

# Gestion des données manquantes en/par analyse factorielle

---

F. Husson

<https://husson.github.io/>

UP de mathématiques appliquées - l'institut Agro

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



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 détruites (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

- 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

The diagram illustrates a data matrix with 'Individus' (Individuals) on the vertical axis and 'Variables' on the horizontal axis. The vertical axis is labeled with '1', 'i', and 'n'. The horizontal axis is labeled with '1', 'j', and 'p'. The matrix itself is a light blue rectangle containing several black question marks ('?') representing missing data points. The labels 'Individus' and 'Variables' are placed to the left and above the matrix, respectively.

	Variables		
	1	j	p
1	?		
	?		
		?	
			?
	?		
		?	
i			
		?	
			?
	?		
		?	
			?
n			?

# Exemple sur des données ozone

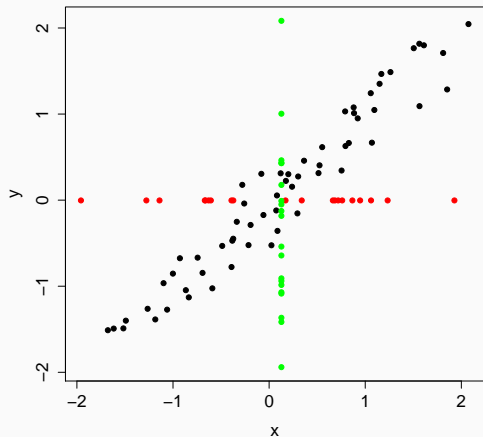
Code disponible : <http://factominer.free.fr/missMDA/ozone.R>

```
> 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)
- Imputation par la moyenne (option par défaut dans de nombreux logiciels)



Distorsion très importante des liaisons  
entre variables

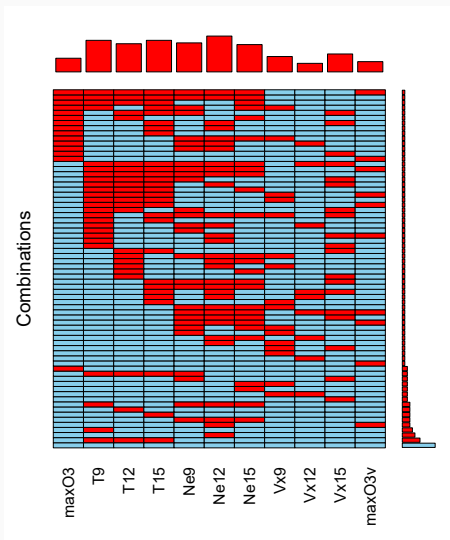
Traitement des données manquantes dépend du :

- 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)

⇒ Visualisation des données manquantes

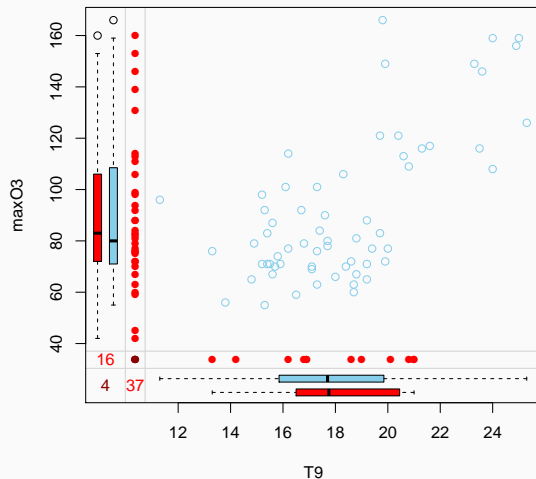
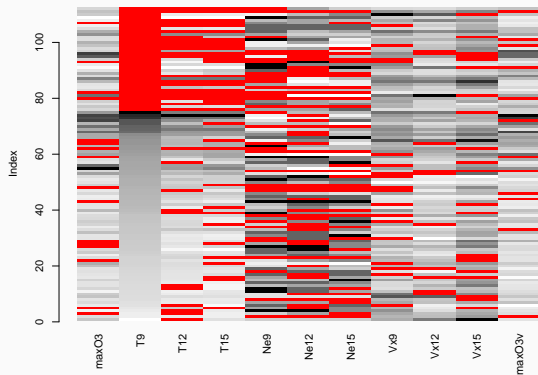
# Visualisation du dispositif de données manquantes



```
> 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")])
```

# 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

# Approche par maximum de vraisemblance : 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
```

## Variances : approche bootstrap

- Bootstrap les lignes :  $\mathbf{X}^1, \dots, \mathbf{X}^B$
- Algorithme EM :  $(\hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\Sigma}}^1), \dots, (\hat{\boldsymbol{\mu}}^B, \hat{\boldsymbol{\Sigma}}^B)$

**Problème** : développer une méthode spécifique pour chaque méthode statistique

- 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

⇒ Hypothèse  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

## Cas bivarié avec données manquantes sur $Y$ (régression aléatoire)

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

## Extension au cas multivarié

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

```
> 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)
```

⇒ 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](#)

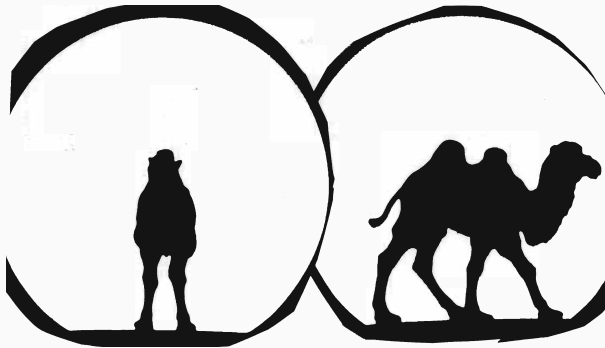
⇒ [R-miss-tastic](#)

⇒ Imputation par ACP



# Ajustement du nuage en ACP

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



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

⇒ Meilleure approximation par projection

⇒ Meilleure représentation de la diversité, de la variabilité

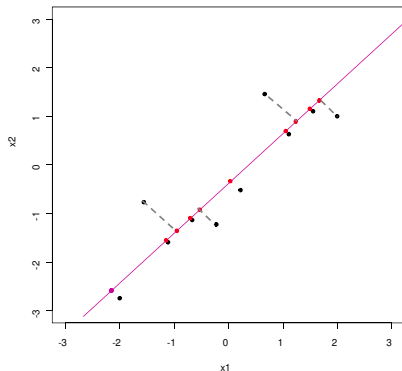
# Ajustement du nuage en ACP

**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

-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

**$\hat{X}$**



**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

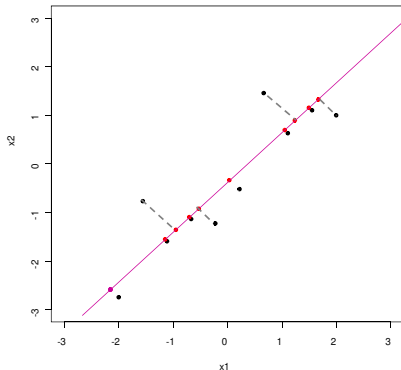
**V'**

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



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

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

## ACP : cas complet

⇒ Point de vue géométrique : minimiser l'erreur de reconstitution

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

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

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

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

⇒ 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

⇒ ACP : moindres carrés

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

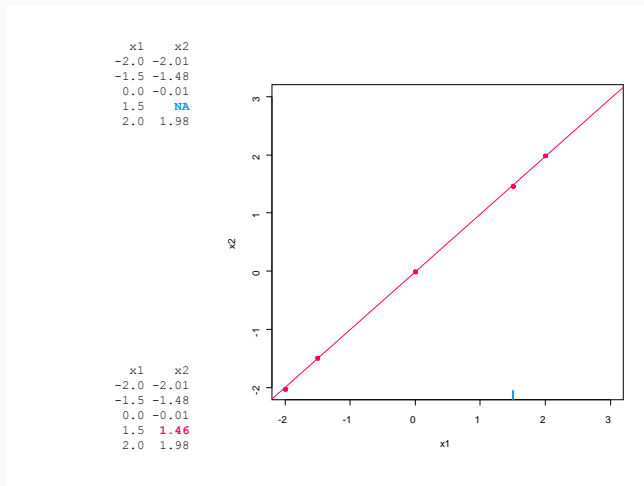
⇒ ACP avec données manquantes : moindres carrés pondérés

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

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



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

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

- ① initialisation  $\ell = 0$  :  $\mathbf{X}^0$  (imputation par la moyenne)
- ② step  $\ell$  :
  - (a) ACP sur le tableau complété  $\rightarrow (\mathbf{U}^\ell, \mathbf{D}^\ell, \mathbf{V}^\ell)$ ;  
**S dimensions conservées**
  - (b) valeurs manquantes imputées par  $\hat{\mathbf{X}}^\ell = \mathbf{M}^\ell + \mathbf{U}^\ell \mathbf{D}^\ell \mathbf{V}^{\ell'}$  ;  
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  $\mathbf{M} + \mathbf{UDV}'$ )

- Résultats de l'ACP obtenus à partir des données observées uniquement : graphe des individus et graphe des variables

⇒ On "saute" les données manquantes, l'ACP itérative minimise

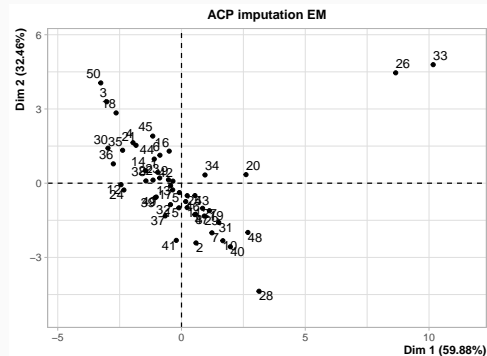
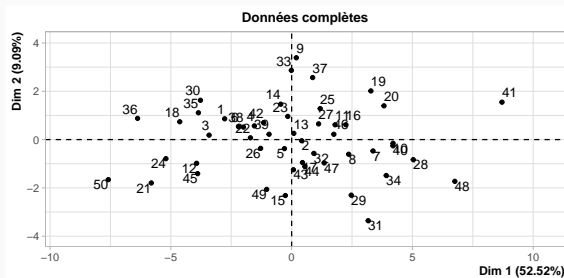
$$\| \mathbf{R} * (\mathbf{X} - (\mathbf{M} + \mathbf{UDV}')) \|^2$$

- 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

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



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

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

⇒ 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 ⇒ ACP itérative régularisée

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

⇒ 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 :

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

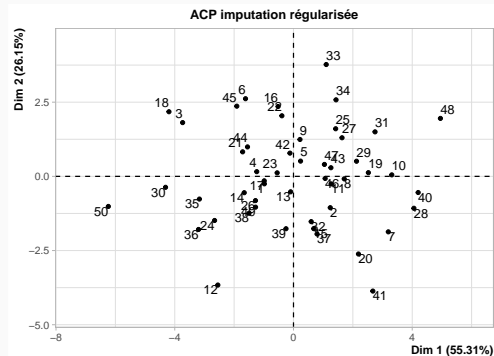
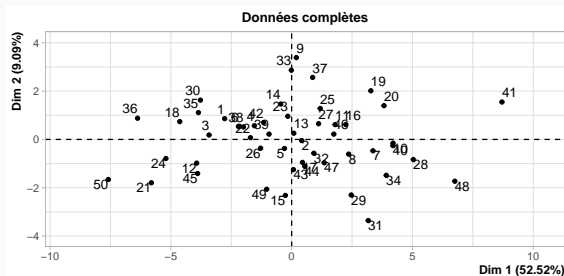
$$\hat{\sigma}^2 = \frac{RSS}{ddl} = \frac{n \sum_{s=S+1}^p d_s^2}{(n-1-S)(p-S)}$$

$\sigma^2$  petit → ACP régularisée  $\approx$  ACP

$\sigma^2$  grand → imputation par la moyenne

# Surajustement

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



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

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

## 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

## Quid des éléments supplémentaires ?

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

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

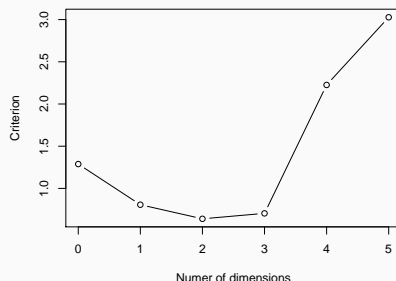
## Tutoriel sur l'ACP avec données manquantes

(données ozone, lignes de code)

⇒ 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")
```



⇒ Etape 2 : Imputation des données manquantes

```
> res.comp <- imputePCA(don, ncp = 2)
```

```
> res.comp$completeObs[1:3,]
```

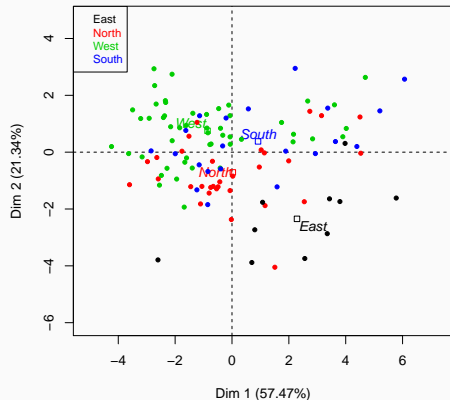
	max03	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	max03v
0601	87	15.60	18.50	20.47	4	4.00	8.00	0.69	-1.71	-0.69	84
0602	82	18.51	20.88	21.81	5	5.00	7.00	-4.33	-4.00	-3.00	87
0603	92	15.30	17.60	19.50	2	3.98	3.81	2.95	1.97	0.52	82



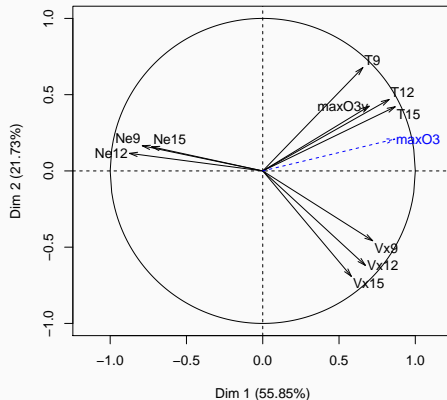
# ACP sur le tableau complété

⇒ Etape 3 : ACP sur le tableau complété

Individuals factor map (PCA)



Variables factor map (PCA)



```
> 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)
```

```
> Factshiny(ozone)
```

Graphes
Valeurs
Description automatique des axes
Résumé du jeu de données
Données

☒ Paramètres de l'ACP

**Choix de variables quantitatives supplémentaires**

**Choix de variables qualitatives supplémentaires**

vent pluie

**Choix d'individus supplémentaires**

☒ Réduire les variables

**Gestion des données manquantes**

☐ Imputer par la moyenne (rapide - non recommandé)

☒ Imputer par un modèle d'ACP à 2 dimensions (bon compromis)

☐ Imputer par un modèle d'ACP à k dimensions (estime k, chronophage)

Soumettre

**Graphes des individus de l'ACP**

**Graphes des variables de l'ACP**

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)

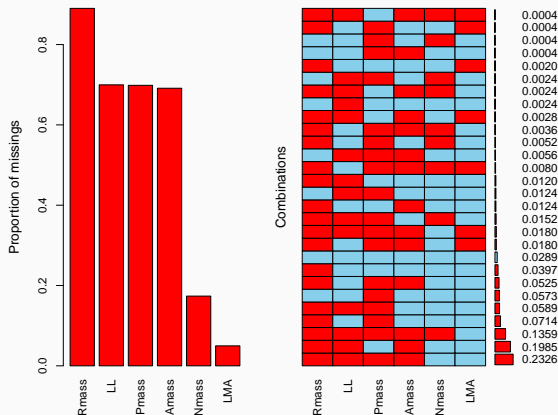
et 1 variable qualitative : le biome (macro-écosystème)

Wright IJ, et al. (2004). The worldwide leaf economics spectrum. *Nature*, 428 :821.

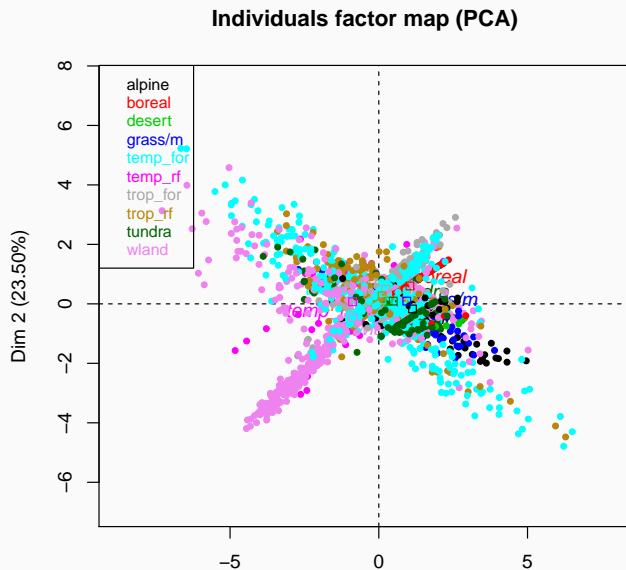
[www.nature.com/nature/journal/v428/n6985/extref/nature02403-s2.xls](http://www.nature.com/nature/journal/v428/n6985/extref/nature02403-s2.xls)

# 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)
```

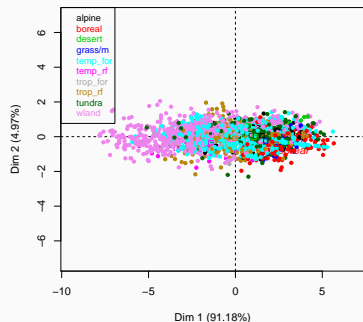


Quid de l'imputation par la moyenne ?

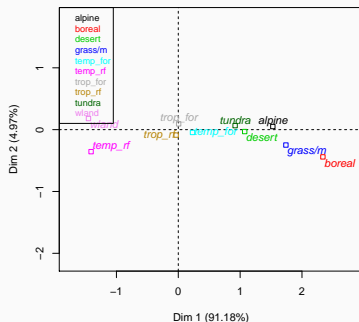


# Jeu de données en écologie

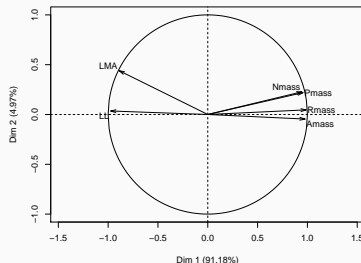
Individuals factor map (PCA)



Individuals factor map (PCA)



Variables factor map (PCA)



```
> 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)
```

- 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)*



# Rappels d'ACM

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

$T =$

1	0	0	1	0	0	1	...	0	1	$p$
1	0	0	1	0	1	0	...	NA	NA	$p$
NA	NA	NA	0	1	0	0	...	0	1	
1	0	0	1	0	0	1	...	0	1	
$t_{ik}$										$p$
0	0	1	NA	NA	0	1	...	0	1	
1	0	0	1	0	0	1	...	0	1	$p$
$n_1$	$n_2$	$n_3$	.....						$n_q$	$np$

$$P_{\Sigma} = \begin{matrix} & n_1 & n_2 & n_3 & & & 0 \\ & & & & \dots & & \\ & & & & & \dots & \\ 0 & & & & & & \dots \\ & & & & & & \dots \\ & & & & & & n_q \end{matrix}$$

## L'ACM comme une ACP pondérée

ACM vue comme l'ACP du triplet

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

# Traitement d'un questionnaire avec *missing single*

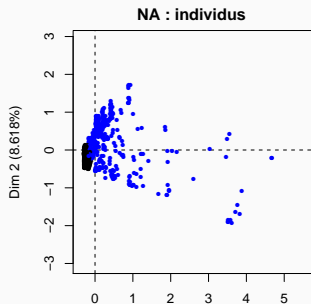
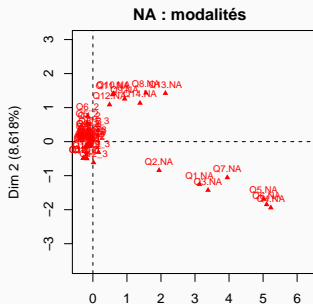
## Les données

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

## Création de nouvelles modalités

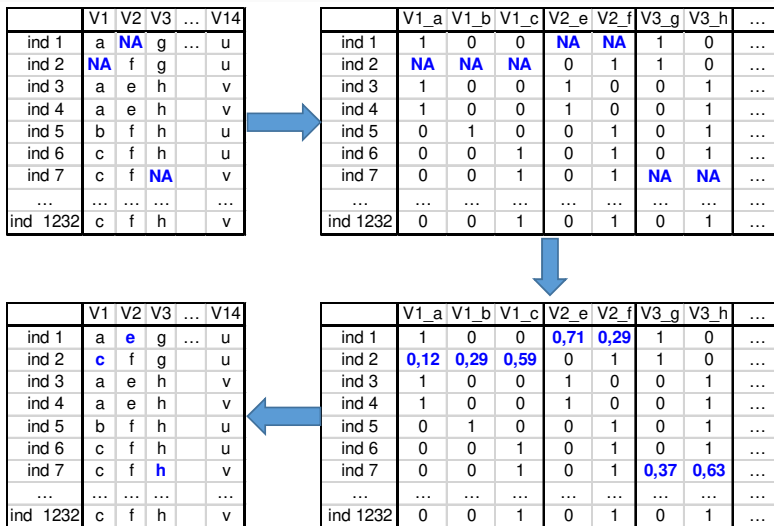
Création d'une modalité NA pour chaque variable ayant au moins une valeur manquante

	V1	V2	V3		V1_a	V1_b	V1_c	V1_NA	V2_e	V2_f	V2_NA	V3_g	V3_h
ind 1	a	NA	g		1	0	0	0	0	0	1	1	0
ind 2	NA	f	g		0	0	0	1	0	1	0	1	0
ind 3	a	e	h		1	0	0	0	1	0	0	0	1
ind 4	a	e	h		1	0	0	0	1	0	0	0	1
ind 5	b	f	h		0	1	0	0	0	1	0	0	1
ind 6	c	f	h		0	0	1	0	0	1	0	0	1
ind 7	c	f	h		0	0	1	0	0	1	0	0	1



- 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)



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

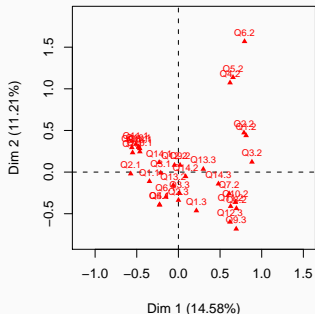
## Imputation du tableau disjonctif

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

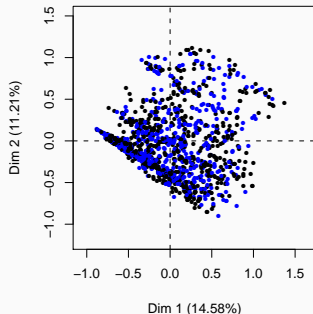
## ACM sur le tableau complété (utilisation de l'argument tab.disj)

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

ACM itérative régularisée : modalités



ACM itérative régularisée : individus



- 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

- General location model (Schafer, 1997)  $\implies$  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*)

$\implies$  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

⇒ Propriétés :

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

⇒ Meilleur que plus proches voisins et mice



# Analyse Factorielle de Données Mixtes (cas complet)

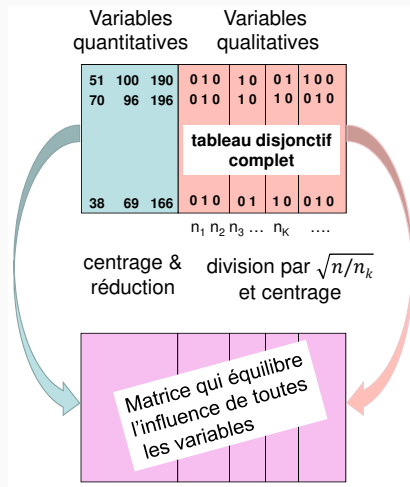
AFDM (Escofier, 1979), PCAMIX (Kiers, 1991)

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

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

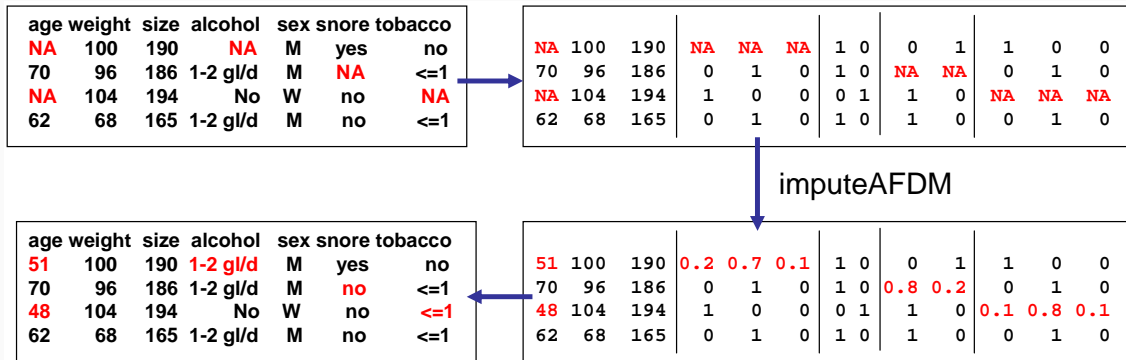
- Les composantes principales  $\mathbf{F}_s$  maximisent :

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



# 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

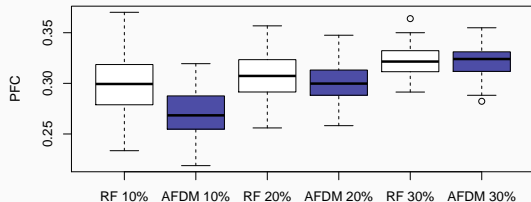
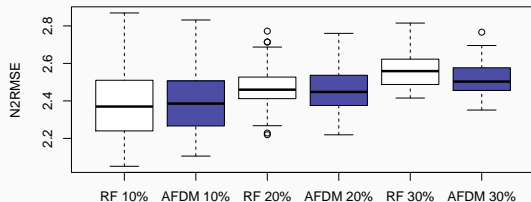


Les valeurs imputées peuvent être vues comme des degrés d'appartenance

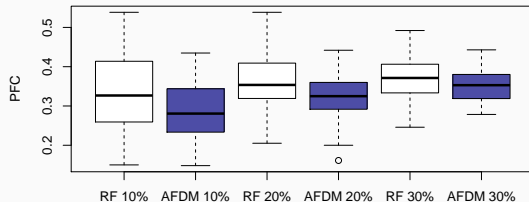
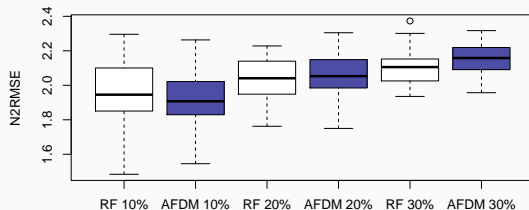
# Comparaison avec forêts aléatoires

Imputations obtenues par forêts aléatoires & ACP itérative

**GBSG2**



**Ozone**



# 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
```

## 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

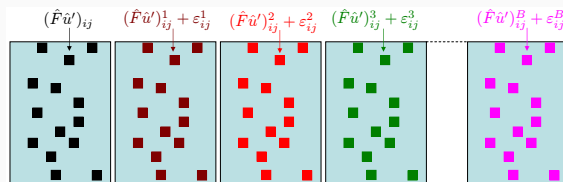
- basée sur la reconstitution de l'ACP (axes et composantes)
- prise en compte des liaisons entre var. quantitatives et qualitatives
- bonne alternative aux méthodes d'imputation (forêts aléatoires, etc.) si liaisons linéaires, pour les variables qualitatives (notamment les modalités rares)

- 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 :  $\hat{\theta}_m, \widehat{Var}(\hat{\theta}_m)$

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

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

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

## Imputation multiple propre

- ❶ Créer des jeux de données bootstrap (autre possibilité régression Bayésienne)
- ❷ Estimer sur chaque jeu de données les paramètres du modèle :  $(\hat{\beta})^1, \dots, (\hat{\beta})^M \implies$  variabilité sur le modèle
- ❸ Ajouter du bruit en imputant pour  $m = 1, \dots, M$  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

⇒ Hypothèse  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Algorithme :

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

Facile à paralléliser

Implémenté dans **Amelia** ([website](#))



Amelia Earhart



James Honaker



Gary King



Matt Blackwell

# Modèle conditionnel

⇒ 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  $\mathcal{N}(\mathbf{X}_{-j}\beta^{-j}, \sigma^{-j})$
- 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

⇒ 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

⇒ Inconvénients : 1 modèle/variable... fastidieux...

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

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

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

- ❶ Variabilité des paramètres,  $M$  jeux possibles :  $(\hat{x}_{ij})^1, \dots, (\hat{x}_{ij})^M$   
Bootstrap des résidus :  $\mathbf{X}^1 = \hat{\mathbf{X}} + \varepsilon^1, \dots, \mathbf{X}^M = \hat{\mathbf{X}} + \varepsilon^M$   
ACP itérative :  $\hat{\mathbf{X}}^1 = \mathbf{M} + \mathbf{U}^1 \mathbf{D}^1 \mathbf{V}^{1'}$ , ...,  $\hat{\mathbf{X}}^M = \mathbf{M}^M + \mathbf{U}^M \mathbf{D}^M \mathbf{V}^{M'}$
- ❷ Bruit : pour  $m = 1, \dots, M$ , 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](#))

## Modèle joint, modèle conditionnel et ACP

⇒ 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)

⇒ Etape 1 : Générer  $M$  jeux de données imputés

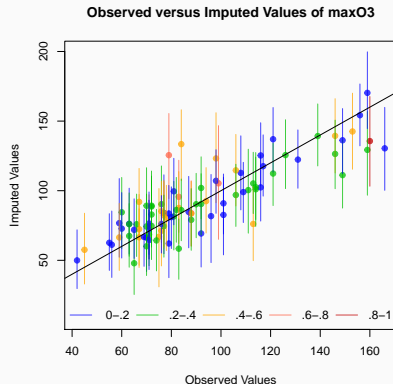
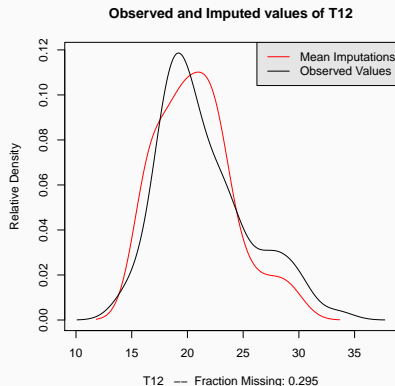
```
> library(Amelia)
> res.amelia <- amelia(don,m=100) ## avec package zelig

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

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

# Imputation multiple en pratique

## Etape 2 : visualisation



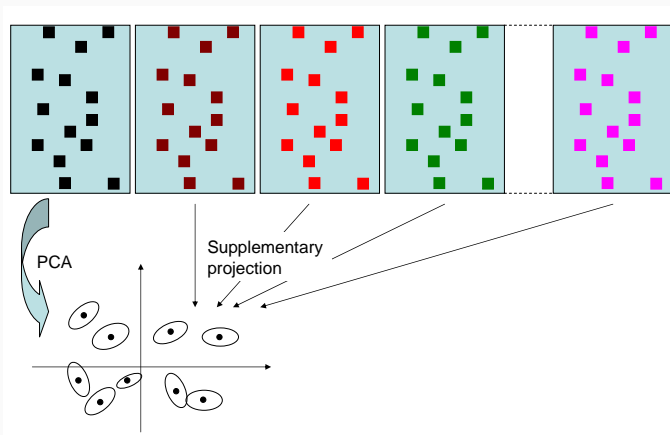
```
> library(Amelia)
> res.amelia <- amelia(don,m=100)
> compare.density(res.amelia, var="T12")
> overimpute(res.amelia, var="maxO3")
```

fonction stripplot dans mice

# Imputation multiple en pratique

Etape 2 : visualisation de l'incertitude liée aux NA

Quelle confiance accorder aux représentations ? Notion de variance



ACP itérative régularisée

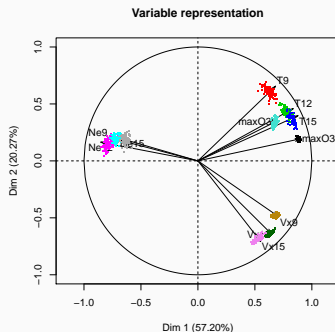
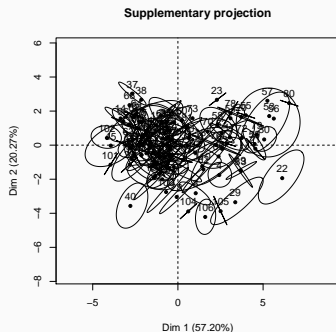
⇒ configuration de référence



# Imputation multiple en pratique

⇒ 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 ")
```



# Imputation multiple en pratique

⇒ Etape 3. Régression par tableau et combinaison des résultats

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^M \hat{\beta}_m$$

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

```
> 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 et mixtes

⇒ 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**)

⇒ Modèle conditionnel : logistique, multinomial, forêts (**mice**)

⇒ **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 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 ?

[http://factominer.free.fr/missMDA/index\\_fr.html](http://factominer.free.fr/missMDA/index_fr.html)



## > Le package missMDA

Le package **missMDA** est complémentaire de FactoMineR. Il permet de gérer les données manquantes pour les méthodes d'analyses factorielles (ACP, AFC, ACM, AFDM, AFM). Il permet de faire de l'imputation simple et multiple.

L'imputation simple consiste à remplacer les valeurs manquantes par des valeurs plausibles. Cela revient à compléter le jeu de données qui peut ensuite être analysé par n'importe quelle méthode d'analyse factorielle.

**missMDA** impute les valeurs manquantes de sorte que les valeurs imputées n'ont aucune influence sur les résultats de l'analyse factorielle (pas d'influence dans le sens où les valeurs imputées n'ont aucun poids, et donc les résultats de l'analyse factorielle sont obtenus uniquement avec les valeurs observées).

**missMDA** utilise des méthodes de réduction de données, ce qui lui permet d'imputer de façon satisfaisante de gros jeux de données contenant des variables quantitatives et/ou qualitatives. En effet, il impute par ACP (ou ACM, ou AFDM ou AFM) en prenant en compte à la fois les similarités entre individus et les liens entre variables.

Voir cette vidéo si vous voulez comprendre le principe de **missMDA** quelque soit les jeux de données (quantitatifs et/ou qualitatifs).

Les imputations sont très bonnes comparées aux méthodes classiques permettant d'imputer des tableaux incomplets (forêts aléatoires par exemple).

- **missMDA** gère les données manquantes dans:
  - les jeux de données avec variables quantitatives grâce à l'ACP (Voir la vidéo)
  - les jeux de données avec variables qualitatives grâce à l'ACM (Voir la vidéo)
  - les tableaux de contingence grâce à l'AFC
  - les données mixtes grâce à l'AFDM
  - les jeux de données où les variables sont structurées par groupe grâce à l'AFM
- **missMDA** permet de faire de l'imputation multiple:
  - pour les variables quantitatives grâce à l'ACP: Voir la vidéo
  - pour les variables qualitatives grâce à l'ACM

## > Menu sur les données manquantes

### Le package missMDA

ACP avec données manquantes

ACM avec données manquantes

Imputation multiple

Peut-on croire dans les valeurs imputées ?

Références - Conférences

## > Les auteurs de missMDA

François Husson

Julie Josse

⇒ Logiciels :

- [R CRAN task View: Missing Data](#)
- [R-miss-tastic](#)

⇒ Articles :

- Gégout-Petit, A., Maumy-Bertrand, M., Saporta, G., & Thomas-Agnan, C. (2022). *Données manquantes*. Edition Technip.
- 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.