



# Données Manquantes que peut-on faire ?

Hadrien Lorenzo

DU BDSI 18 et 20/05/2022

- 1 Introduction
- 2 De la régression à l'imputation
- 3 Imputation et vraies données
- 4 Conclusion

# Section 1

## Introduction

## Analyse cas complet

On dit qu'une donnée est manquante lorsque l'observation d'une variable pour une observation n'est pas accessible.

On peut choisir de retirer l'observation associée aux NA mais peu indiqué si :

- petit échantillon ( $n$  faible),
- forte proportion de NA.

On peut aussi retirer les variables associées aux NA si :

- la structure des données le permet ( $p$  important)
- peu de variables sont atteintes par des NA

Le jeu de données résultant est appelé **cas complet**, en pratique il n'est jamais conseillé d'utiliser cette approche.

# Solutions restantes

Il reste de remplir les cases, c'est l'**imputation**.

- Estimer les NA avec des valeurs fixées (**Imputation simple**) :
  - Conditionnellement à la seule variable considérée :
    - **moyenne/médiane**,
    - **Last Observation Carried Forward (LOCF)**,
  - Sur l'ensemble des variables :
    - k-plus proches voisins (**kNN** via *FNN*),
    - forêts aléatoires (**missForest** (Stekhoven & Buehlmann, 2012), itératif),
    - SVD (**missMDA** (Josse & Husson, 2016), itératif),
    - ...
- **Imputation multiple** :
  - Modèles conditionnels (**mice** (van Buuren & Groothuis-Oudshoorn, 2011) )
  - SVD (**missMDA**)
  - ...

# Une classification des processus de perte de données

Due à Little & Rubin (1976).

- MCAR** (*Missing Completely At Random*), si la probabilité de perte de l'information est la même pour toutes les observations et donc ne dépend ni des données observées ni des autres données manquantes.
- MAR** (*Missing At Random*), si la probabilité que la donnée soit manquante est associée à une ou plusieurs autres variables, observées.
- MNAR** (*Missing Not At Random*), si la probabilité de ne pas observer cette valeur dépend de cette valeur et ou dépend de variables non observées.

# Exemples

- MCAR** Un capteur qui s'éteint et se rallume sans raison. Un patient malade qui se rend à l'hôpital seulement lorsqu'il reçoit du courrier (dur à prédire/observer donc...)
- MAR** Un végétarien qui ne donne pas son avis (gustatif) sur un morceau de viande qui lui est présenté. Un nourrisson qui ne renseigne pas la couleur du cheval blanc d'Henri IV.
- MNAR** Qui gagne bien sa vie répond moins facilement à toute question sur son salaire.

# Commentaires

**MCAR** Le plus "simple" à gérer, mais peu réaliste.

**MAR** Peut aussi être géré si approche multivariée.

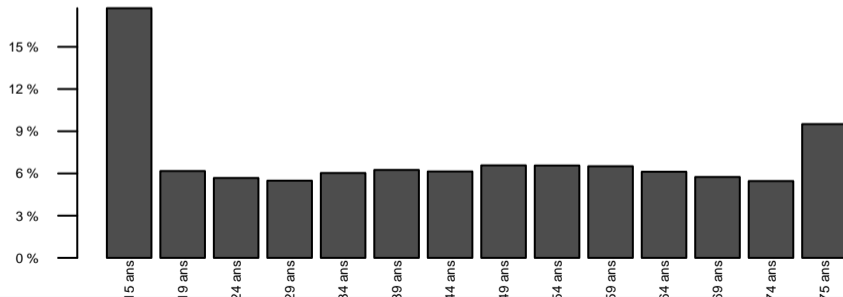
**MNAR** Difficile à gérer sans information supplémentaire (modèles, variables,...)



# Exemple, l'âge de la population française selon l'Insee

On peut télécharger ces données sur le site de l'[Insee](#)<sup>2</sup>.

```
insee <- read.csv("files/insee.csv", sep=";")[-15,1:4]
prop <- insee$Ensemble/sum(insee$Ensemble)*100
barplot(t(prop), yaxt="n", xaxt="n")
axis(1, line = -1, tick = F, at = (1:nrow(insee))*1.2-1/2, insee$Groupe.d.âges, las=2, cex.axis=0.4)
axis(2, at = seq(0,20,by = 3), paste(seq(0,20,by = 3), "%"), las=2, cex.axis=0.4)
```



## Exemple, l'âge de la population française selon l'Insee (2)

### Exercice

Vous devez générer un jeu de données de taille  $n = 1000$  touché par un des processus de données manquantes proposés :

**Sc<sub>1</sub>** "10% des gens refusent de donner leur âge."

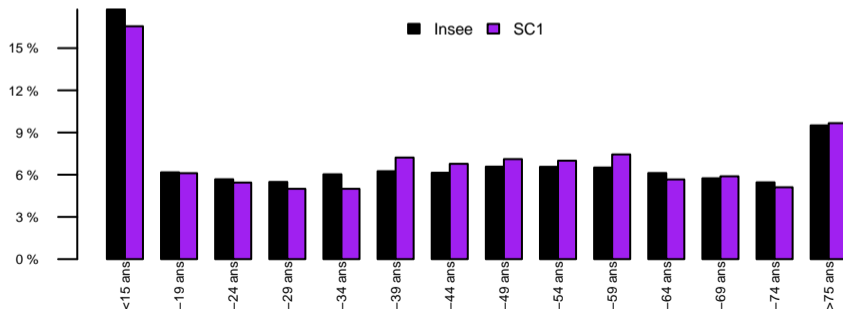
**Sc<sub>2</sub>** "Les moins de 5 ans ne parviennent pas à renseigner leur âge."

**Sc<sub>3</sub>** "C'est vendredi, les étudiants ne sont pas là!"

Vous discuterez des hypothèses prises et des résultats obtenus.

# Exemple, l'âge de la population française selon l'Insee, $Sc_1$

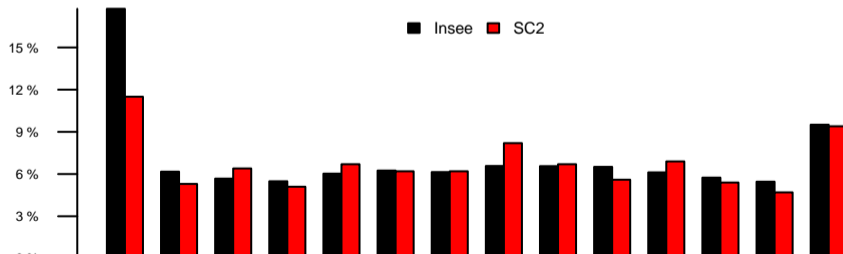
```
propNA <- 0.9
sampSC1 <- sample(x = 1:length(prop), size = propNA*n_pop, prob = prop, replace = T)
propSC1 <- table(sampSC1)/(propNA*n_pop)*100
barplot(t(cbind(prop, propSC1)), beside = T, col=c(1, "purple"), yaxt="n", xaxt="n")
axis(1, line = -1, tick = F, at = (1:nrow(insee))*3-1, insee$Groupe.d.âges, las=2, cex.axis=0.4)
axis(2, at = seq(0,20,by = 3), paste(seq(0,20,by = 3), "%"), las=2, cex.axis=0.4)
legend("top", c("Insee", "SC1"), fill=c(1, "purple"), ncol=2, bty="n", cex=1/2)
```



# Exemple, l'âge de la population française selon l'Insee, $Sc_2$

Un tiers des individus de classe 1 est à retirer du jeu de données (hyp 1)

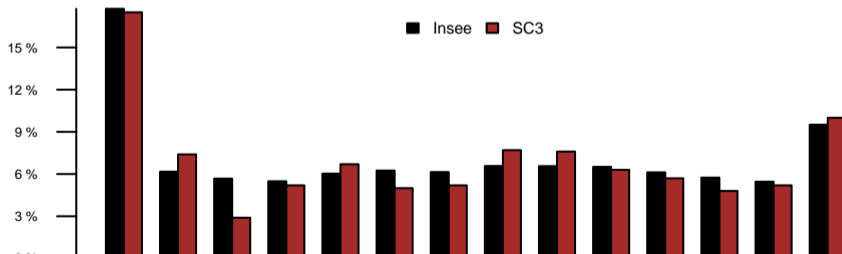
```
propSC2 <- sample(x = 1:length(prop),size = n_pop,prob = prop,replace = T)
class1 <- which(propSC2==1);class1_no5 <- class1[1:(length(class1)/3)];propSC2[class1_no5] <- NA
propSC2 <- table(propSC2)/n_pop*100
barplot(t(cbind(prop,propSC2)),beside = T,col=c(1,"red"),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*3-1,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
legend("top",c("Insee", "SC2"),fill=c(1,"red"),ncol=2,bty="n",cex=1/2)
```



# Exemple, l'âge de la population française selon l'Insee, $Sc_3$

La moitié des individus de classe **3** est à retirer du jeu de données (hyp 2)

```
propSC3 <- sample(x = 1:length(prop),size = n_pop,prob = prop,replace = T)
class1 <- which(propSC3==3);class1_stu <- class1[1:(length(class1)/2)];propSC3[class1_stu] <- NA
propSC3 <- table(propSC3)/n_pop*100
barplot(t(cbind(prop,propSC3)),beside = T,col=c(1,"brown"),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*3-1,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
legend("top",c("Insee", "SC3"),fill=c(1,"brown"),ncol=2,bty="n",cex=1/2)
```



## Exemple, l'âge de la population française selon l'Insee, commentaires

- L'analyse est robuste à l'hypothèse MCAR  $\mathbf{Sc}_1$
- $\mathbf{Sc}_2$  peut être facilement géré
- $\mathbf{Sc}_3$  est plus difficile à gérer

## Section 2

# De la régression à l'imputation

# Modèle linéaire simple à une covariable

Soit le modèle de simulation

$$y = x + \epsilon$$

où  $x \sim \mathcal{N}(0, 1)$ ,  $\epsilon \sim \mathcal{N}(0, 1/4)$  et  $\epsilon \perp\!\!\!\perp x$ . On a accès à :

- un échantillon d'entraînement  $S = (x_i, y_i)_{i=1, \dots, n}$  de taille  $n = 50$ ,
- un échantillon de test  $\tilde{S} = (x_i)_{i=n+1, \dots, n+\tilde{n}}$  de taille  $\tilde{n} = 30$ .

En régression, on réalise deux opérations successives :

- construction d'un modèle de régression sur  $S$ , noté  $\mathcal{P}$ ,
- estimation de la réponse de  $\mathcal{P}$  à une covariable  $x_0$ , notée  $\hat{y}_0 = \mathcal{P}(x_0)$ .

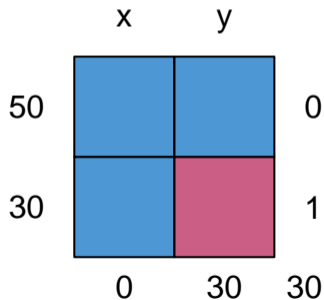


## Lien avec les données manquantes

En combinant  $S$  et  $\tilde{S}$ , on a accès à un jeu de données de dimensions  $(n + \tilde{n}) \times (2)$  avec  $\tilde{n}$  données manquantes : les valeurs  $(y_i)_{i=n+1, \dots, n+\tilde{n}}$ .

On peut observer la structure des données manquantes grâce aux commandes suivantes

```
x <- rnorm(n,0,1) ; y <- x + rnorm(n,0,sigma)
x_t <- rnorm(n_tilde,0,1) ; y_t <- rep(NA,n_tilde)
pattern_miss <- mice::md.pattern(rbind( cbind(x,y) , cbind(x_t,y_t) ))
```



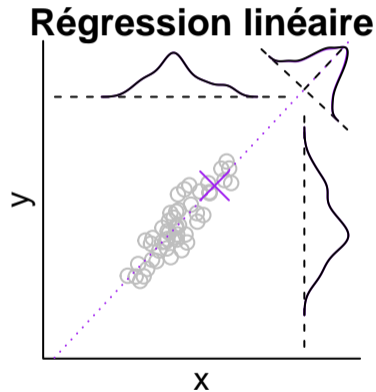
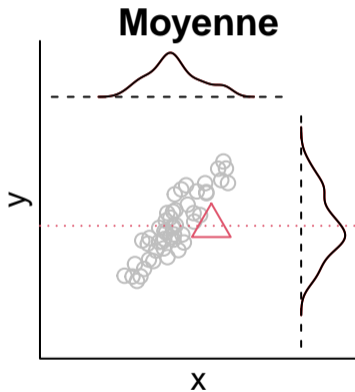
# Analyse du jeu de données simulées

Comparaison de 2 méthodologies de régression :

- à la moyenne,
- en utilisant le modèle de régression linéaire.

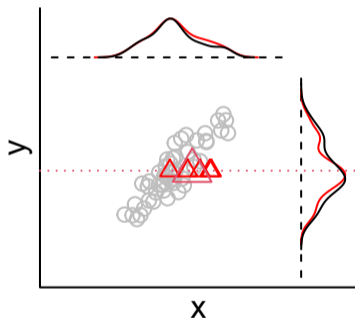
Dans chaque cas on observera attentivement le comportement des distributions après imputations.

# Après imputation d'une observation

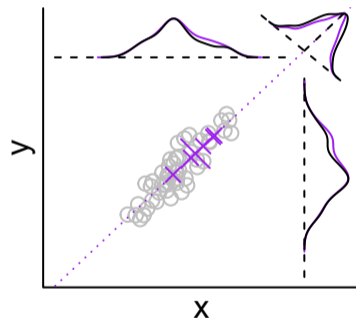


# Après imputation de 5 observations

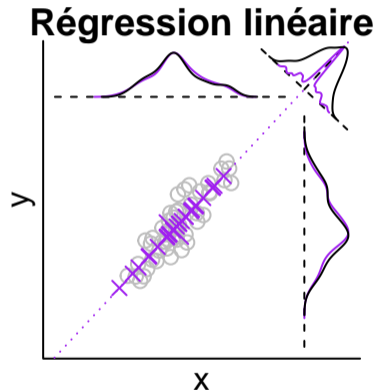
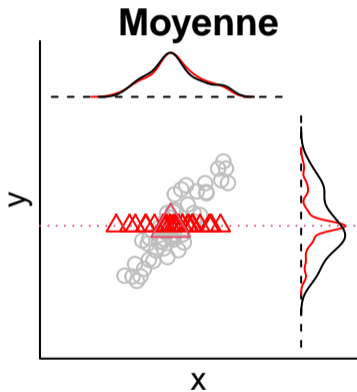
## Moyenne



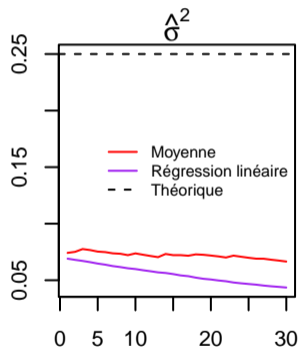
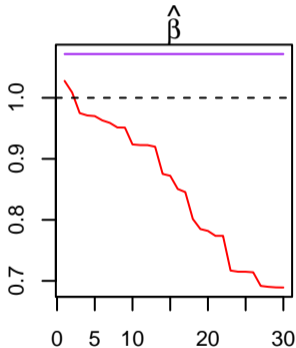
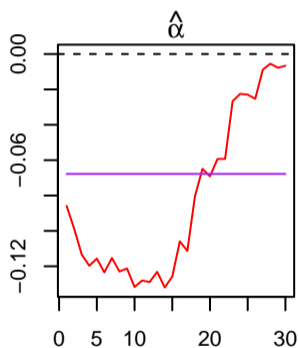
## Régression linéaire



# Après imputation de 30 observations



# En global



# Observations

Deux impératifs apparaissent :

- Conditionner l'estimation des données manquantes sur les données observées.
- Utiliser un modèle de régression adapté.
- Garder en tête que l'ensemble reconstruit doit être réutilisé : c'est la grosse différence avec la régression/classification/analyse classique.

Une solution ?

# Observations

Deux impératifs apparaissent :

- Conditionner l'estimation des données manquantes sur les données observées.
- Utiliser un modèle de régression adapté.
- Garder en tête que l'ensemble reconstruit doit être réutilisé : c'est la grosse différence avec la régression/classification/analyse classique.

## Une solution ?

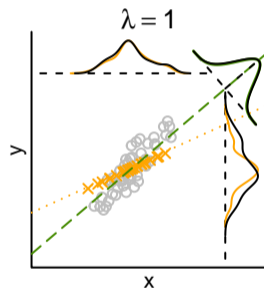
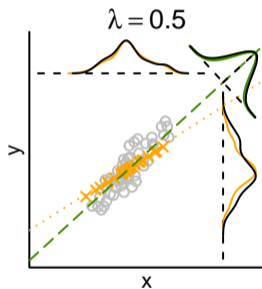
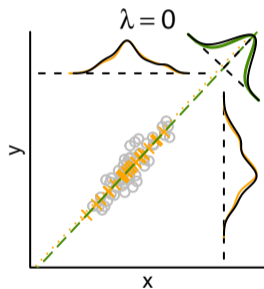
Pourquoi ne pas faire de la régularisation ?

Régularisation Ridge :  $\min_{\beta \in \mathbb{R}} (y - x\beta)^2 + \lambda\beta^2$



# La régularisation (1)

- Régression linéaire pénalisée Ridge pour l'imputation.
- Régression linéaire pour l'estimation des paramètres.



$\lambda$	0	0.5	1
$\hat{\alpha}$	0.114	0.114	0.114
$\hat{\beta}$	1.072	0.917	0.852
$\hat{\sigma}^2$	0.146	0.177	0.21

## La régularisation (2)

### Conclusion

La régularisation *sous* – *estime* la variance de  $y$ .

Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de  $y$ .

## La régularisation (2)

### Conclusion

La régularisation *sous – estime* la variance de  $y$ .

Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de  $y$ .

### Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

## La régularisation (2)

### Conclusion

La régularisation *sous – estime* la variance de  $y$ .

Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de  $y$ .

### Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

### Une solution ?

## La régularisation (2)

### Conclusion

La régularisation *sous – estime* la variance de  $y$ .

Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de  $y$ .

### Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

### Une solution ?

Générer  $M$  jeux de données imputés pour obtenir cette variabilité.  $\implies$  C'est **la régression stochastique**.

## La régression stochastique - *improper imputation*

Alors que le modèle de prédiction était précédemment

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

où  $\hat{\beta}$  était estimé sur le jeu de données  $S$ , nous allons maintenant utiliser l'estimateur

$$\tilde{y} = \hat{\alpha} + \hat{\beta}x + \eta,$$

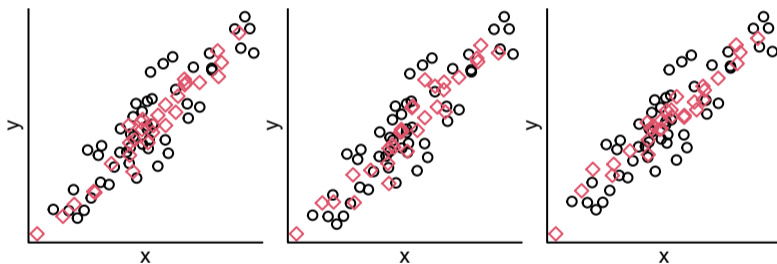
avec  $\eta \sim \mathcal{N}(0, \hat{\sigma}^2)$  et  $\hat{\sigma}^2$  est estimé sur  $S$  via

$$\hat{\sigma}^2 = \frac{1}{n + \tilde{n} - p - 1} \sum_{i=1}^{n+\tilde{n}} (y_i - \hat{y}_i)^2,$$

ici  $n = 1$

## La régression stochastique - *improper imputation* (2)

On répète un nombre  $M$  20 d'imputations du jeu de données initiales. En voici 3 versions.



La moyenne empirique des  $M = 20$  estimations pour chacun des 3 coefficients:

$$\hat{\alpha}_N \approx -0.066 \pm 0.047, \quad \hat{\beta}_N \approx 1.078 \pm 0.054, \quad \hat{\sigma}_N^2 \approx 0.162$$

## La régression stochastique - *improper imputation* (3)

### Problème

Cette solution ne prend pas en compte la variabilité sur les paramètres.

### Deux solutions *proper*

- Estimer  $M$  jeux de données imputés en intégrant une variabilité des paramètres à chaque fois. Plusieurs solutions ont été imaginées :
  - Tirer  $M$  paramètres sur la distribution a posteriori. Estimer  $M$  jeux de données différents en suivant ces paramètres Tanner & Wong (1987), utilisée dans **MICE** van Buuren & Groothuis-Oudshoorn (2011)
  - *Bootstrapper* ( $S, \tilde{S}$ ) pour créer  $M$  jeux de données dans lesquels il y a potentiellement des données manquantes, approche de **missMDA** Josse & Husson (2016).
- Ensuite appliquer la méthode d'analyse sur chacun des jeux de données séparément.
- Estimer les paramètres par agrégation des  $M$  modèles.



## Vers l'imputation multiple - *proper*

On obtient des estimations :

- Via bootstrap :

$$\hat{\alpha}_{\text{Boot}} \approx -0.08 \pm 0.05, \quad \hat{\beta}_{\text{Boot}} \approx 1.069 \pm 0.059, \quad \hat{\sigma}_{\text{Boot}}^2 \approx 0.19$$

- Via la postérieure/bayésienne :

$$\hat{\alpha}_{\text{Post}} \approx -0.077 \pm 0.055, \quad \hat{\beta}_{\text{Post}} \approx 1.085 \pm 0.064, \quad \hat{\sigma}_{\text{Post}}^2 \approx 0.232$$

Pour rappel, dans les cas impropres :

Moyenne	$\hat{\alpha}_0 \approx -0.007 \pm 0.067$	$\hat{\beta}_0 \approx 0.689 \pm 0.077$	$\hat{\sigma}_0^2 \approx 0.339$
Régression linéaire	$\hat{\alpha}_l \approx -0.068 \pm 0.044$	$\hat{\beta}_l \approx 1.072 \pm 0.05$	$\hat{\sigma}_l^2 \approx 0.146$
Régression stochastique	$\hat{\alpha}_N \approx -0.066 \pm 0.047$	$\hat{\beta}_N \approx 1.078 \pm 0.054$	$\hat{\sigma}_N^2 \approx 0.162$

## Règle de Rubin (*Rubin's Rule*)

Soit  $M$  jeux de données imputés alors la règle de Rubin stipule que

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m,$$

où  $\hat{\theta}_m$  est la valeur du paramètre estimé pour le jeu de données d'indice  $m$ . Il vient alors :

$$\hat{\sigma}_{IM}^2 = \hat{\sigma}_W^2 + \left(1 + \frac{1}{M}\right) \hat{\sigma}_B^2,$$

$$\text{variance imputée : } \hat{\sigma}_W^2 = \frac{1}{M} \sum_{m=1}^M \hat{\sigma}_m^2,$$

$$\text{variance inter-imputation : } \hat{\sigma}_B^2 = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta}_{IM})^2.$$

Sachant que les variances par modèle sont estimées via

$$\begin{aligned} \hat{\sigma}_m^2(\hat{\alpha}) &= \hat{\sigma}^2 \left( \frac{1}{n+\bar{n}} + \frac{\bar{x}^2}{\sum_{i=1}^{n+\bar{n}} (x_i - \bar{x})^2} \right) \\ \hat{\sigma}_m^2(\hat{\beta}) &= \hat{\sigma}^2 \frac{1}{\sum_{i=1}^{n+\bar{n}} (x_i - \bar{x})^2} \end{aligned}$$

## Maximisation de vraisemblance

Soit une fonction de densité  $p(x|\theta)$  qui dépend du paramètre  $\theta$ , potentiellement multi-dimensionnel. Dans le cas d'un échantillon de  $n$  observations indépendantes  $X = (x_1, \dots, x_n)$  il vient

$$p(X|\theta) = \prod_{i=1}^n p(x_i|\theta).$$

Plus cette valeur est importante et plus le modèle  $\theta$  est vraisemblable suivant les données  $X$ , pour cette interprétation nous utilisons souvent la notation

$$\mathcal{L}(\theta|X) = \log(p(X|\theta)),$$

que nous appelons **log-likelihood**. Bien noter l'inversion  $\theta|X$  et  $X|\theta$ . La fonction  $\mathcal{L}$  n'est pas une densité.

## Maximisation de vraisemblance (2)

Le meilleur paramètre pour nos données  $X$  sera donc celui qui maximise cette vraisemblance.

$$\max_{\theta \in \Theta} \mathcal{L}(\theta|X).$$

Dans le cas simple, sans données manquantes et en supposant que  $x_i \sim \mathcal{N}(\mu, \Sigma)$ , alors  $\mathcal{L}(\cdot|X)$  est une fonction de classe  $\mathcal{C}^1$  et par annulation de son gradient on trouve que l'estimateur **au maximum de vraisemblance** correspond au couple *moyenne empirique* et *matrice de variance empirique* sur l'échantillon  $X$ .

Le problème se pose lorsque des données manquent dans  $X$ ...

## Algorithme **EM** (Intro)

Si l'on suppose des données observées  $X^{(obs)}$  et des données manquantes  $X^{(mis)}$  d'un échantillon  $X$  de loi de paramètre  $\theta$ , alors il vient

$$p(X^{(obs)}, X^{(mis)}|\theta) = p(X^{(mis)}|X^{(obs)}, \theta)p(X^{(obs)}|\theta),$$

ce qui implique, par passage au logarithme, pour une valeur courante de  $\theta$  notée  $\theta^{(i)}$  par multiplication par  $p(X^{(mis)}|X^{(obs)}, \theta^{(i)})$  et intégration sur le support de  $X^{(mis)}$  noté  $\mathcal{X}$

$$\mathcal{L}(\theta|X^{(obs)}) = \int_{x \in \mathcal{X}} p(x|X^{(obs)}, \theta^{(i)}) \ln \left( p(x, X^{(obs)}|\theta) \right) dx \quad (1)$$

$$- \int_{x \in \mathcal{X}} p(x|X^{(obs)}, \theta^{(i)}) \ln \left( p(x|X^{(obs)}, \theta) \right) dx, \quad (2)$$

$$= Q(\theta|\theta^{(i)}) + H(\theta|\theta^{(i)}) \quad (3)$$

## Algorithme **EM** (Intro 2)

La fonction  $H(\cdot|\theta^{(i)}) \rightarrow \int_{x \in \mathcal{X}} p(x|X^{(obs)}, \theta^{(i)}) \ln \left( p(x|X^{(obs)}, \cdot) \right) dx$  vérifie  $\forall \theta \in \Theta$   
 $H(\theta|\theta^{(i)}) \geq H(\theta^{(i)}|\theta^{(i)})$  et donc

$$\mathcal{L}(\theta|X^{(obs)}) - \mathcal{L}(\theta^{(i)}|X^{(obs)}) \geq Q(\theta|\theta^{(i)}) - Q(\theta^{(i)}|\theta^{(i)}),$$

ainsi en améliorant  $Q(\theta|\theta^{(i)})$  para rapport à  $\theta$ , on améliore  $\mathcal{L}(\theta|X^{(obs)})$  au moins aussi bien. En plus,  $Q(\theta|\theta^{(i)})$  est bien souvent plus simple à calculer que  $\mathcal{L}(\theta|X^{(obs)})$ .

L'algorithme est tout trouvé : partir d'un point  $\theta^{(0)}$  et itérer l'estimation de Q et la maximisation de Q... algorithme EM!

# Algorithme **EM**

**EM** (pour **E**stimation-**M**aximization) permet d'estimer des paramètres  $\theta$  dans un modèle à vraisemblance. Soit l'espérance conditionnelle, pour une itération  $i > 0$  :

$$Q\left(\theta|\theta^{(i)}\right) = \mathbb{E}_{X^{(mis)}} \left[ \mathcal{L}\left(\theta|X^{(mis)}, X^{(obs)}\right) | X^{(obs)}, \theta^{(i)} \right],$$

Dans le cas normale multivarié de paramètre  $\theta = (\mu, \Sigma)$ , cette fonction s'écrit

$$Q\left(\theta|\theta^{(i)}\right) = -\ln(|\Sigma|) - \sum_{i=1}^{n+\tilde{n}} (x_i - \mu)' \Sigma^{-1} (x_i - \mu).$$

## Algorithme **EM** (2)

**Résultat théorique**  $Q(\theta|\theta^{(i)})$  est un fonctionnelle qui montre la vraisemblance d'un jeu de données (imputé conditionnellement à  $\theta^{(i)}$ ) pour un paramètre  $\theta$ . Par construction si

$$\theta^{(i+1)} = \operatorname{argmax}_{\theta} Q(\theta|\theta^{(i)}),$$

alors estimer les données manquantes conditionnellement à  $\theta^{(i+1)}$  est plus intéressant :  $\mathcal{L}(\theta^{(i+1)}|X^{(obs)}) \geq \mathcal{L}(\theta^{(i)}|X^{(obs)})$ . Le nouveau paramètre est donc plus vraisemblable.

**Idée de l'algorithme** Alternier des étapes d'estimation de  $\theta^{(i)}$  et d'estimation de données manquantes.



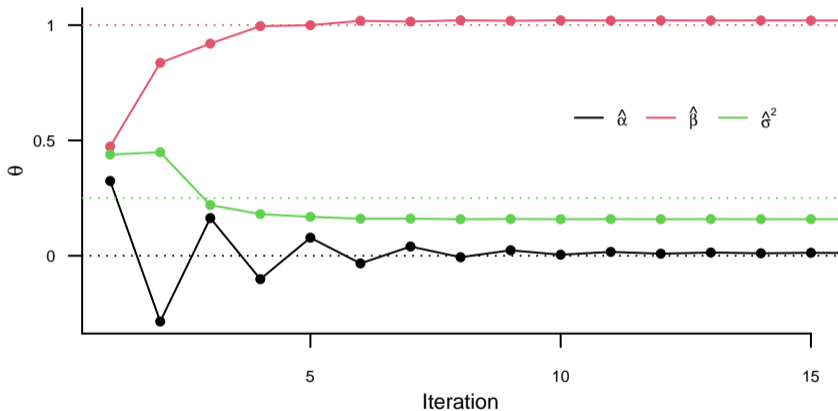
## Algorithme **EM** (3)

Soit pour l'itération  $i$  en ayant initialisé  $\theta^{(i)}$ . On répète les deux étapes suivantes jusqu'à convergence :

**Etape E** Estimation des données manquantes et donc de  $Q(\theta|\theta^{(i)})$ .

**Etape M** Maximisation de  $Q(\theta|\theta^{(i)})$  obtention de  $\theta^{(i+1)}$ .

# Algorithme EM (4)



Les valeurs estimées sont alors

# Au total

$$\hat{\alpha}_{EM} \approx 0.012 \pm 0.041, \hat{\beta}_{EM} \approx 1.02 \pm 0.047, \hat{\sigma}_{EM}^2 \approx 0.159$$

$$\hat{\alpha}_{Boot} \approx -0.08 \pm 0.05, \hat{\beta}_{Boot} \approx 1.069 \pm 0.059, \hat{\sigma}_{Boot}^2 \approx 0.19$$

$$\hat{\alpha}_{Post} \approx -0.077 \pm 0.055, \hat{\beta}_{Post} \approx 1.085 \pm 0.064, \hat{\sigma}_{Post}^2 \approx 0.232$$

$$\hat{\alpha}_0 \approx -0.007 \pm 0.067, \hat{\beta}_0 \approx 0.689 \pm 0.077, \hat{\sigma}_0^2 \approx 0.339$$

$$\hat{\alpha}_l \approx -0.068 \pm 0.044, \hat{\beta}_l \approx 1.072 \pm 0.05, \hat{\sigma}_l^2 \approx 0.146$$

$$\hat{\alpha}_N \approx -0.066 \pm 0.047, \hat{\beta}_N \approx 1.078 \pm 0.054, \hat{\sigma}_N^2 \approx 0.162$$

## Section 3

# Imputation et vraies données

## Un passage à l'échelle nécessaire

Nous avons vu les principales caractéristiques de l'imputation par comparaison à la régression classique.

En pratique, les jeux de données ne sont pas bivariés et les NA ne sont pas que dans une seule variable. . .

Domage. . .

Des modifications de ce que nous venons de voir ont été imaginées afin de gérer la présence de NA dans le cas général.

# missMDA

- Utilisation de méthodes factorielles.
- Variables qualitatives/quantitatives/mixtes.
- Imputation simple/multiple.
- Echantillonnage bootstrap/bayésienne.
- Beaucoup de choses à dire. . .

## Amelia II (algorithme **EMB**), Honaker & King (2010)

- Hypothèse d'un jeu de donnée complété qui suit une distribution normale multivariée ( $p$  variables) de paramètre  $(\mu, \Sigma)$ .
- Bootstrap ou bayésienne

Echantillonnage bootstrap et imputation via algorithme EM... EMB(ootstrap).

# mice

Modèles conditionnels chaînés.

Utilise une pénalisation Ridge dans l'imputation, paramètre  $\kappa$  de l'ordre de  $\kappa = 0.0001$  ( $\kappa = 0.1$  est grand dans ce cas et *“may introduce a systematic bias toward the null, and should thus be avoided”*)

4 modèles d'imputation :

- `norm.predict` : imputation sans alea,
- `norm.nob` : imputation stochastique,
- `norm` : imputation stochastique et postérieure/bayésienne,
- `norm.boot` : Estimation des paramètres sur un échantillon bootstrap des données observées.



## mice (2)

Les variables sont organisées par nombre d'obs. NA croissant. Un modèle sur la première variable conditionnellement aux autres est construit. Les données manquantes sont imputées grâce à ce modèle. C'est au tour de la variable suivante etc... On recommence jusqu'à convergence.

Voir **Vignette**<sup>3</sup>.

---

<sup>3</sup><https://stefvanbuuren.name/fimd/sec-linear-normal.html#def:norm>

# missForest

- Utilisation de “random forests”
- Variables qualitatives/quantitatives/mixtes.
- Imputation simple.

Les variables sont organisées par nombre d'obs. NA croissant. Un modèle sur la première variable conditionnellement aux autres est construit, sur les observations présentes pour cette variable. Les données manquantes sont imputées grâce à ce modèle. C'est au tour de la variable suivante etc. . . On recommence jusqu'à convergence.

## k-NN, Troyanskaya *et al.* (2001)

- Non-itérative.
- Imputation simple. Pour une observation avec des NA, utilise les observations La fonction `impute.knn` du package `impute`. Le paramètre le plus important est `k`, le nombre de plus proches voisins.
- Package `impute` : **kNN** réalisé sur les variables et non les observations. Données quantitatives. Distance euclidienne.
- Package `VIM`. Données mixtes. Distance de Gower.

# Imputation en grande dimension

L'imputation en grande dimension a tendance à ne faire ressortir que les premières dimensions du jeu de donnée (nombre de composantes en ACP).

Or en grande dimension, les dimensions associées avec la réponse ne sont pas forcément dans les premières dimensions du jeu de données.

**Exemple** : Ebola et données génétiques, dataset rVSV\_ZEBOV Rechtien *et al.* (2017).

Il convient d'imputer en prenant en compte l'information que l'on souhaite retrouver ( $\mathbf{y}$ ). C'était l'objet de ma thèse et de la méthode **Koh-Lanta**, voir LORENZO *et al.* (2019).

## Section 4

### Conclusion

# Conclusion

C'est un sujet difficile et/car computationnel, lire :

- Imbert & Vialaneix (2018), une biblio très fournie méthodo et implémentations.
- [Imputation multiple & analyse factorielle, François Husson<sup>4</sup>](#).
- [Utilisation d'Amelia II et imputation multiple<sup>5</sup>](#)

---

<sup>4</sup>[http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/105543\\_museum\\_hist\\_nat.pdf](http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/105543_museum_hist_nat.pdf)

<sup>5</sup><https://cran.r-project.org/web/packages/Amelia/vignettes/intro-mi.html>

# Activités

- Lire les vignettes associées aux implémentations.
- Analyser un jeu de données avec (ou sans) des NA avec l'une (ou plusieurs) des méthodes discutées.

## Références I

HONAKER, J. & KING, G. (2010) What to do about missing values in time series cross-section data. *American Journal of Political Science*, **54**, 561–581.

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 Société Française de Statistique*, **159**, 1–55.

JOSSE, J. & HUSSON, F. (2016) missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, **70**, 1–31.

LITTLE, R.J. & RUBIN, D.B. (1976) *Statistical analysis with missing data*, vol. 793, ed. John Wiley & Sons.



## Références II

LORENZO, H., SARACCO, J., & THIÉBAUT, R. (2019) Supervised learning for multi-block incomplete data. *arXiv preprint arXiv:1901.04380*.

RECHTIEN, A., RICHERT, L., LORENZO, H., MARTRUS, G., HEJBLUM, B., DAHLKE, C., KASONTA, R., ZINSER, M., STUBBE, H., MATSCHL, U., & OTHERS (2017) Systems vaccinology identifies an early innate immune signature as a correlate of antibody responses to the ebola vaccine rVSV-zebov. *Cell reports*, **20**, 2251–2261.

STEKHOVEN, D.J. & BUEHLMANN, P. (2012) MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.

TANNER, M.A. & WONG, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

## Références III

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., & ALTMAN, R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

VAN BUUREN, S. & GROOTHUIS-OUUDSHOORN, K. (2011) mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, **45**, 1–67.