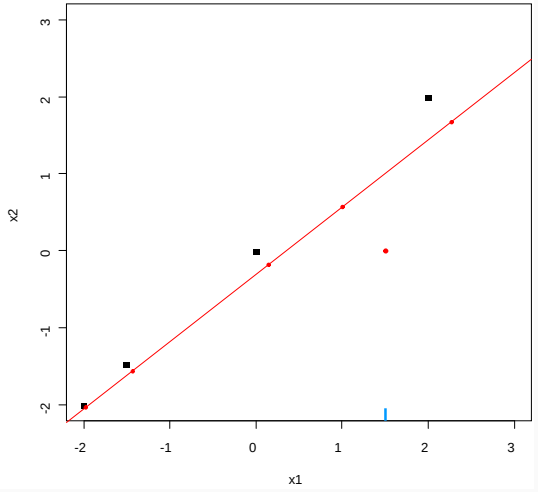
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98



Initialisation $\ell = 0$: $\mathbf{X}^0$ (imputation par la moyenne)

ACP itérative

```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98

   ^     ^
   x1    x2
-1.98 -2.04
-1.44 -1.56
 0.15 -0.18
 1.00  0.57
 2.27  1.67
```



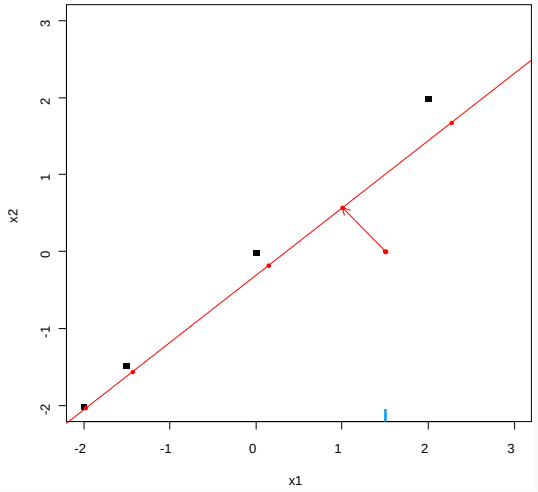
ACP sur le jeu de données complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$;

ACP itérative

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\langle$ x1	$\langle$ x2
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67



Valeurs manquantes imputées par le modèle $\hat{\mathbf{X}}^{\ell} = \mathbf{M}^{\ell} + \mathbf{U}^{\ell}\mathbf{D}^{\ell}\mathbf{V}^{\ell\prime}$

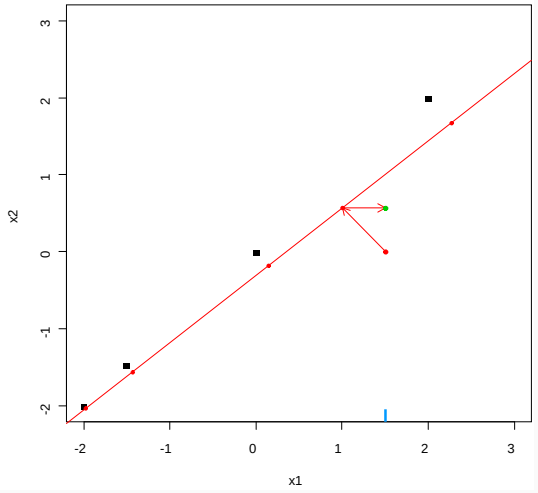
ACP itérative

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



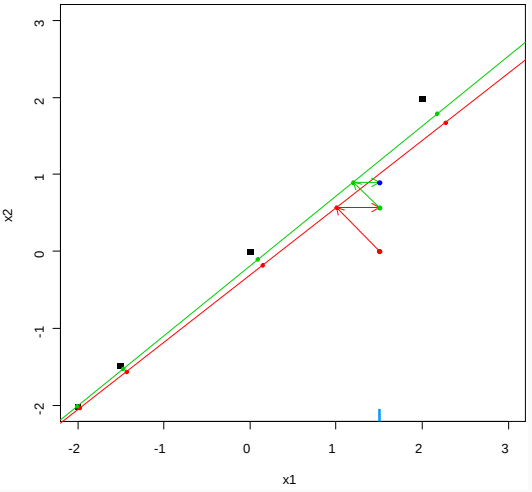
Nouveau jeu de données imputé $\mathbf{X}^{\ell} = \mathbf{R} * \mathbf{X} + (1 - \mathbf{R}) * \hat{\mathbf{X}}^{\ell}$

ACP itérative

```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98
```

```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98
```

```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98
```



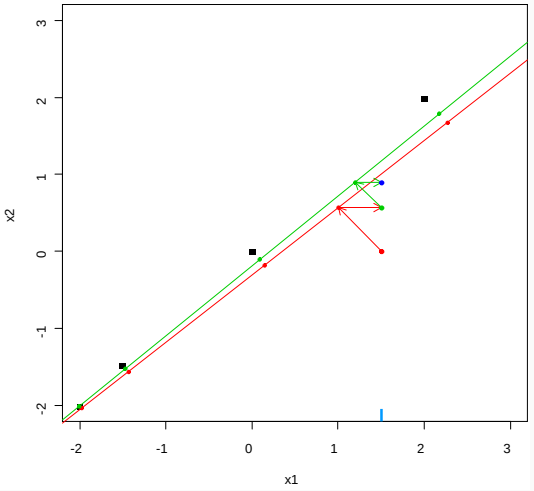
ACP itérative

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

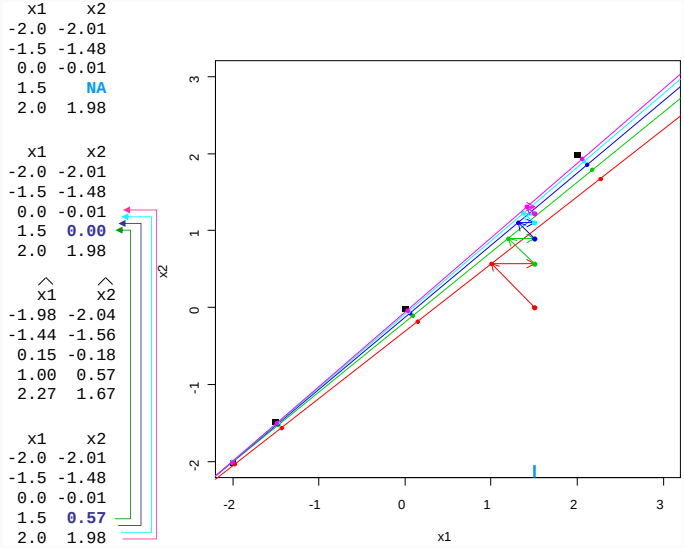
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	**0.90**
2.18	1.78

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.90**
2.0	1.98



ACP itérative

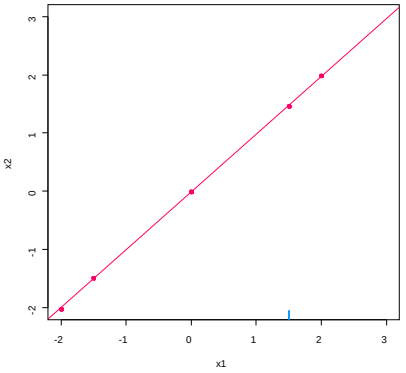


Les étapes sont répétées jusqu'à convergence

ACP itérative

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	1.46
2.0	1.98



ACP sur le jeu de données complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$

Valeurs manquantes imputées par le modèle $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell\prime}$

ACP itérative

1. initialisation $\ell = 0$: $\mathbf{X}^0$ (imputation par la moyenne)
2. step ℓ :
   (a) ACP sur le tableau complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$;
   S dimensions conservées
   (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell\prime}$;
   nouveau tableau imputé $\mathbf{X}^\ell = \mathbf{R} * \mathbf{X} + (1 - \mathbf{R}) * \hat{\mathbf{X}}^\ell$
3. étapes répétées jusqu'à convergence

ACP itérative

1. initialisation $\ell = 0$: $\mathbf{X}^0$ (imputation par la moyenne)
2. step ℓ :
   (a) ACP sur le tableau complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$;
   S dimensions conservées
   (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell\prime}$;
   nouveau tableau imputé $\mathbf{X}^\ell = \mathbf{R} * \mathbf{X} + (1 - \mathbf{R}) * \hat{\mathbf{X}}^\ell$
   (c) moyennes (et écarts-types) sont mis à jour
3. étapes répétées jusqu'à convergence

ACP itérative

1. initialisation $\ell = 0$: $\mathbf{X}^0$ (imputation par la moyenne)
2. step ℓ :
   (a) ACP sur le tableau complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$;
   S dimensions conservées
   (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell\prime}$;
   nouveau tableau imputé $\mathbf{X}^\ell = \mathbf{R} * \mathbf{X} + (1 - \mathbf{R}) * \hat{\mathbf{X}}^\ell$
   (c) moyennes (et écarts-types) sont mis à jour
3. étapes répétées jusqu'à convergence

ACP itérative

1. initialisation $\ell = 0$: $\mathbf{X}^0$ (imputation par la moyenne)
2. step ℓ :
   (a) ACP sur le tableau complété $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$;
   S dimensions conservées
   (b) valeurs manquantes imputées par $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell\prime}$;
   nouveau tableau imputé $\mathbf{X}^\ell = \mathbf{R} * \mathbf{X} + (1 - \mathbf{R}) * \hat{\mathbf{X}}^\ell$
   (c) moyennes (et écarts-types) sont mis à jour
3. étapes répétées jusqu'à convergence

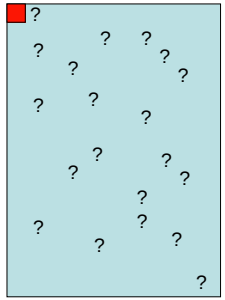
$\Rightarrow$ algorithme EM pour le modèle à effets fixes

$\Rightarrow$ Imputation (complétion de matrice, Netflix)

$\Rightarrow$ Réduction de la variabilité (imputation par $\mathbf{M} + \mathbf{UDV}'$)

Choix du nombre de composantes

Choix du nombre de composantes

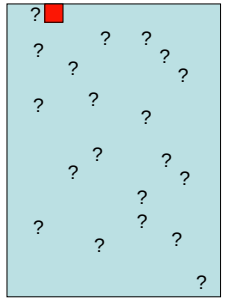


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}(s) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}^{s;-\{ij\}})^2$$

$\Rightarrow$ Très coûteux en temps de calcul

Choix du nombre de composantes

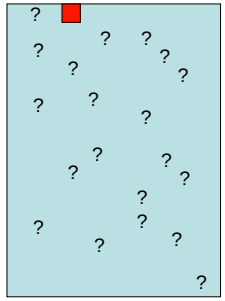


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}(s) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}^{s;-\{ij\}})^2$$

$\Rightarrow$ Très coûteux en temps de calcul

Choix du nombre de composantes

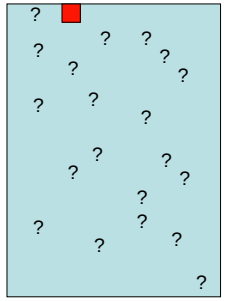


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}(s) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}^{s;-\{ij\}})^2$$

$\Rightarrow$ Très coûteux en temps de calcul

Choix du nombre de composantes



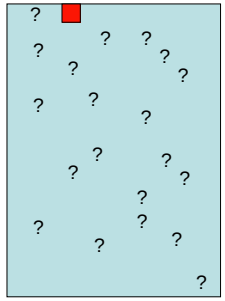
$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}(s) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}^{s;-\{ij\}})^2$$

$\Rightarrow$ Très coûteux en temps de calcul

Ajouter plusieurs valeurs manquantes supplémentaires simultanément

Choix du nombre de composantes



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}(s) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}^{s;-\{ij\}})^2$$

$\Rightarrow$ Très coûteux en temps de calcul

Ajouter plusieurs valeurs manquantes supplémentaires simultanément

Approximation possible par validation croisée généralisée $\Longrightarrow$ gain en temps de calcul

Propriétés

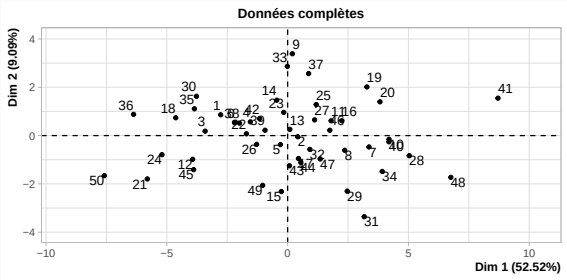
- Résultats de l'ACP obtenus à partir des données observées uniquement : graphe des individus et graphe des variables
  $\Rightarrow$ On "saute" les données manquantes, l'ACP itérative minimise
  $\|\mathbf{R} * (\mathbf{X} - (\mathbf{M} + \mathbf{UDV}'))\|^2$

- Imputation :
  - prend en compte les ressemblances entre individus et les liaisons entre variables
  - le tableau imputé peut être utilisé (avec précaution) pour réaliser d'autres analyses

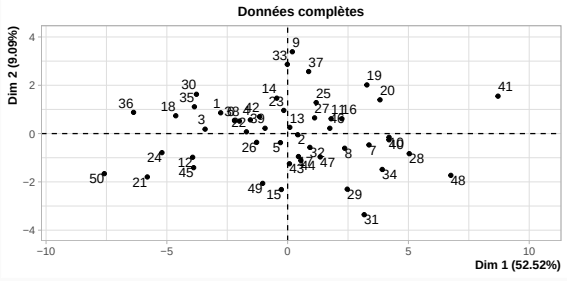
- Problème de surajustement

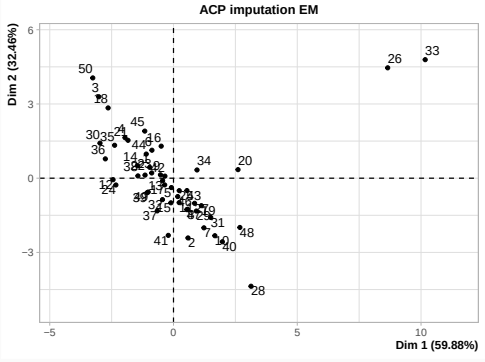
Surajustement

$$X_{50\times 10} = \mathbf{U}_{50\times 2}\mathbf{D}\mathbf{V}'_{10\times 2} + \mathcal{N}(0, 0.5)\,;$$



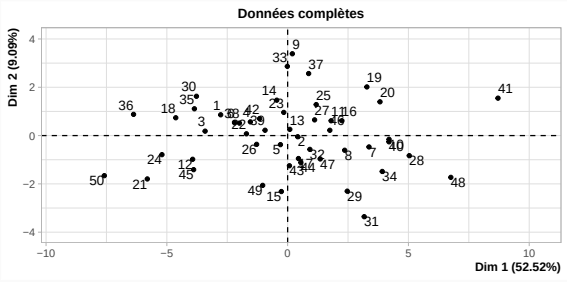
$X_{50\times 10} = \mathbf{U}_{50\times 2}\mathbf{D}\mathbf{V}'_{10\times 2} + \mathcal{N}(0, 0.5)$; 50% of NA

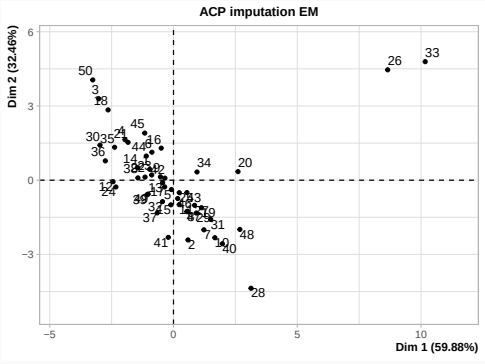




Surajustement

$X_{50\times 10} = \mathbf{U}_{50\times 2}\mathbf{D}\mathbf{V}'_{10\times 2} + \mathcal{N}(0, 0.5)$; 50% of NA





$\Rightarrow$ erreur d'ajustement faible : $||\mathbf{R} * (\mathbf{X} - \hat{\mathbf{X}})||^2 = 0.50$

$\Rightarrow$ erreur de prédiction élevée : $||(1 - \mathbf{R}) * (\mathbf{X} - \hat{\mathbf{X}})||^2 = 16.98$

Surajustement

$\Rightarrow$ Bon ajustement et mauvaise prédiction

- Trop de paramètres sont estimés par rapport au nombre de données observées : le nombre de dimension S et le nombre de données manquantes sont grands
- Faibles liaisons entre variables

1. Diminuer le nombre S
2. Early stopping
3. Régularisation $\Rightarrow$ ACP itérative régularisée

ACP itérative régularisée (Josse *et al.*, 2009)

$\Rightarrow$ Initialisation - étape d'estimation - étape d'imputation

L'étape d'imputation :

$$\hat{x}_{ij}^{\text{ACP}} = \sum_{s=1}^{S} d_s u_{is} v_{js}$$

est remplacée par une étape d'imputation régularisée :

$$\hat{x}_{ij}^{\text{rACP}} = \sum_{s=1}^{S} \left(\frac{d_s^2 - \hat{\sigma}^2}{d_s^2} \right) d_s u_{is} v_{js} = \sum_{s=1}^{S} \left(d_s - \frac{\hat{\sigma}^2}{d_s} \right) u_{is} v_{js}$$

ACP itérative régularisée (Josse *et al.*, 2009)

$\Rightarrow$ Initialisation - étape d'estimation - étape d'imputation

L'étape d'imputation :

$$\hat{x}_{ij}^{\text{ACP}} = \sum_{s=1}^{S} d_s u_{is} v_{js}$$

est remplacée par une étape d'imputation régularisée :

$$\hat{x}_{ij}^{\text{rACP}} = \sum_{s=1}^{S} \left(\frac{d_s^2 - \hat{\sigma}^2}{d_s^2} \right) d_s u_{is} v_{js} = \sum_{s=1}^{S} \left(d_s - \frac{\hat{\sigma}^2}{d_s} \right) u_{is} v_{js}$$

$$\hat{\sigma}^2 = \frac{RSS}{\text{ddl}} = \frac{n \sum_{s=S+1}^{p} d_s^2}{(n-1-S)(p-S)}$$

ACP itérative régularisée (Josse *et al.*, 2009)

$\Rightarrow$ Initialisation - étape d'estimation - étape d'imputation

L'étape d'imputation :

$$\hat{x}_{ij}^{\text{ACP}} = \sum_{s=1}^{S} d_s u_{is} v_{js}$$

est remplacée par une étape d'imputation régularisée :

$$\hat{x}_{ij}^{\text{rACP}} = \sum_{s=1}^{S} \left(\frac{d_s^2 - \hat{\sigma}^2}{d_s^2} \right) d_s u_{is} v_{js} = \sum_{s=1}^{S} \left(d_s - \frac{\hat{\sigma}^2}{d_s} \right) u_{is} v_{js}$$

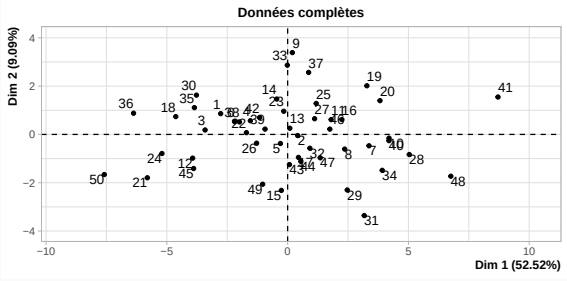
$$\hat{\sigma}^2 = \frac{RSS}{\text{ddl}} = \frac{n \sum_{s=S+1}^{p} d_s^2}{(n-1-S)(p-S)}$$

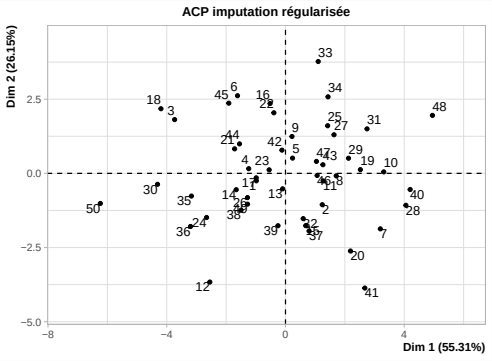
Compromis seuillage doux/dur (Mazumder, Hastie & Tibshirani, 2010)

σ^2 petit $\rightarrow$ ACP régularisée $\approx$ ACP

σ^2 grand $\rightarrow$ imputation par la moyenne

$X_{50\times 10} = \mathbf{U}_{50\times 2}\mathbf{D}\mathbf{V}'_{10\times 2} + \mathcal{N}(0, 0.5)$; 50% of NA





$\Rightarrow$ erreur d'ajustement : $||\mathbf{R} * (\mathbf{X} - \hat{\mathbf{X}})||^2 = 0.56$ (EM= 0.50)

$\Rightarrow$ erreur de prédiction : $||(1 - \mathbf{R}) * (\mathbf{X} - \hat{\mathbf{X}})||^2 = 2.28$ (EM= 16.98)

Propriétés de l'imputation par ACP

Bilan

- L'ACP itérative régularisée permet d'imputer les valeurs manquantes d'un jeu incomplet
- Le tableau imputé peut être directement utilisé par un algorithme classique d'ACP
- Les valeurs imputées n'ont aucun poids dans le critère utilisé pour construire axes et composantes d'une ACP
- Bonne qualité d'imputation quand la structure du jeu de données est forte (imputation utilisant les ressemblances entre individus et les liaisons entre variables)
- Bien meilleur que l'algorithme Nipals (encore trop utilisé)
- Compétitif par rapport aux forêts aléatoires

Gestion des éléments supplémentaires dans l'ACP

Quid des éléments supplémentaires ?

Idée : pondérer les éléments supplémentaires (variables quantitatives, individus supplémentaires)

1. Mettre un poids nul aux éléments supplémentaires qui ne contribueront pas à la construction des dimensions
2. Lancer l'algorithme d'ACP itérative régularisée avec ces poids : l'imputation n'utilise pas l'information portée par les éléments supplémentaires
3. Lancer ensuite l'ACP sur le tableau complété en utilisant la fonction classique d'ACP avec éléments supplémentaires

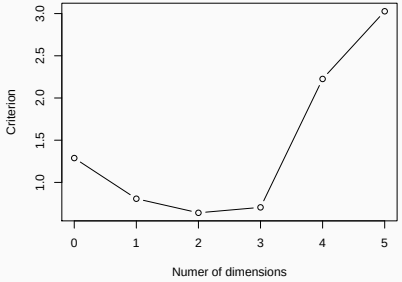
Tutoriel sur l'ACP avec données manquantes

(données ozone, lignes de code)

$\Rightarrow$ Etape 1 : Estimation du nombre de dimensions

(Validation croisée, Bro, 2008 ; GCV, Josse & Husson, 2011)

```
> library(missMDA)
> nb <- estim_ncpPCA(don, method.cv="Kfold")
> nb$ncp    #2
> plot(0:5, nb$criterion, xlab="nb dim", ylab="MSEP")
```

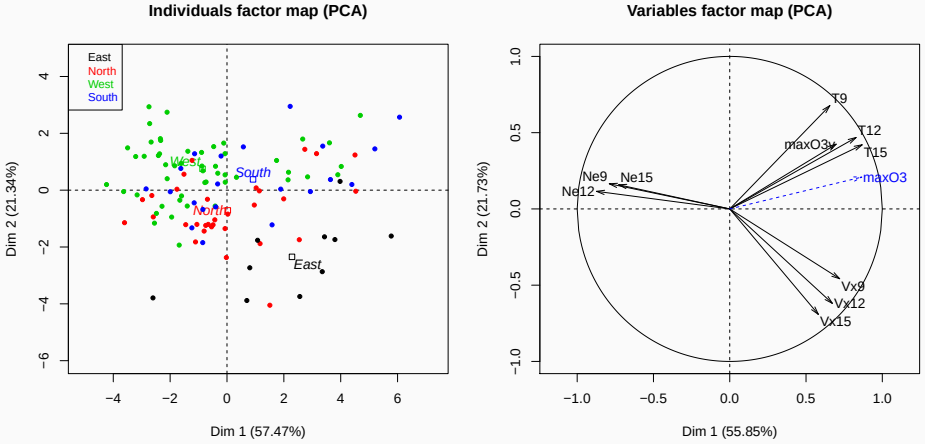


Imputation par ACP en pratique

⇒ Etape 2 : Imputation des données manquantes

```
> res.comp <- imputePCA(don, ncp = 2)
> res.comp$completeObs[1:3,]
     maxO3    T9   T12   T15 Ne9 Ne12 Ne15   Vx9  Vx12  Vx15 maxO3v
0601    87 15.60 18.50 20.47   4 4.00 8.00  0.69 -1.71 -0.69     84
0602    82 18.51 20.88 21.81   5 5.00 7.00 -4.33 -4.00 -3.00     87
0603    92 15.30 17.60 19.50   2 3.98 3.81  2.95  1.97  0.52     82
```

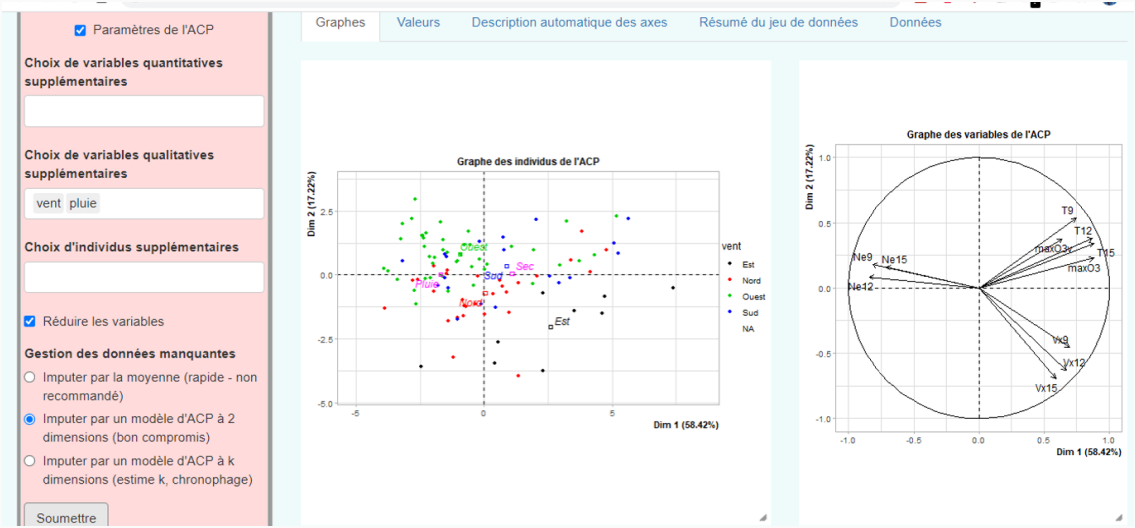
⇒ Etape 3 : ACP sur le tableau complété




```
> imp <- cbind.data.frame(res.comp$completeObs, ozone[,12])
> res.pca <- PCA(imp, quanti.sup=1, quali.sup=12)
> plot(res.pca, hab=12, lab="quali")
> plot(res.pca, choix="var")
```

3 en 1 avec le package Factoshiny

```
> library(Factoshiny)
> Factoshiny(ozone)
```



Jeu de données en écologie

Données Glopnet : 2494 espèces décrites par 6 variables quantitatives (données, lignes de code)

- LMA (leaf mass per area)
- LL (leaf lifespan)
- Amass (photosynthetic assimilation)
- Nmass (leaf nitrogen)
- Pmass (leaf phosphorus)
- Rmass (dark respiration rate)

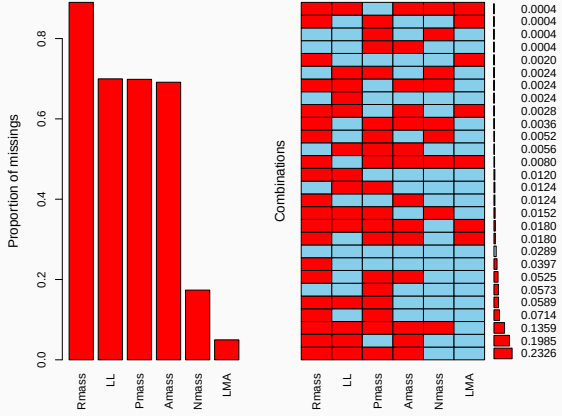
et 1 variable qualitative : le biome (macro-écosystème)

Wright IJ, et al. (2004). The worldwide leaf economics spectrum. *Nature*, 428 :821.
www.nature.com/nature/journal/v428/n6985/extref/nature02403-s2.xls

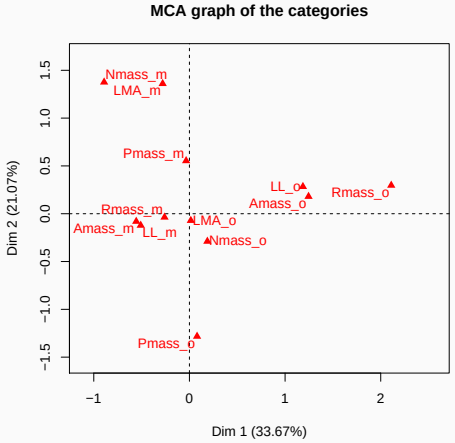
Jeu de données en écologie

```
> sum(is.na(don))/(nrow(don)*ncol(don)) # 53% de données manquantes
[1] 0.5338145
> dim(na.omit(don))  ## suppression des espèces avec données manquantes
[1] 72  6             ## reste seulement 72 espèces!
```

```
> library(VIM)
> aggr(don,numbers=TRUE,sortVar=TRUE)
```

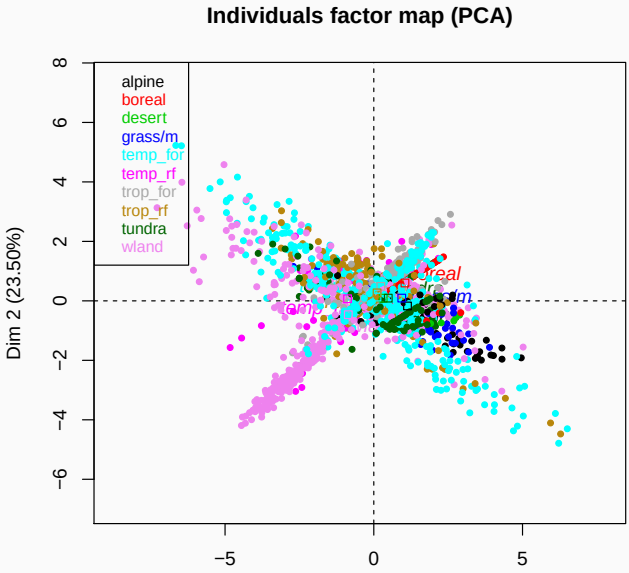


Jeu de données en écologie

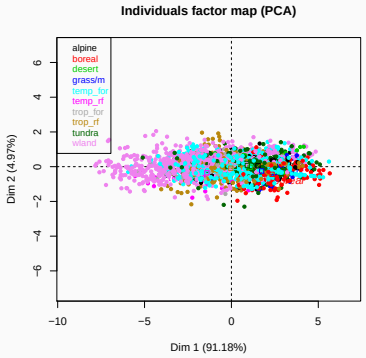


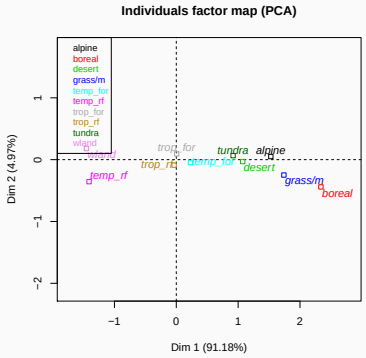

```
> mis.ind <- matrix("o",nrow=nrow(don),ncol=ncol(don))
> mis.ind[is.na(don)] <- "m"
> dimnames(mis.ind) <- dimnames(don)
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA,invis="ind",title="MCA graph of the categories")
```

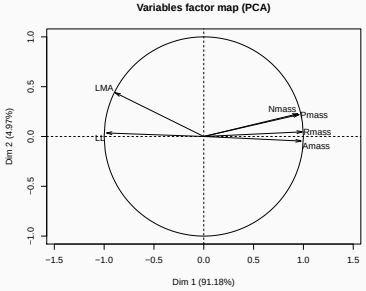
Quid de l'imputation par la moyenne ?



Jeu de données en écologie








```
> library(missMDA)
> nb <- estim_ncpPCA(don,method.cv="Kfold",nbsim=100)
> res.comp <- imputePCA(don,ncp=2)
> imp <- cbind.data.frame(res.comp$completeObs,tab.init[,1:4])
> res.pca <- PCA(imp,quanti.sup=1,quali.sup=12)
> plot(res.pca, hab=12, lab="quali")
> plot(res.pca, choix="var")
> res.pca$ind$coord #scores (principal components)
```

Plan

1. Introduction
2. Imputation simple pour variables quantitatives
3. Imputation simple pour variables qualitatives
4. Imputation simple pour données mixtes
5. Imputation multiple

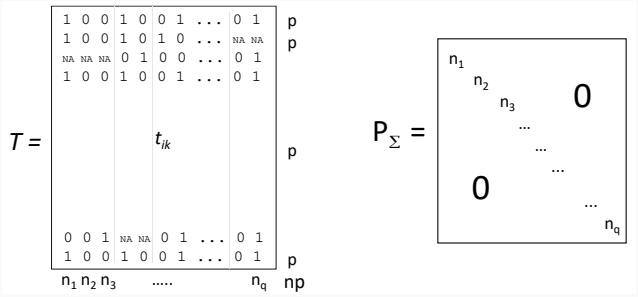
Les méthodes d'analyse factorielle

- Analyse exploratoire de tableaux de données
- Dépend de la structure et de la nature 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

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

Traitement d'un questionnaire avec *missing single*

Les données

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

Traitement d'un questionnaire avec *missing single*

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

Traitement d'un questionnaire avec *missing single*

Les données

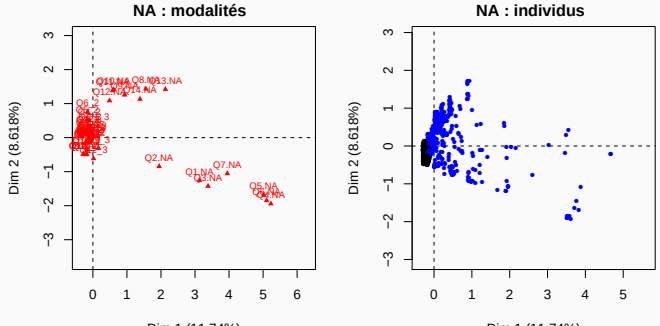
1232 répondants, 14 questions, 35 modalités, 9% de NA pour 42% des répondants

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 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
   (b) Imputation des données manquantes par les données reconstituées
   (c) Mise à jour des marges

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

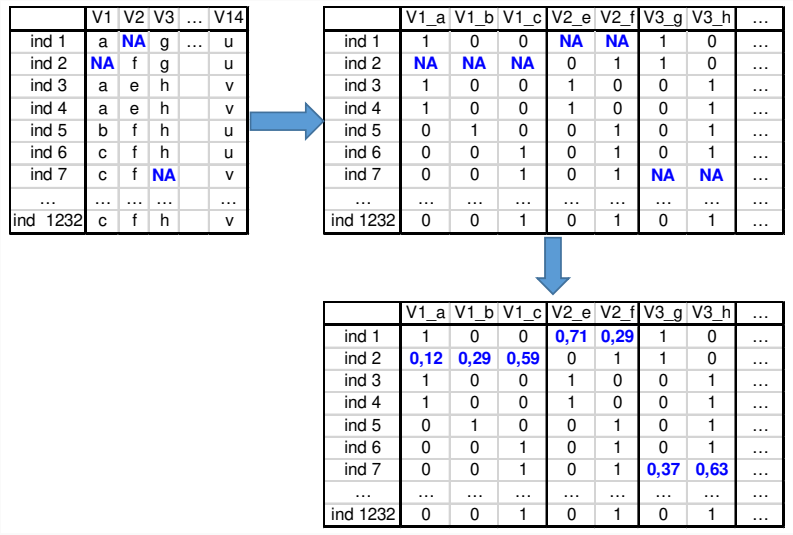
	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

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)



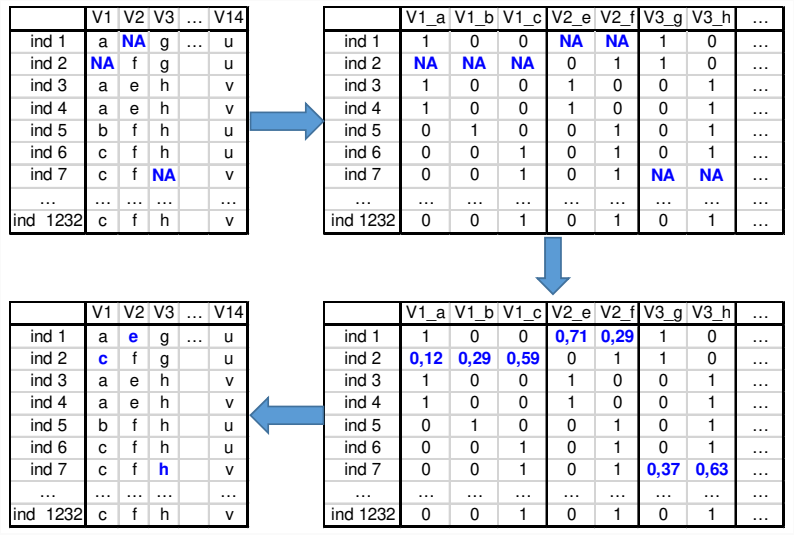
	V1	V2	V3	…	V14
ind 1	a	**NA**	g	…	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
…	…	…	…		…
ind 1232	c	f	h		v

	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

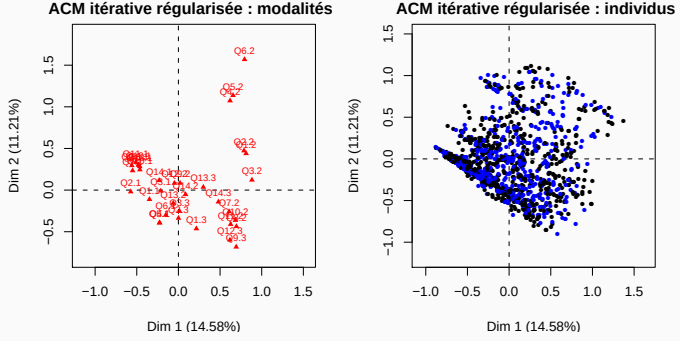
Les valeurs imputées peuvent être vues comme des degrés d'appartenance

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

1 Introduction

2 Imputation simple pour variables quantitatives

3 Imputation simple pour variables qualitatives

4 Imputation simple pour données mixtes

5 Imputation multiple

Données mixtes

Modèle joint

- General location model (Schafer, 1997) $\Longrightarrow$ problème quand beaucoup de modalités
- Transformer les variables qualitatives en indicatrices et faire comme si les variables étaient continues (`Amelia`)
- Modèle à classes latentes (Vermunt) – modèles Bayésien non paramétrique (Dunson, Reiter, Duke University)

Modèle conditionnel

- Linéaire, logistique, multinomial, logit (`mice`)
- Forêts aléatoires (Stekhoven & Bühlmann, 2012, `missForest`)

Données mixtes

Modèle joint

- General location model (Schafer, 1997) $\Longrightarrow$ problème quand beaucoup de modalités
- Transformer les variables qualitatives en indicatrices et faire comme si les variables étaient continues (`Amelia`)
- Modèle à classes latentes (Vermunt) – modèles Bayésien non paramétrique (Dunson, Reiter, Duke University)

Modèle conditionnel

- Linéaire, logistique, multinomial, logit (`mice`)
- Forêts aléatoires (Stekhoven & Bühlmann, 2012, `missForest`)

$\Rightarrow$ Analyse factorielle de données mixtes (Audigier, Husson & Josse, 2014, `missMDA`)

Imputation itérative par forêts aléatoires

1. Imputation initiale : moyenne - modalité au hasard
   Trier les variables en fonction du nombre de valeurs manquantes
2. Ajuster une forêt aléa $\mathbf{X}_j^{obs}$ en fct de $\mathbf{X}_{-j}^{obs}$ puis prédire $\mathbf{X}_j^{miss}$
3. Boucler sur les variables jusqu'à un critère d'arrêt

Imputation itérative par forêts aléatoires

1. Imputation initiale : moyenne - modalité au hasard
   Trier les variables en fonction du nombre de valeurs manquantes
2. Ajuster une forêt aléa $\mathbf{X}_j^{obs}$ en fct de $\mathbf{X}_{-j}^{obs}$ puis prédire $\mathbf{X}_j^{miss}$
3. Boucler sur les variables jusqu'à un critère d'arrêt

$\Rightarrow$ Propriétés :

- Relations non-linéaires, interactions complexes
- $n << p$
- erreur out-of-bag : approximation de l'erreur d'imputation

$\Rightarrow$ Meilleur que plus proches voisins et `mice`

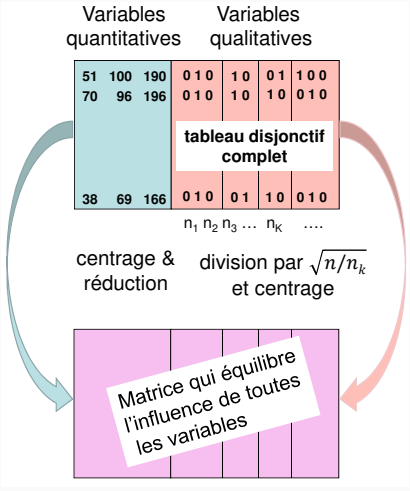
AFDM (Escofier, 1979), PCAMIX (Kiers, 1991)

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

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

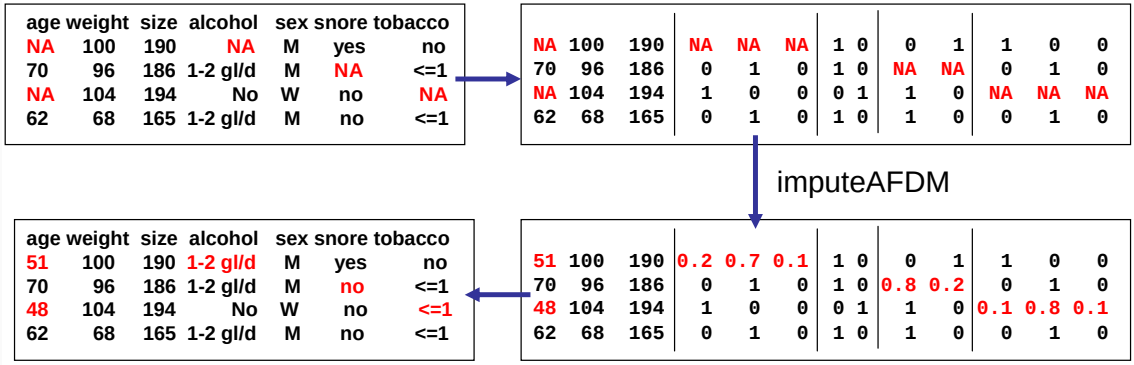
- Les composantes principales $\mathbf{F}_s$ maximisent :

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



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



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

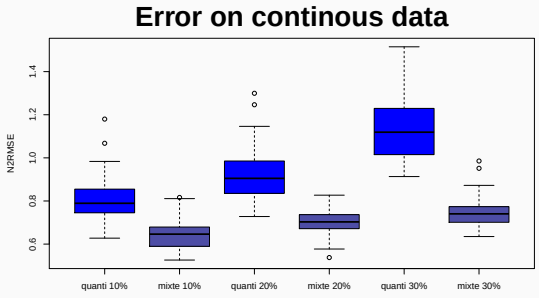
- Critère
  - pour données quantitatives :

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

  - pour données qualitatives : proportion de modalités mal prédites

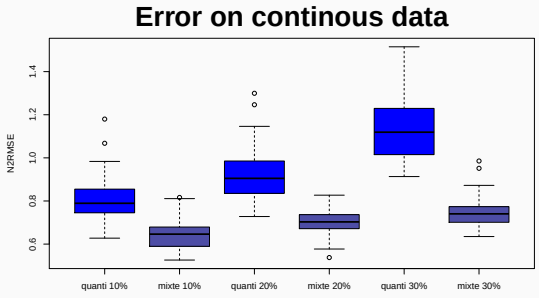
Imputation avec var. quanti uniquement

Imputation avec variables quanti et quali



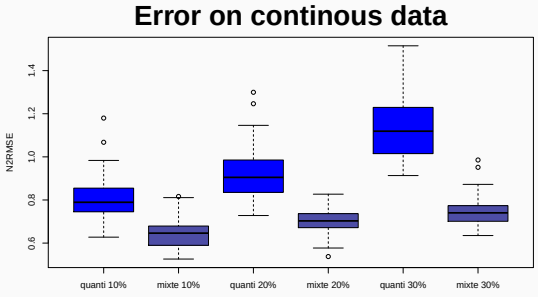
Imputation avec var. quanti uniquement

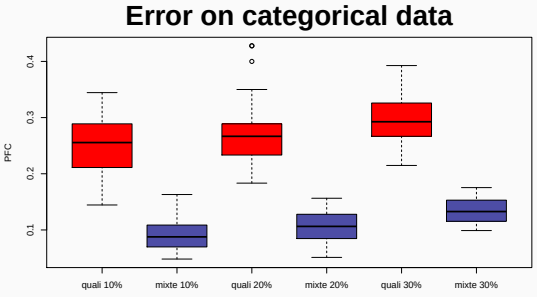
Imputation avec variables quanti et quali



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

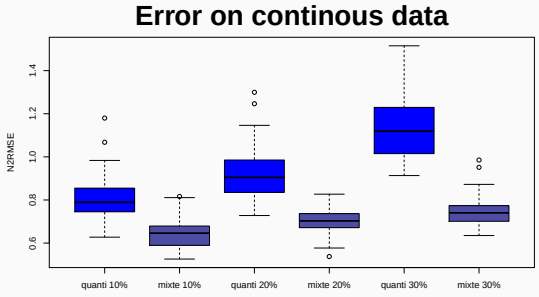
Imputation avec var. quanti uniquement Imputation avec var. quali uniquement
Imputation avec variables quanti et quali

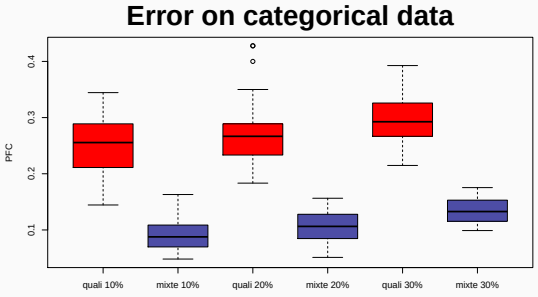




Variables quali améliorent
l'imputation sur variables quanti ...

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

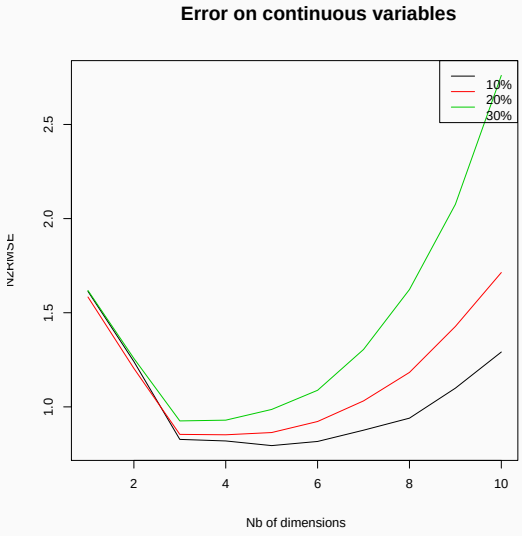


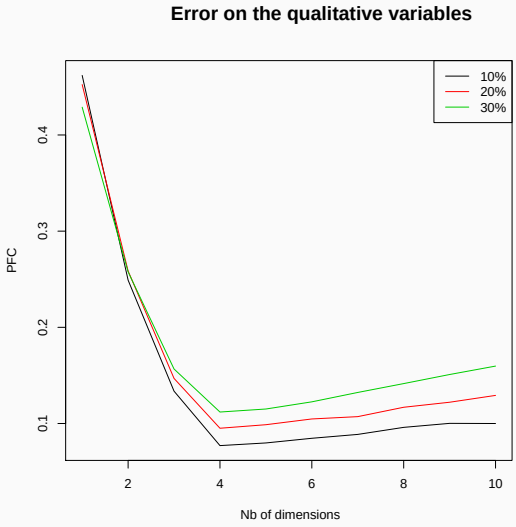


Variables quali améliorent
l'imputation sur variables quanti ...

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

Simulations



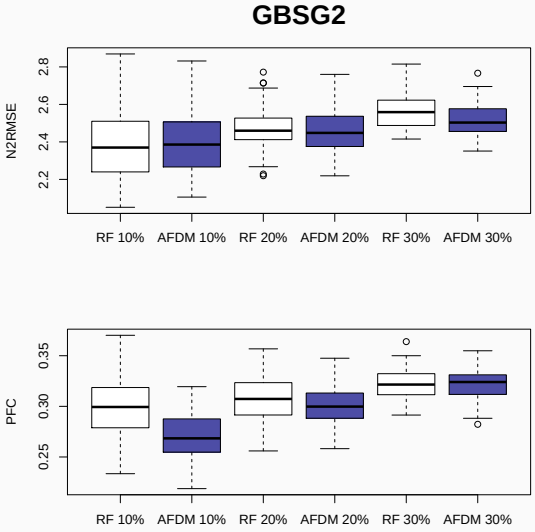


$\Rightarrow$ L'erreur sur le choix du nombre de dimensions a un impact faible sur l'erreur d'imputation

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

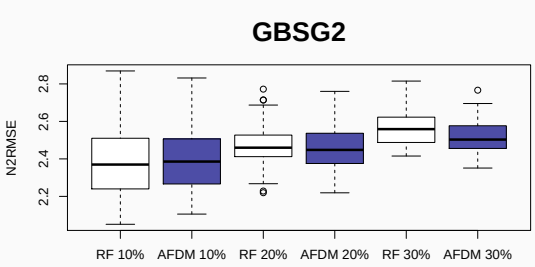
Comparaison avec forêts aléatoires

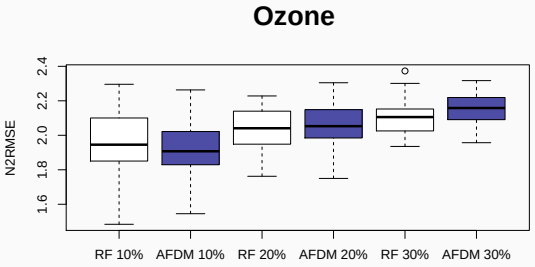
Imputations obtenues par forêts aléatoires & ACP itérative

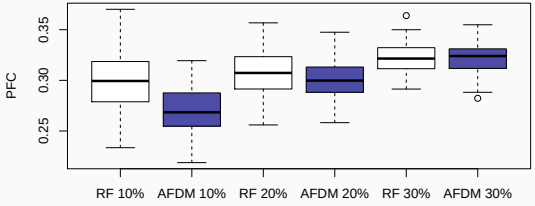


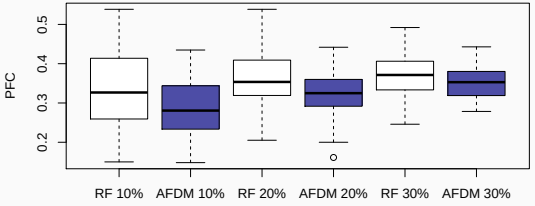
Comparaison avec forêts aléatoires

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` – `Factoshiny`

⇒ 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)

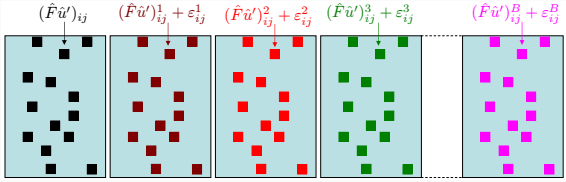
Plan

1. Introduction
2. Imputation simple pour variables quantitatives
3. Imputation simple pour variables qualitatives
4. Imputation simple pour données mixtes
5. Imputation multiple

Imputation multiple

Imputation simple : une valeur unique ne peut pas refléter l'incertitude sur la prédiction $\Rightarrow$ sous-estimation de l'écart-type

1. Générer M valeurs possibles pour chaque valeur manquante



2. Faire l'analyse sur chaque tableau imputé : $\hat{\theta}_m$, $\widehat{Var}\left(\hat{\theta}_m\right)$
3. Combiner les résultats : $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$
   $T = \frac{1}{M}\sum_m \widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_m\left(\hat{\beta}_m - \hat{\beta}\right)^2$

$\Rightarrow$ Objectif : fournir une estimation des paramètres et de leur variabilité (prendre en compte la variabilité due aux données manquantes)

Imputation multiple propre

1. Créer des jeux de données bootstrap (autre possibilité régression Bayésienne)
2. Estimer sur chaque jeu de données les paramètres du modèle : $(\hat{\beta})^1, ..., (\hat{\beta})^M \Longrightarrow$ variabilité sur le modèle
3. Ajouter du bruit en imputant pour $m = 1, ..., M$ valeurs manquantes y_i^m en tirant dans la distribution prédictive $\mathcal{N}(x_i\hat{\beta}^m, (\hat{\sigma}^2)^m)$

2 sources de variabilité : dans les paramètres du modèle & dans le bruit ajouté
Variance de prédiction = variance d'estimation + bruit

Modèle joint

$\Rightarrow$ Hypothèse $\mathbf{x}_{i.} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Algorithme :

1. Bootstrap des lignes : $\mathbf{X}^1$, ... , $\mathbf{X}^M$
   Algorithme EM : $(\hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\Sigma}}^1)$, ... , $(\hat{\boldsymbol{\mu}}^M, \hat{\boldsymbol{\Sigma}}^M)$
2. Imputation : x_{ij}^m tirée depuis $\mathcal{N}\left(\hat{\boldsymbol{\mu}}^m, \hat{\boldsymbol{\Sigma}}^m\right)$

Facile à paralléliser

Implémenté dans `Amelia` (website)

Amelia Earhart

James Honaker

Gary King

Matt Blackwell

Modèle conditionnel

$\Rightarrow$ Hypothèse : un modèle par variable

Algorithme :

1. Imputation initiale : imputation par la moyenne
2. Pour la variable j
   - 2.1 $(\beta^{-j}, \sigma^{-j})$ tirés d'une distribution Bootstrap ou a posteriori
   - 2.2 Imputation : régression aléatoire x_{ij} tiré dans $\mathcal{N}\left(\mathbf{X}_{-j}\beta^{-j}, \sigma^{-j}\right)$
3. Boucler sur les variables
4. Répéter M fois les étapes 2 et 3

Implémenté dans `mice` (website)

"*There is no clear-cut method for determining whether the MICE algorithm has converged*"

Stef van Buuren

Modèle joint versus modèle conditionnel

$\Rightarrow$ Modèle conditionnel prend le leadership ?

- Flexible : un modèle par variable. Facile de gérer les interactions et les variables de natures différentes (binaire, ordinale, quali...)
- Beaucoup de modèles statistiques sont des modèles conditionnels !
- Fonctionne bien en pratique

$\Rightarrow$ Inconvénients : 1 modèle/variable... fastidieux...

Modèle joint versus modèle conditionnel

$\Rightarrow$ Modèle conditionnel prend le leadership ?

- Flexible : un modèle par variable. Facile de gérer les interactions et les variables de natures différentes (binaire, ordinale, quali...)
- Beaucoup de modèles statistiques sont des modèles conditionnels !
- Fonctionne bien en pratique

$\Rightarrow$ Inconvénients : 1 modèle/variable... fastidieux...

$\Rightarrow$ Que faire avec fortes corrélations ou quand $n < p$?

- modèle joint régularise la covariance $\mathbf{\Sigma} + k\mathbb{I}$ (choix de k ?)
- modèle conditionnel : régression ridge ou sélection de variables $\Rightarrow$ beaucoup de paramètres de réglage ... pas facile ...

$$\begin{aligned} x_{ij} &= \tilde{x}_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ &= m_j + \sum_{s=1}^{S} d_s u_{is} v_{js} + \varepsilon_{ij} \end{aligned}$$

1. Variabilité des paramètres, M jeux possibles : $(\hat{x}_{ij})^1, ..., (\hat{x}_{ij})^M$
   Bootstrap des résidus : $\mathbf{X}^1 = \hat{\mathbf{X}} + \varepsilon^1, ..., \mathbf{X}^M = \hat{\mathbf{X}} + \varepsilon^M$
   ACP itérative : $\hat{\mathbf{X}}^1 = \mathbf{M} + \mathbf{U}^1\mathbf{D}^1\mathbf{V}^{1'}$, ..., $\hat{\mathbf{X}}^M = \mathbf{M}^M + \mathbf{U}^M\mathbf{D}^M\mathbf{V}^{M'}$
2. Bruit : pour $m = 1, ..., M$, valeurs manquantes x_{ij}^m sont imputées en choisissant depuis une distribution prédictive $\mathcal{N}(\hat{x}_{ij}^m, \hat{\sigma}^2)$

Implémenté dans `missMDA` (website)

$\Rightarrow$ Bonnes estimations des paramètres et de leur variance à partir d'un jeu incomplet (coverage proche de 0.95)
La variabilité due aux données manquantes est bien prise en compte

`Amelia` & `mice` ont des difficultés avec les fortes corrélations et $n < p$
`missMDA` nécessite un paramètre de réglage : nombre de dim.

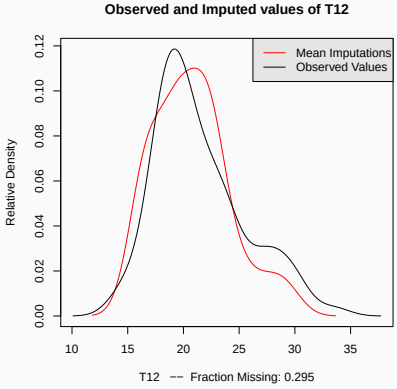
`Amelia` & `missMDA` sont basés sur les liaisons linéaires
`mice` est plus flexible (un modèle par variable)

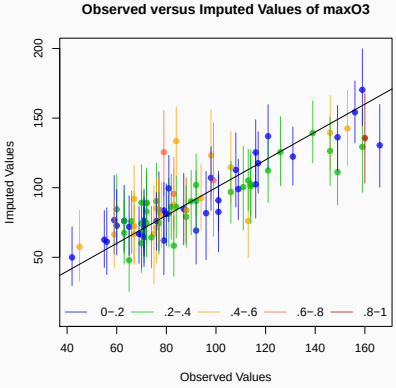
Imputation multiple en pratique

$\Rightarrow$ Etape 1 : Générer M jeux de données imputés

```
> library(Amelia)
> res.amelia <- amelia(don,m=100)  ##  avec package zelig

> library(mice)
> res.mice <- mice(don,m=100,defaultMethod="norm.boot")

> library(missMDA)
> res.MIPCA <- MIPCA(don,ncp=2,nboot=100)
> res.MIPCA$resMI
```

Etape 2 : visualisation



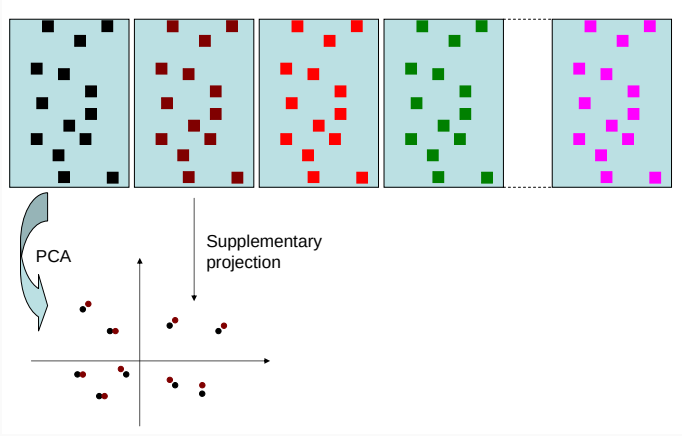



```
> library(Amelia)
> res.amelia <- amelia(don,m=100)
> compare.density(res.amelia, var="T12")
> overimpute(res.amelia, var="maxO3")
```

fonction `stripplot` dans `mice`

Etape 2 : visualisation de l'incertitude liée aux NA

Quelle confiance accorder aux représentations ? Notion de variance

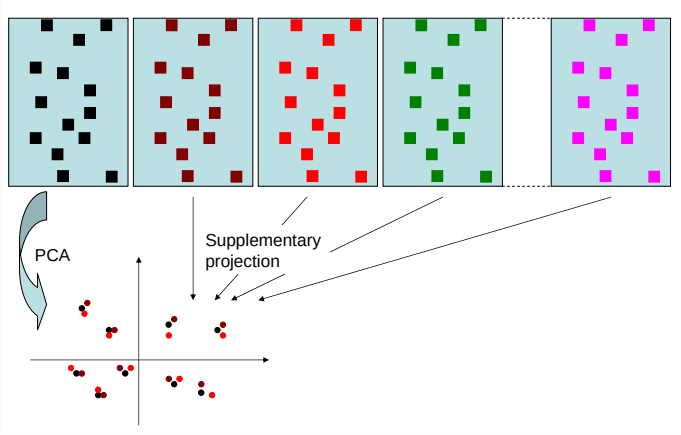


ACP itérative régularisée

$\Rightarrow$ configuration de référence

Etape 2 : visualisation de l'incertitude liée aux NA

Quelle confiance accorder aux représentations ? Notion de variance

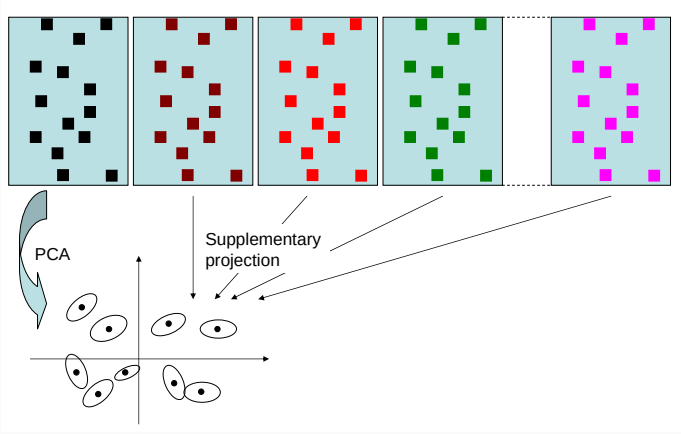


ACP itérative régularisée

$\Rightarrow$ configuration de référence

Etape 2 : visualisation de l'incertitude liée aux NA

Quelle confiance accorder aux représentations ? Notion de variance

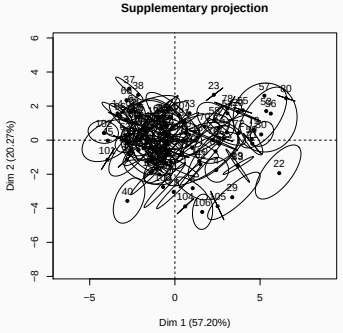


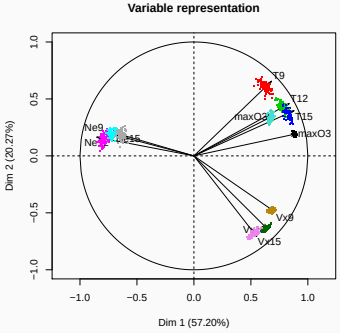
ACP itérative régularisée

$\Rightarrow$ configuration de référence

⇒ Etape 2 : visualisation de l'incertitude liée aux NA

```
> res.MIPCA <- MIPCA(don,ncp=2)
> plot(res.MIPCA,choice= "ind.supp"); plot(res.MIPCA,choice= "var ")
```





$\Rightarrow$ Etape 3. Régression par tableau et combinaison des résultats

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

```
> require(mice)
> imp<-prelim(res.mi=res.MIPCA,X=ozone[,1:11])
> fit <- with(data=imp,exp=lm(maxO3~T9+T12+T15+Ne9+Ne12+...+Vx15+maxO3v))
> res.pool<-pool(fit)
> summary(res.pool)
```

	est	se	t	df	Pr(>|t|)	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	19.31	16.30	1.18	50.48	0.24	-13.43	52.05	NA	0.46	0.44
T9	-0.88	2.25	-0.39	26.43	0.70	-5.50	3.75	37	0.71	0.69
T12	3.29	2.38	1.38	27.54	0.18	-1.59	8.18	33	0.70	0.68
....										
Vx15	0.23	1.33	0.17	39.00	0.87	-2.47	2.93	21	0.57	0.55
maxO3v	0.36	0.10	3.65	46.03	0.00	0.16	0.56	12	0.50	0.48

Imputation multiple pour variables qualitatives

$\Rightarrow$ Modèle joint :

- Modèle log-linear (Schafer, 1997) `(cat)` : pb si bcp de modalités
- Modèles à classes latentes (Vermunt, 2014) - Bayésien non-paramétrique (Si & Reiter, 2014, Murray & Reiter, 2016) `(MixedDataImpute, NPBayesImpute, NestedCategBayesImpute)`

$\Rightarrow$ Modèle conditionnel : logistique, multinomial, forêts `(mice)`

$\Rightarrow$ `MIMCA` fournit des inférences valides (ex. régression logistique avec NA) appliquée à des jeux de données avec bcp de modalités et des modalités rares

Imputation multiple pour données mixtes : `MIFAMD` sur le même principe, et modèles joints et conditionnels

Remarque de Dempster & Rubin (1983)

*"**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."*

Remarques

Remarque de Dempster & Rubin (1983)

*"**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."*

Remarques sur l'imputation multiple

- Théorie de l'IM : bonne pour la régression. Autres méthodes ?
- Modèle d'imputation doit être aussi complexe que le modèle d'analyse (interaction)

Remarques

Remarque de Dempster & Rubin (1983)

*"**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."*

Remarques sur l'imputation multiple

- Théorie de l'IM : bonne pour la régression. Autres méthodes ?
- Modèle d'imputation doit être aussi complexe que le modèle d'analyse (interaction)

Quelques problèmes pratiques encore ouverts : la recherche n'est pas finie !

- Imputation de X et X^2
- Problèmes de bornes (> 0) $\Rightarrow$ tronquer ?
- Comment faire avec des variables temporelles ?
- Comment faire avec des données de grandes dimensions ?

Une page Web et des didacticiels

http://factominer.free.fr/missMDA/index_fr.html

Quelques références supplémentaires

Schafer (1997)

Joseph L. Schafer

Little & Rubin (1987, 2002)

Roderick Little

Donald Rubin

Van Buuren (2012)

Stef van Buuren

chap 25 de Gelman & Hill (2006)

Andrew Gelman

Jennifer L. Hill

Ressources

⇒ Logiciels :

- R CRAN task View: Missing Data
- R-miss-tastic

⇒ Articles :

- Imbert, A., & Vialaneix, N. (2018). Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes. *Journal de la SFdS*, **159(2)**, 1-55.
- Josse J, Husson F. & Pagès J (2009) Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la SFdS*. **150 (2)**, 28-51.