Introduction à une étude statistique avec données manquantes

Matthieu Marbac
Ensai - CREST

"We are surrounded by missing data. Problems created by missing data in statistical analysis have long been swept under the carpet [van Buuren, 2018]."

"Missing data refers to a data value that should have been recorded but, for some reason, was not [Day, 1999]."

"Missing data are unobserved values that would be meaningful for analysis if observed ; in other words, a missing value hides a meaningful value [Little and Rubin, 2019]."

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Terminologie

On dispose d'un échantillon composé de n individus décrits par p variables :

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top.$$

On définit la matrice binaire d'indicateurs de réponse

$$\boldsymbol{R} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)^\top,$$

où $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{ip})^\top \in \{0,1\}^p$ avec $R_{ij} = 1$ si la variable j est observée pour l'individu i et $R_{ij} = 0$ sinon.

$$\mathbf{X} = \begin{bmatrix} 1.1 & 2.4 & 0.5 \\ 1.5 & 3.0 & 1.2 \\ 0.4 & 9.1 & 6.0 \\ 0.3 & 1.5 & 9.2 \\ 0.2 & 4.2 & 5.5 \end{bmatrix} \quad \text{et} \quad \boldsymbol{R} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Terminologie

Il est important de bien faire la distinction entre les éléments suivants :

- *L'ensemble des données* regroupe toutes les variables : $(\boldsymbol{R}, \mathbf{X})$.
- *Les données observées* $(\boldsymbol{R}, \mathbf{X}^{\text{obs}})$ où $\mathbf{X}^{\text{obs}}$ regroupe les données observées pour les individus (*i.e.*, l'ensemble des X_{ij} tels que $R_{ij} = 1$).
- *Les données manquantes* $\mathbf{X}^{\text{miss}}$ sont celles dont on ne dispose pas (*i.e.*, l'ensemble des X_{ij} tels que $R_{ij} = 0$).
- *Les données complètes* sont constituées du sous-échantillon d'individus n'ayant pas d'observation manquante (*i.e.*, un individu i fera partie de ce sous-échantillon uniquement si $\prod_{j=1}^{p} R_{ij} = 1$).

$$\mathbf{X} = \begin{bmatrix} 1.1 & 2.4 & 0.5 \\ 1.5 & 3.0 & 1.2 \\ 0.4 & 9.1 & 6.0 \\ 0.3 & 1.5 & 9.2 \\ 0.2 & 4.2 & 5.5 \end{bmatrix} \quad \text{et} \quad \boldsymbol{R} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Paramètre d'intérêt

La distribution des données observées est obtenue en intégrant la distribution de toutes les données par rapport aux données manquantes, de sorte que

$$f(\mathbf{X}^{\text{obs}}, \boldsymbol{R}) = \int f(\mathbf{X}, \boldsymbol{R}) d\mathbf{X}^{\text{miss}}.$$

Lorsqu'on se place dans un cadre paramétrique, on a

$$f(\mathbf{X}^{\text{obs}}, \boldsymbol{R}; \theta) = \int f(\mathbf{X}; \gamma) f(\boldsymbol{R} \mid \mathbf{X}; \psi) d\mathbf{X}^{\text{miss}},$$

où $\theta = (\gamma^\top, \psi^\top)^\top$ regroupe l'ensemble des paramètres du modèle tels que

- γ groupe les paramètres relatifs à la distribution des variables $\mathbf{X}$.
- ψ groupe les paramètres relatifs à la distribution conditionnelle de la non-réponse sachant les variables $\mathbf{X}$.

Paramètre d'intérêt

La distinction entre paramètre d'intérêt et paramètre de nuisance dépend de l'objectif de l'étude statistique.

Cet objectif doit être parfaitement établi avant la mise en place de l'analyse statistique car il détermine les méthodes à utiliser.

On utilisera des approches différentes si le but est :

- d'obtenir un jeu de données sans valeurs manquantes.
- d'estimer γ pour avoir des informations sur la distribution marginale de $\mathbf{X}_j$.

Il est d'usage de décrire les valeurs manquantes d'un jeu de donnée à partir du *schéma* qu'ont les valeurs manquantes et de leur *mécanisme*. En fonction de ces deux caractéristiques, des méthodes statistiques seront adaptées ou non.

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

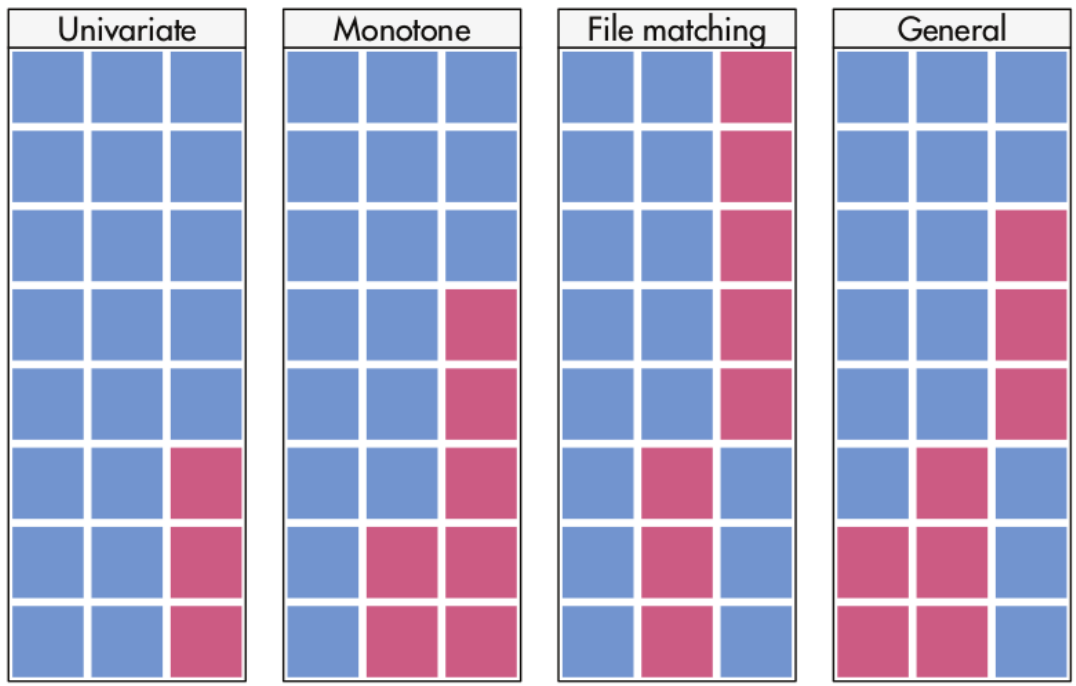
Analyse en Composantes Principales et imputation

Conclusion

Schémas des données manquantes

Définition (Schéma des données manquantes)

Ils indiquent comment sont organisés les emplacements des valeurs manquantes dans les données $\mathbf{X}$ *et se définissent donc uniquement à partir de la matrice* $\boldsymbol{R}$.

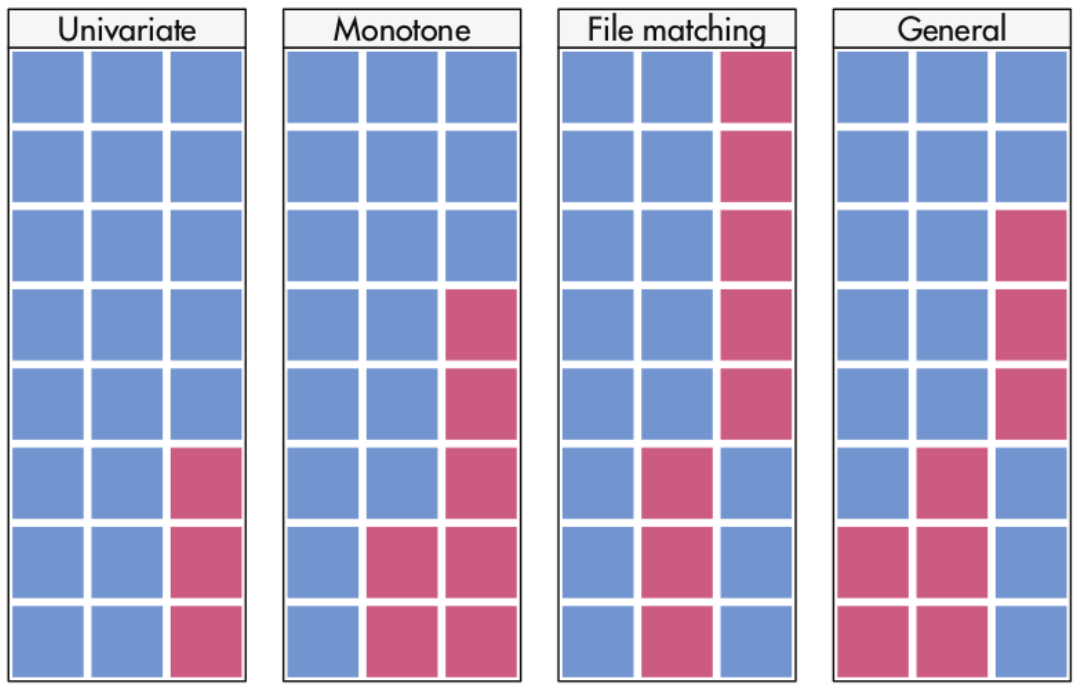


Schémas des données manquantes

Définition (Schéma univarié)

Le schéma des données manquantes est univarié si

$$\exists! j_0 \in [\![1, p]\!], \ \sum_{i=1}^{n} R_{ij_0} < n.$$

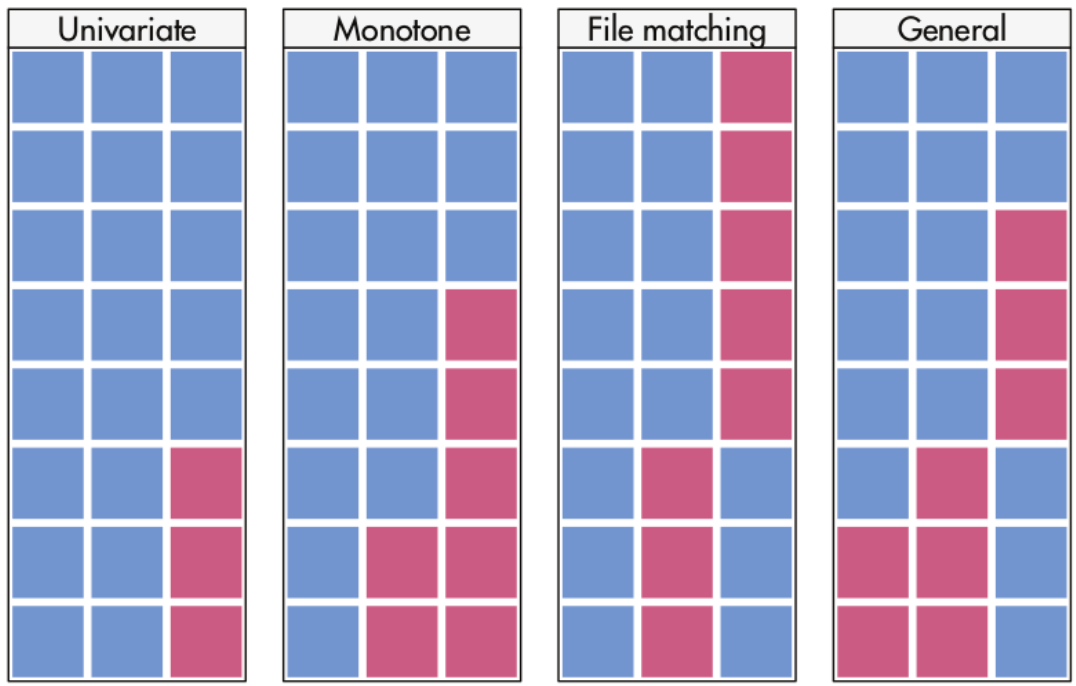


Schémas des données manquantes

Définition (Schéma monotone)

Le schéma des données manquantes est monotone si il existe une permutation des indices des variables $\sigma : [\![1, p]\!] \to [\![1, p]\!]$ *telle que*

$$\forall i,\ R_{i\sigma(j)} = 0 \Longrightarrow R_{i\sigma(j+1)} = 0.$$

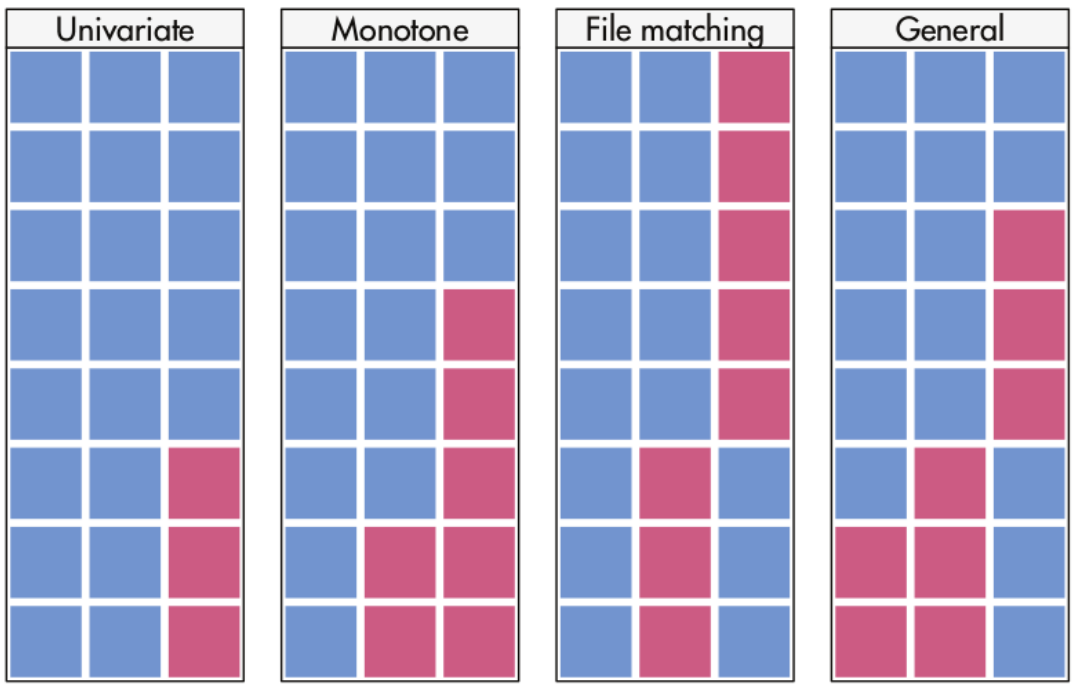


Schémas des données manquantes

Définition (Schéma file matching)

Soit $\Gamma = \{j \in [\![1, p]\!] : \sum_{i=1}^{n} R_{ij} < n\}$ *le sous ensemble des indices de variables présentant des valeurs manquantes. Le schéma des données manquantes est file matching si il existe une partition* $\{\{\Gamma_1\}, \ldots, \{\Gamma_K\}\}$ *en* K *groupes non vides de* Γ *telle que*

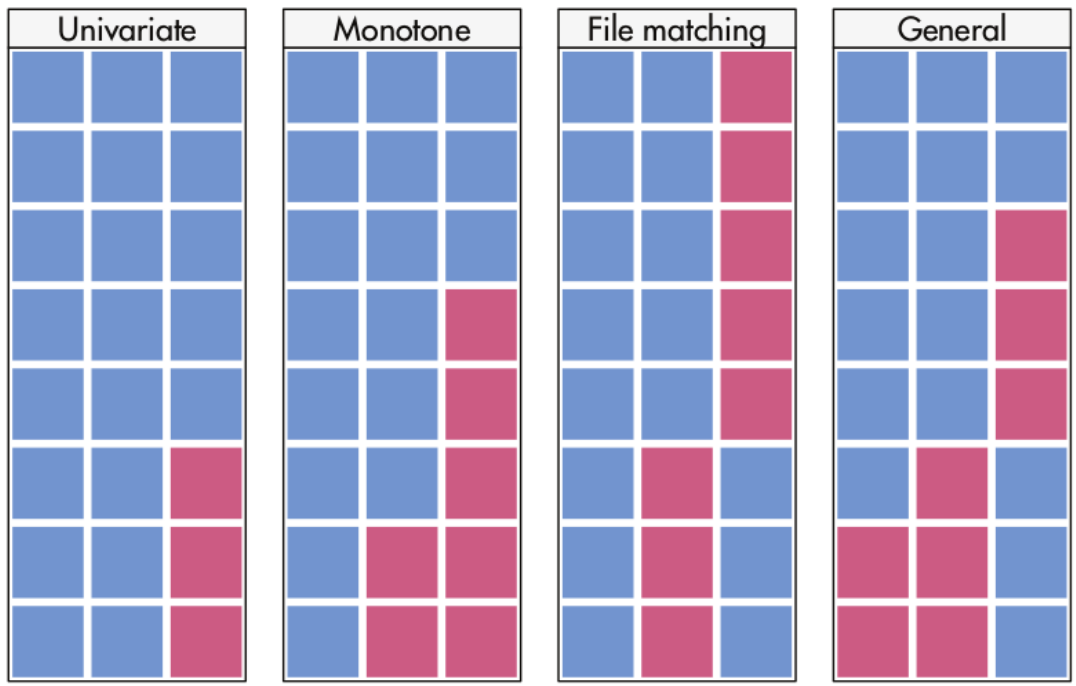
$$\forall i,\ R_{ij} = 1 \text{ avec } j \in \Gamma_k \Longrightarrow R_{ij'} = \begin{cases} 1 & \text{si } j' \in \Gamma_k \\ 0 & \text{si } j' \notin \Gamma_k \end{cases} .$$



Schémas des données manquantes

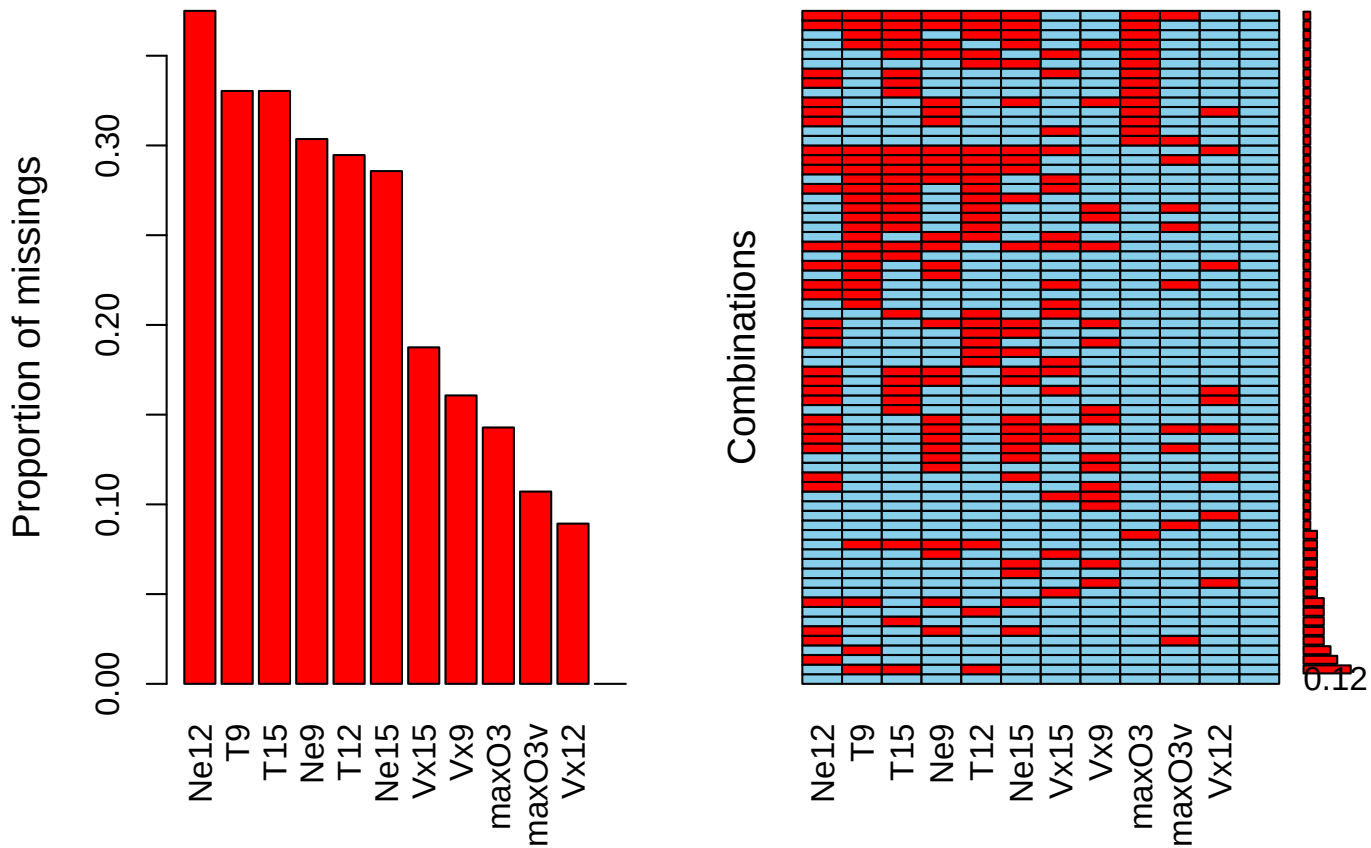
Définition (Schéma général)

Les données manquantes suivent un schéma général si elles ne suivent pas les schémas décrits précédemment.



Schémas des données manquantes sous R

```
don <- read.table("http://factominer.free.fr/missMDA/ozoneNA.csv",
                  header=TRUE, sep=",", row.names=1)
library(VIM)
aggr(don,only.miss=TRUE,sortVar=TRUE)
```



Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Mécanismes de données manquantes

Définition (Mécanismes de données manquantes)

C'est le processus qui produit les valeurs manquantes. Il détermine le lien entre $\mathbf{X}$ *et* $\boldsymbol{R}$.

Le mécanisme engendrant les données manquantes a un rôle majeur, puisque son type détermine si la méthode statistique a ou non de bonnes propriétés.

On divise généralement les mécanismes en trois catégories [Rubin, 1976]

- les données manquantes générées complètement au hasard (MCAR).
- les données manquantes générées au hasard (MAR).
- les données manquantes générées autrement (MNAR).

Mécanisme MCAR

Un mécanisme MCAR est caractérisé par le fait que la probabilité qu'une donnée soit manquante n'est pas liée avec les données $\mathbf{X}$.

Le fait qu'une donnée soit manquante ne dépend donc ni de la valeur de cette donnée ni des données observées (*i.e.*, $\boldsymbol{R} \perp\!\!\!\perp \mathbf{X}$).

Definition (Mécanisme MCAR)

Un mécanisme est de type MCAR si

$$f(\boldsymbol{R} \mid \mathbf{X}) = f(\boldsymbol{R}).$$

Sous un mécanisme MCAR, les observations complètes (sous-échantillon composé des individus n'ayant pas de valeurs manquantes) forment un sous-échantillon représentatif de toutes les observations.

Ainsi n'importe quelle méthode statistique valide dans le cas où toutes les données sont observées restera valide si on l'applique à ce sous-échantillon.

Mécanisme MCAR

Bien qu'aucun biais ne soit ajouté en considérant uniquement le sous-échantillon des données complètes, la réduction du nombre d'observations utilisées pour l'étude statistique fera augmenter la variance des estimateurs.

Ainsi, même sous un mécanisme MCAR, les méthodes de gestion des données manquantes ont un intérêt. Ces méthodes deviennent même indispensables lorsque le schéma est de type *file sending* puisque le nombre d'individus complets est nul par construction (*i.e.*, $n_{\text{full}} = \sum_{i=1}^{n} \prod_{j=1}^{p} R_{ij} = 0$).

Mécanisme MCAR avec covariables

Il peut arriver que des covariables supplémentaires soient accessibles, notamment dans le cas de suivi de cohortes.

Par exemple, $\mathbf{X}_i$ est une donnée longitudinale qui comporte les mesures réalisées sur l'individu i à différents instants et pouvant présenter des données manquantes. On dispose également de covariables donnant des informations supplémentaires (sans valeurs manquantes) $\mathbf{V}_i$ mesurées sur l'individu i (*e.g.*, genre, age,...).

On trouve deux définitions de MCAR dans la littérature

- le terme MCAR est restreint au cas où l'indicateur de réponse est indépendant de toutes les variables (*i.e.*, $\boldsymbol{R}_i \perp \mathbf{X}_i, \mathbf{V}_i$; Little [1995])
- le terme MCAR est étendu au cas où l'indicateur de réponse peut dépendre des covariables mais est conditionnellement indépendant à $\mathbf{X}_i$ (*i.e.*, $\boldsymbol{R}_i \perp \mathbf{X}_i \mid \mathbf{V}_i$; Daniels and Hogan [2008]).

Pour la seconde définition, on retrouve également le terme de mécanisme de données manquantes MCAR-covariate-dependent comme proposé par Little [1995].

Mécanisme MAR

Une hypothèse moins restrictive que MCAR consiste à supposer que la probabilité que des données soient manquantes ne dépend pas des données elles-mêmes mais peut dépendre des données observées. Dans ce cas, le mécanisme est de type MAR.

Définition (Mécanisme MAR)

Un mécanisme est de type MAR si

$$f(\boldsymbol{R} \mid \mathbf{X}) = f(\boldsymbol{R} \mid \mathbf{X}^{obs}).$$

Sous l'hypothèse MAR, les observations complètes ne suivent pas la même distribution que la population. Ainsi, on ne peut pas considérer ces observations comme un sous-échantillon aléatoire de $\mathbf{X}$.

Mécanisme MAR

L'hypothèse MAR est une condition permettant de mettre un place une estimation valide sans modéliser le mécanisme de génération des valeurs manquantes.

Dans un cadre paramétrique, pour un mécanisme MAR, on a la décomposition

$$f(\mathbf{X}^{\text{obs}}, \boldsymbol{R}; \theta) = f(\boldsymbol{R} \mid \mathbf{X}^{\text{obs}}; \psi) f(\mathbf{X}^{\text{obs}}; \gamma).$$

L'estimation du paramètre d'intérêt γ peut se faire sans tenir compte de la distribution conditionnelle de l'indicateur de réponse sachant les données observées.

L'estimation du paramètre d'intérêt γ ne nécessite pas de connaître le processus de génération des données manquantes.

Les valeurs manquantes peuvent être extrapolées en utilisant les données observées et la distribution du vecteur $\mathbf{X}_i$.

Mécanisme MNAR

Tout mécanisme qui ne vérifie pas l'hypothèse MAR est un mécanisme MNAR.

Pour les mécanismes MNAR, la probabilité d'apparition des données manquantes dépend de la partie non-observée des données.

Définition (Mécanisme MNAR)

Un mécanisme est de type MNAR si

$$f(\boldsymbol{R} \mid \mathbf{X}) \neq f(\boldsymbol{R} \mid \mathbf{X}^{obs}).$$

Sous l'hypothèse MNAR, presque toutes les méthodes d'analyse classiques ne sont pas valides.

Par exemple, les méthodes ignorant le modèle de génération des données manquantes (*i.e.*, la distribution conditionnelle de $\boldsymbol{R}$ sachant $\mathbf{X}$) produisent des résultats biaisés.

Cela s'explique par le fait que la distribution des données observées ne se simplifie pas contrairement au cas MAR.

Il faut généralement de modéliser la distribution du couple $(\mathbf{X}, \boldsymbol{R})$ pour l'inférence.

Impact du type de mécanisme

- $\mathbf{X}$ est composé de n réalisations indépendantes issues d'une loi normale bivariée centrée, de variances égales à 1 et de corrélation égale à 0,5.
- X_{i1} est toujours observé et X_{i2} peut avoir des valeurs manquantes
- On considère trois mécanismes différents

MCAR $P[R_{i\mathbf{2}} = 1 \mid \mathbf{X}_i] = 0.5$
MAR $P[R_{i\mathbf{2}} = 1 \mid \mathbf{X}_i] = (1 + e^{\mathbf{10}X_{i\mathbf{1}}})$
MNAR $P[R_{i\mathbf{2}} = 1 \mid \mathbf{X}_i] = (1 + e^{\mathbf{10}X_{i\mathbf{2}}})$.

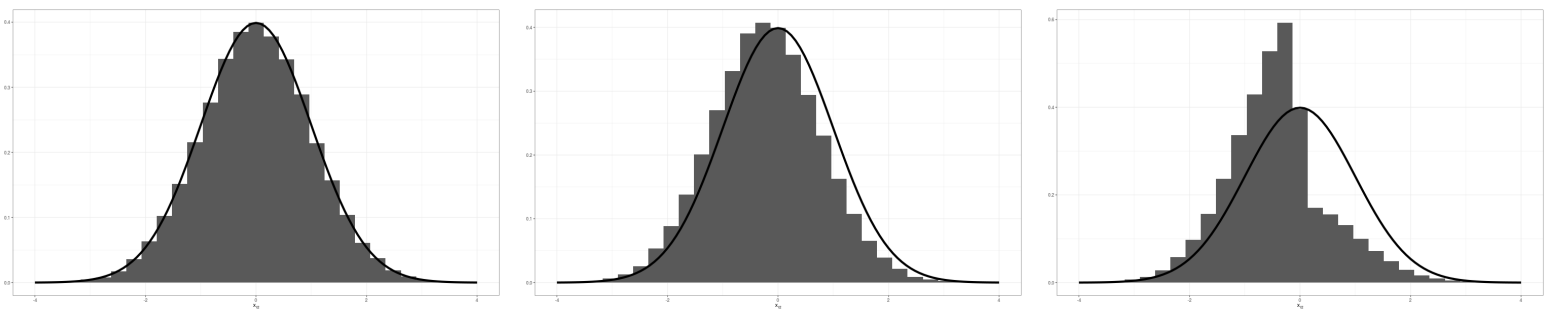


Figure – Densité marginale de $X_{i\mathbf{2}}$ comparée à l'histogramme des réalisations sans valeur manquante de $X_{i\mathbf{2}}$ obtenue pour les mécanismes MCAR (à gauche), MAR (au centre) et MNAR (à droite).

Impact du type de mécanisme

La modélisation de la loi de l'ensemble des données est plus complexe car il faut définir la relation entre les données $\mathbf{X}$ et les indicateurs de réponse $\boldsymbol{R}$.

Le mécanisme déterminant la validité des estimations statistiques, nous discutons maintenant comment celui-ci peut être diagnostiqué.

On ne parle pas de test car cela nécessiterait d'avoir accès aux valeurs manquantes.

Diagnostic du type de mécanisme

On peut chercher à valider empiriquement la pertinence d'une hypothèse MCAR contre une alternative MAR, mais uniquement sous la condition (invérifiable en pratique) que le mécanisme n'est pas MNAR.

Cette validation peut se faire au moyen d'un test sur l'égalité des distributions entre

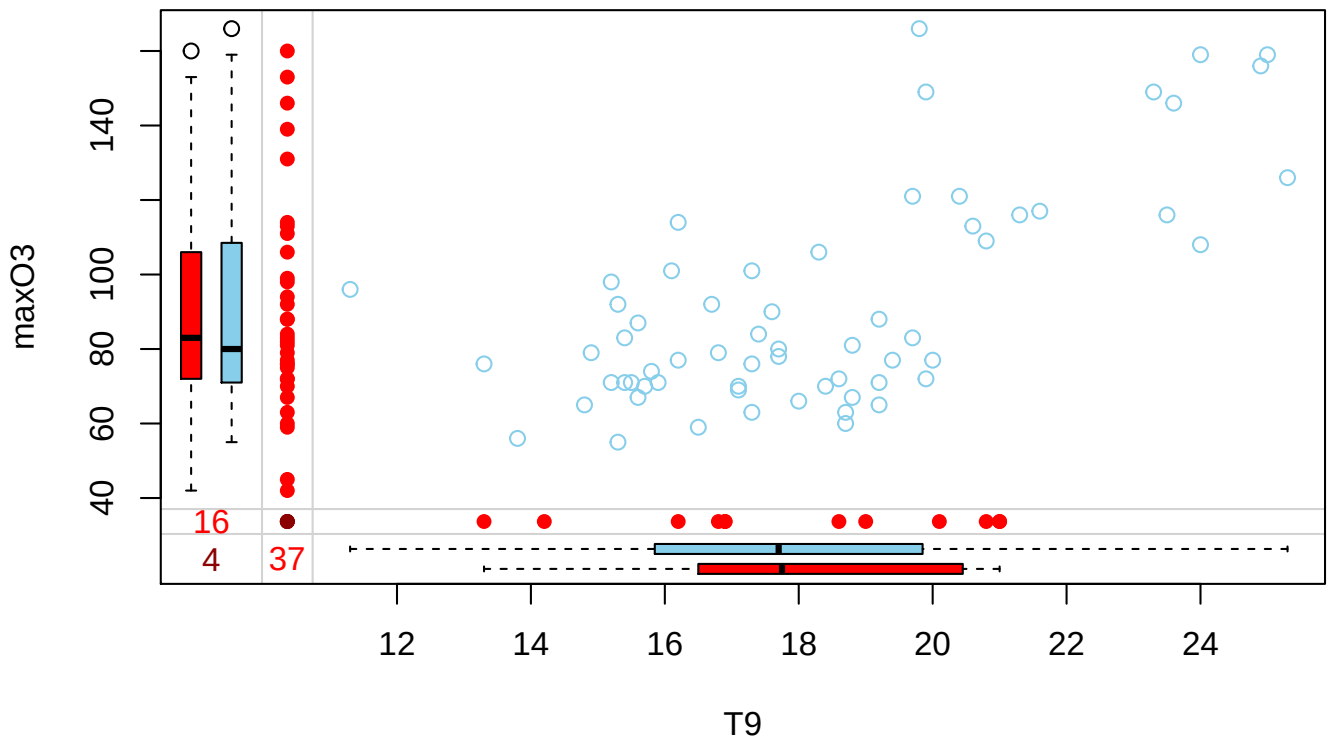
$$X_{ij} \mid R_{ik} = 1 \text{ et } X_{ij} \mid R_{ik} = 0,\ j \neq k$$

en utilisant par exemple des tests de Student ou du chi-deux ou un test global suivant le même principe [Little, 1988].

Cependant, l'hypothèse MCAR étant fort restrictive, il est courant que le test rejette cette hypothèse de modélisation.

Diagnostic du type de mécanisme

```
marginplot(don[,c("T9","maxO3")])
```



Diagnostic du type de mécanisme

L'hypothèse MAR permet elle aussi de mettre en place des méthodes d'estimation simples pour γ, car elle ne nécessite pas de modéliser la loi conditionnelle de $\boldsymbol{R} \mid \mathbf{X}^{\text{obs}}$.

Or, on ne peut pas tester l'hypothèse MAR contre MNAR.

Molenberghs et al. [2008] explique que pour chaque modèle MNAR appliqué à un ensemble de données observées, on peut obtenir la même vraisemblance par un modèle MAR.

Notons que les prédictions des valeurs manquantes conditionnellement aux données observées seront en général différentes pour les deux modèles.

Ce travail pose la questions cruciale de l'identifiabilité de la modélisation mais non pas au niveau des paramètres du modèle, comme c'est souvent le cas en statistiques, mais au niveau des modèles eux-mêmes.

Cas de non-identifiabilité entre MAR et MNAR

Comme le mécanisme MAR permet une estimation plus facile et comme un test statistique ne peut pas être mis en place pour distinguer les mécanismes MAR et MNAR, les hypothèses posées sont généralement vérifiées par d'autres études empiriques sur des sujets similaires.

En pratique, il est également d'usage d'inclure un maximum de variables (liées aux variables présentant des valeurs manquantes) de sorte que l'hypothèse MAR devienne le plus raisonnable possible (exemple : mesure du salaire).

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Ignorabilité

La notion d'ignorabilité permet de caractériser les cas où l'estimation de la loi de $\mathbf{X}$ peut se faire sans information sur le mécanisme engendrant les données manquantes.

Définition (Ignorabilité)

Le mécanisme des données manquantes est ignorable pour l'inférence fréquentiste si

MAR *: le mécanisme est MAR ;*

Distinction *: L'espace des paramètres est le produit de l'espace de γ et de l'espace de ψ.*

Remarque sur la propriété d'ignorabilité

- L'estimation de γ peut se faire sans modéliser le lien entre $\mathbf{X}$ et $\boldsymbol{R}$.
- L'ignorabilité n'implique pas qu'il faille négliger les données manquantes.
- L'imputation des valeurs manquantes est facile car

$$f(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \boldsymbol{R}) = f(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}).$$

Exemples de cas où la propriété d'ignorabilité est crédible

Voici deux exemples [Schafer, 1997] considèrent des cas où les données sont manquantes par construction (*missing by design*) puisque la collecte des données est organisée de sorte qu'on ne mesure pas toutes les variables sur tous les sujets. Dans ce cas, il est raisonnable de considérer que le mécanisme est MAR.

Exemple (Mécanisme ignorable et double échantillonnage)

Dans les plans de sondage qui considèrent un double sous-échantillonnage, s variables sont mesurées sur tous les sujets de l'échantillon et $p - s$ variables sont ensuite mesurées sur une sous-partie de l'échantillon sélectionnée aléatoirement uniquement à partir des s variables qui sont mesurées sur tous les sujets.

Exemple (Mécanisme ignorable et tests multiples partiels)

Le cas de tests multiples partiellement effectués est courant pour les études médicales où l'on dispose de deux tests de coûts (financiers ou humain) différents. Ainsi, on effectue le test peu coûteux sur un grand panel de sujets, ensuite le test le plus coûteux est effectué sur un sous-échantillon du panel (nommé échantillon de calibration). L'échantillon de calibration peut être choisi complètement au hasard ou en fonction des résultats obtenus lors du premier test.

Ignorabilité pour un paramètre

Pour certaines situations, le paramètre d'intérêt réside uniquement en une partie de γ.

Little et al. [2017] ont introduit la notion de mécanisme ignorable pour un paramètre.

Définition (Ignorabilité pour un paramètre)

Soit $\gamma = (\gamma_1^\top, \gamma_2^\top)^\top$ où γ_1 et γ_2 sont deux sous-ensembles de paramètres. Le mécanisme est ignorable pour le paramètre γ_1 si

- *γ_1 et $(\gamma_2^\top, \psi^\top)$ vérifient la propriété de distinction.*
- *La log-vraisemblance des données observées peut se décomposer comme*

$$\ell(\mathbf{X}^{obs}, \boldsymbol{R}; \theta) = \ell_1(\mathbf{X}^{obs}; \gamma_1) + \ell_2(\mathbf{X}^{obs}, \boldsymbol{R}; \gamma_2, \psi),$$

où $\ell_1(\mathbf{X}^{obs}; \gamma_1)$ ne tient pas compte du mécanisme des données manquantes et où $\ell_2(\mathbf{X}^{obs}, \boldsymbol{R}; \gamma_2, \psi)$ ne fait pas intervenir le paramètre γ_1.

Exemple d'ignorabilité pour un paramètre

On considère

- n réalisations indépendantes $\mathbf{X}_1, \ldots, \mathbf{X}_n$ telles que $\mathbf{X}_i = (X_{i1}, X_{i2})$ suit une loi normale centrée en $\gamma = (\gamma_1, \gamma_2)^\top$, de variances 1 et de corrélation ρ.
- $\mathbb{P}[R_{i1} = 1 \mid \mathbf{X}_i] = \psi_1$ et $\mathbb{P}[R_{i2} = 1 \mid \mathbf{X}_i] = \mathbb{1}_{\{X_{i2} \geqslant \psi_2\}}$

Le mécanisme est MNAR, donc non ignorable.

Cependant, γ_1 peut être estimé de manière consistante sans modéliser le mécanisme des valeurs manquantes, en considérant la décomposition de la log-vraisemblance observée avec

$$\ell_1(\mathbf{X}^{\text{obs}}; \gamma_1) = \sum_{i=1}^{n} R_{i1} \ln \phi(X_{i1}; \gamma_1).$$

Dans le cas où le mécanisme n'est pas ignorable pour le paramètre d'intérêt, il est généralement d'usage de modéliser la loi du couple $(\mathbf{X}, \boldsymbol{R})$ ce que nous pouvons faire par les trois approches que nous décrivons maintenant.

Modélisation de la loi de l'ensemble des données

Lorsque la propriété d'**ignorabilité** n'est pas satisfaite, on doit alors modéliser la loi de l'ensemble des données.

La modélisation de la loi de l'ensemble des données se fait par

- modèle de sélection (selection model)
- modélisation par mélange (pattern-mixture)
- paramètre partagé (shared parameter)

Notons qu'il existe cependant des méthodes valides, sous l'hypothèse MAR, qui spécifient la loi de l'ensemble des données.

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Méthodes basées sur la vraisemblance

Les méthodes basées sur la vraisemblance se placent dans un cadre paramétrique.

Rappelons que, si dans le cas ignorable, on ne fait des hypothèses paramétriques que sur la loi de $\mathbf{X}$, une modélisation de la loi du couple $(\mathbf{X}, \boldsymbol{R})$ est nécessaire dans le cas non ignorable.

L'inférence est faite par maximisation de la fonction de log-vraisemblance observée définie par

$$\ell(\mathbf{X}^{\text{obs}}, \boldsymbol{R}; \theta) = \sum_{i=1}^{n} \left[\ln f(\mathbf{X}_i^{\text{obs}}; \gamma) + \ln f(\boldsymbol{R}_i \mid \mathbf{X}_i^{\text{obs}}; \psi) \right].$$

Dans le cas ignorable, les estimateurs du maximum de vraisemblance peuvent être explicites.

Dans le cas non-ignorable, la maximisation de la vraisemblance se fera par des algorithmes itératifs de type EM.

Méthodes basées sur la vraisemblance

Dans un cadre Bayésien, on considére une loi *a priori* $\pi(\theta)$ sur le paramètre.

L'estimation bayésienne est basée sur la loi *a posteriori* des paramètres

$$\pi(\theta \mid \mathbf{X}^{\text{obs}}, \boldsymbol{R}) \propto f(\mathbf{X}^{\text{obs}}, \boldsymbol{R} \mid \theta)\pi(\theta).$$

Si une estimation ponctuelle est souhaitée, on peut chercher à déterminer l'estimateur maximisant la loi *a posteriori*. Cette maximisation peut se faire par un algorithme EM.

Les méthodes bayésiennes permettent de tenir compte de l'incertitude sur l'estimateur, notamment à travers la dispersion donnée par $\pi(\theta \mid \mathbf{X}^{\text{obs}}, \boldsymbol{R})$.

Si la loi *a posteriori* des paramètres est rarement explicite, il est possible de mettre en place des algorithmes de type Markov Chain Monte Carlo (*MCMC* ; Robert and Casella [2013]) pour générer un échantillon de paramètres selon la loi *a posteriori*.

Méthodes basées sur la vraisemblance

Les approches basées sur la vraisemblance permettent de gérer tous les schémas et mécanismes de données manquantes.

Des modélisations paramétriques devront être faites, notamment sur le mécanisme de génération des données manquantes dans le cas MNAR.

Ainsi, les conclusions statistiques seront à mettre en perspective en fonction de la validité de la famille de modèles paramétriques choisit.

Si des estimateurs explicites peuvent être obtenus dans le cas MAR, cela n'est en général pas le cas pour les mécanismes MNAR (l'intégrale induite dans la log-vraisemblance observée n'étant pas explicite) ce qui implique l'utilisation d'algorithmes itératifs.

Méthodes basées sur des pondérations

Le principe des méthodes de pondération est d'estimer la quantité d'intérêt sur le sous-échantillon d'individus ne comportant pas de valeurs manquantes.

Ces méthodes sont inspirées de la théorie des sondages [Horvitz and Thompson, 1952].

Afin de corriger le biais induit par cette restriction, on attribue un poids aux données sans valeurs manquantes.

En introduisant un poids reflétant la probabilité de non-réponse, les méthodes de pondérations permettent de corriger les sous/sur-représentations des données complètes et cherchent à rendre leur distribution égale à celle de la population.

Méthodes basées sur des pondérations

Les méthodes de pondérations nécessitent de modéliser le mécanisme à l'origine des données manquantes car c'est lui qui définit les poids.

Dans le cas d'un mécanisme MAR, on affecte à l'observation i le poids ω_i suivant

$$\omega_i = 1/\psi(\mathbf{X}_i^{\text{obs}}).$$

On calcule les statistiques nécessaires sur l'échantillon complet pondéré par les ω_i.

Cependant, il n'est pas nécessaire de spécifier la distribution de $\mathbf{X}$ (ni celle de $\mathbf{X}^{\text{obs}}$).

Exemple de méthode de pondération

On considère

- n observations indépendantes issues d'une loi bivariée d'espérance $\mu = (\mu_1, \mu_2)^\top \in \mathbb{R}^2$ et de variances σ_1^2 et σ_2^2 strictement positives.
- X_{i1} est toujours observé et $\mathbb{P}[R_{i2} = 1 \mid \mathbf{X}_i] = \psi(x_{i1}) > 0$.

On s'intéresse à l'estimation de μ_2 et de σ_2^2 par une méthode de pondération.

On affecte à l'observation i le poids ω_i suivant

$$\omega_i = 1/\psi(X_{i1}).$$

On a alors les estimateurs suivants

$$\hat{\mu}_2 = \frac{1}{n_\omega} \sum_{i=1}^{n} \omega_i R_{i2} X_{i2} \text{ et } \hat{\sigma}_2^2 = \frac{1}{n_\omega} \sum_{i=1}^{n} \omega_i R_{i2} (X_{i2} - \hat{\mu}_2)^2,$$

où $n_\omega = \sum_{i=1}^{n} R_{i2} \omega_i$.

Méthodes basées sur des pondérations

Les poids des observations doivent être estimés car la distribution conditionnelle de $\boldsymbol{R}$ sachant $\mathbf{X}$ est inconnue (c'est une différence par rapport aux sondages).

Sous un mécanisme MAR, la probabilité de non-réponse peut être estimée à partir des données observées $\mathbf{X}^{\text{obs}}$.

Les méthodes sont faciles à mettre en place dans le cas d'un schéma univarié ou si la probabilité conditionnelle de non-réponse sachant les données observées est une fonction monotone des données observées.

Une partie des développements actuels des méthodes de pondérations porte sur les méthodes *doublement robustes* [Tsiatis, 2007] qui modélisent la distribution de $\mathbf{X}$ en plus de la modélisation du mécanisme générant les données manquantes.

L'avantage de cette double modélisation est de pouvoir construire un estimateur consistant de la quantité d'intérêt dès lors qu'une des deux modélisations est correcte.

Méthodes basées sur l'imputation

Les méthodes basées sur l'imputation fonctionnent à partir de deux étapes qui permettent de

1. Compléter le jeu de données (prédire une valeur pour les données non observées).
2. Utilisée une méthode classique d'estimation pour un jeu de données complet.

On distinguera les méthodes d'imputation simple des méthodes d'imputation multiple.

Méthodes d'imputation simple

- Un seul jeu de données complet.
- Imputation selon $\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \boldsymbol{R}$.
- Méthodes d'imputation déterministe ou stochastique.
- Cette approche ne permet pas de prendre en compte la variabilité des données. Selon la quantité qu'on souhaite estimer, l'estimateur peut être biaisé.

Comparaison des méthodes d'imputation simple

On considère

- n observations indépendantes $\mathbf{X}_1, \ldots, \mathbf{X}_n$ telles que $\mathbf{X}_i$ suit une loi normale centrée bivariée, de variance 1 et de corrélation ρ.
- X_{i1} est toujours observé et $\mathbb{P}[R_{i2} = 0 \mid \mathbf{X}_i] = (1 + e^{1 - X_{i1}})^{-1}$.

On considère trois méthodes d'imputation

- imputation par l'espérance.
- imputation par l'espérance conditionnelle.
- imputation stochastique.

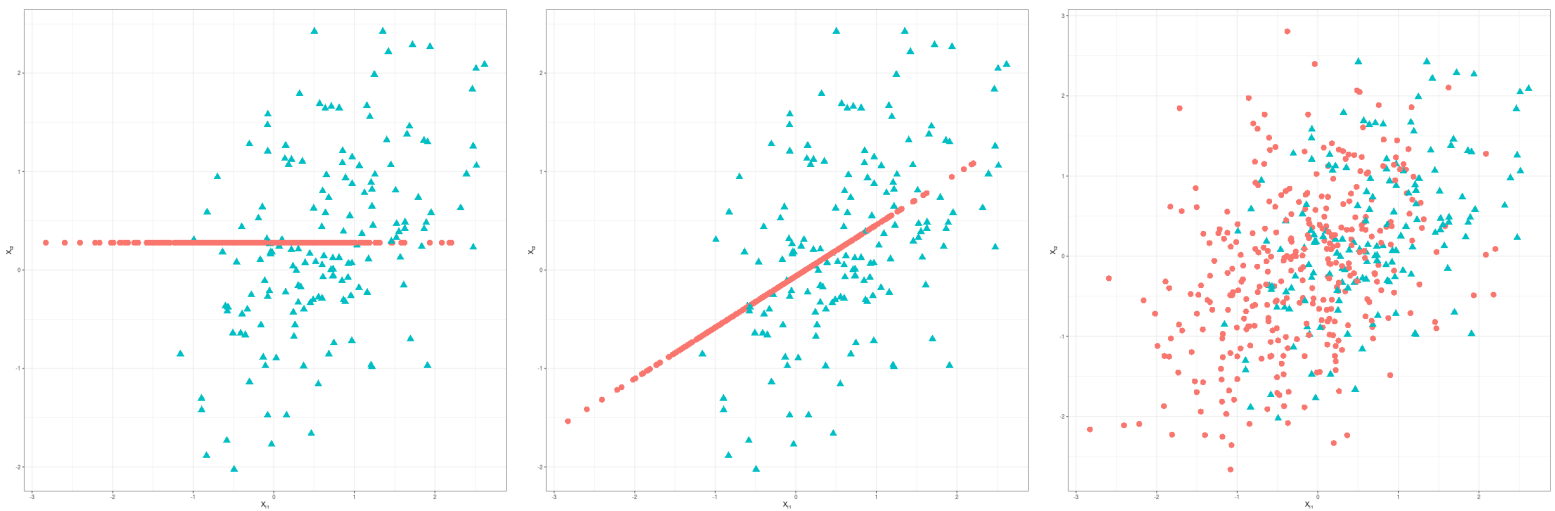


Figure – Jeu de données avec trois imputations différentes : imputation par l'espérance (gauche), imputation par l'espérance conditionnelle (centre), imputation stochastique par la loi conditionnelle (droite). Les données complètes sont en bleu et les données imputées en rose.

Méthodes d'imputations multiples

Les méthodes d'imputation stochastiques permettent d'éviter le problème d'incertitude sur la valeur réelle de la variable à imputer.

Il est préférable de considérer plusieurs imputations du même jeu de données. On parle alors d'imputations multiples.

L'imputation multiple est constituée de trois étapes.

1. On génére M tableaux de données complets à partir de $\mathbf{X}^{\text{obs}}$ et de générations aléatoires de $\mathbf{X}^{\text{miss}}$ (généré selon $\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \boldsymbol{R}$).
2. On calcule, sur chaque tableau imputé, l'estimateur de la quantité d'intérêt.
3. On agrège les estimateurs pour obtenir l'estimateur final.

Remarquons que la génération de $\mathbf{X}^{\text{miss}}$ est plus confortable sous un mécanisme MAR (elle se résume à générer selon $\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}$).

Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Objectif

Le but de l'analyse en composantes principales (ACP) est d'obtenir une représentation approchée d'un nuage de points de $\mathbb{R}^p$ dans un sous-espace de dimension plus faible.

On dispose d'un triplet (X, D, M) où :

- $X \in \mathbb{R}^{n \times p}$ est la matrice des données
- D est la matrice des poids
- M la métrique..

L'idée est de projeter le nuage de points de $\mathbb{R}^p$ dans des sous-espaces affines (droites, plans,...) de dimension réduite. Cette projection doit être fidèle..

Formulation du problème

Idée : la projection doit être fidèle..

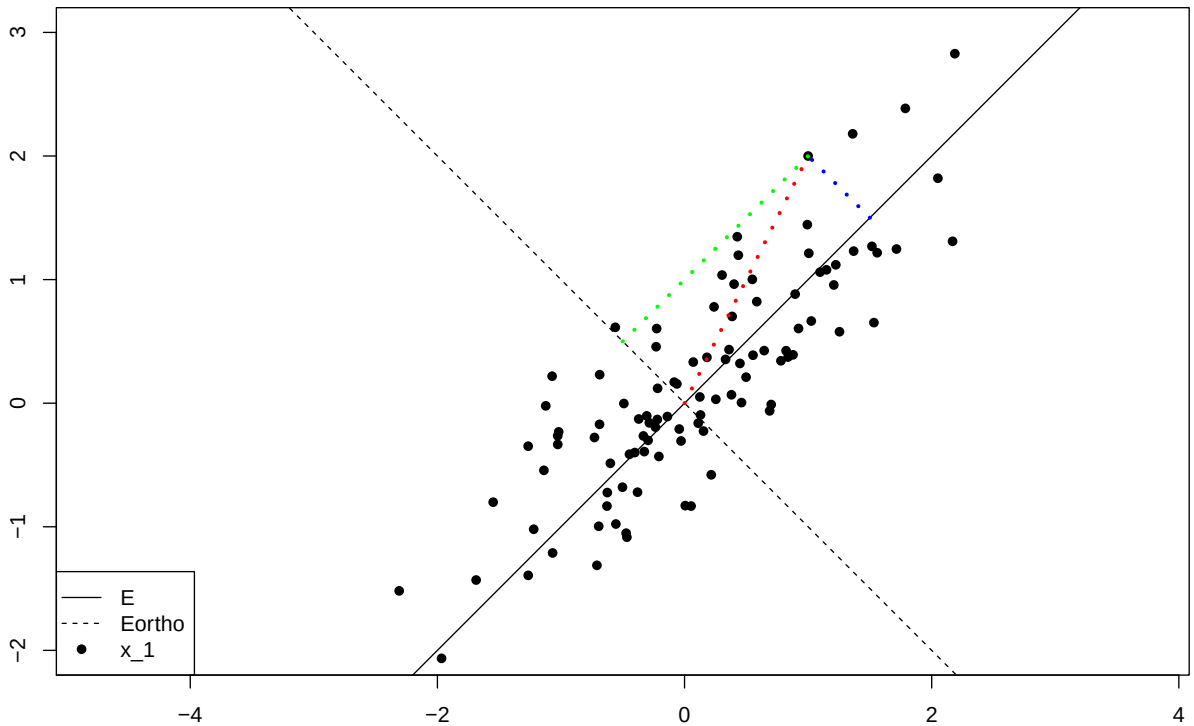
Formulation minimiser les écarts entre les points de N et leurs projections. Cela fait donc appel à l'inertie projetée du nuage..

Il faut trouver le sous-espace affine E_k de dimension k ($k < p$) tel que I_{E_k} (inertie du nuage N par rapport à E_k) soit minimale avec

$$I_{E_k} = \sum_{i=1}^{n} p_i d_M^2(x_i, E_k).$$

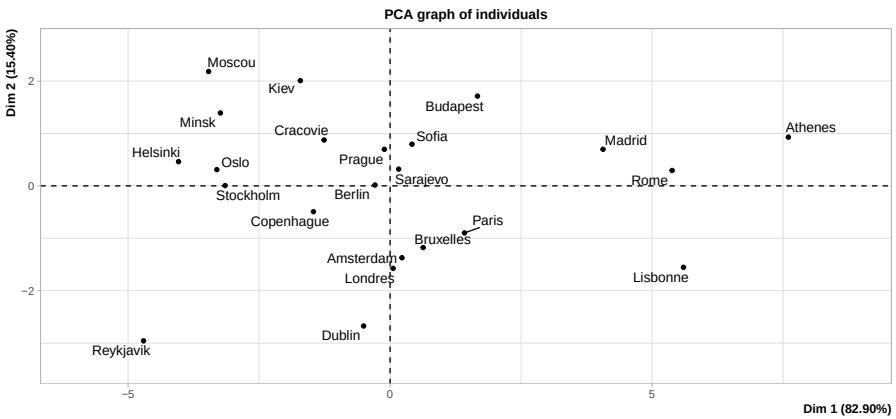
.

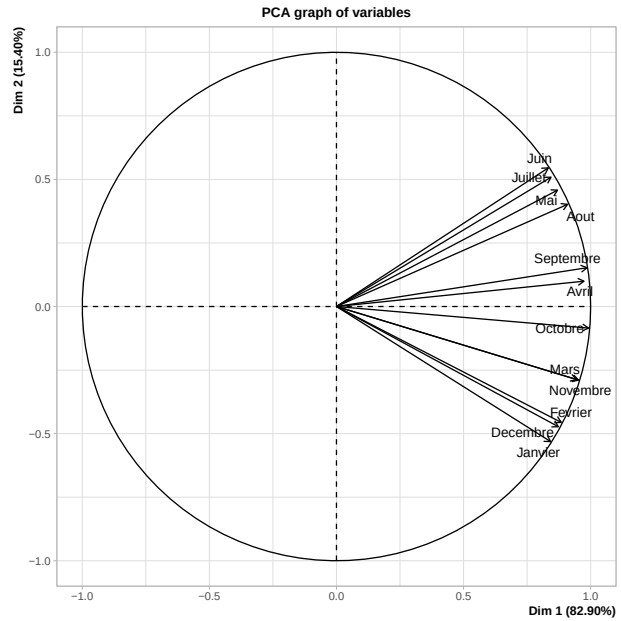
Représentation fidèle



Exemple

On souhaite visualiser les données des températures des différentes villes européénnes.





Formule de reconstitution

Lorsque l'on conserve tous les axes factoriels non triviaux, l'ACP nous donne une nouvelle base pour exprimer l'observation x_i.

En effet, dans l'ancienne base canonique de $\mathbb{R}^p$ notée $(e_1, \ldots, e_p)$, on a

$$x_i = \sum_{j=1}^{p} x_{ij} e_j .$$

Dans la nouvelle base données par l'ACP, on a

$$x_i = \sum_{k=1}^{r} c_{ik} u_k .$$

Formule de reconstitution

Ainsi, en notant que par propriété des vecteurs propres $(MU)^{-1} = U^\top$, on a

$$X = CU^\top = \sum_{k=1}^{r} c_{ik}\, u_k^\top .$$

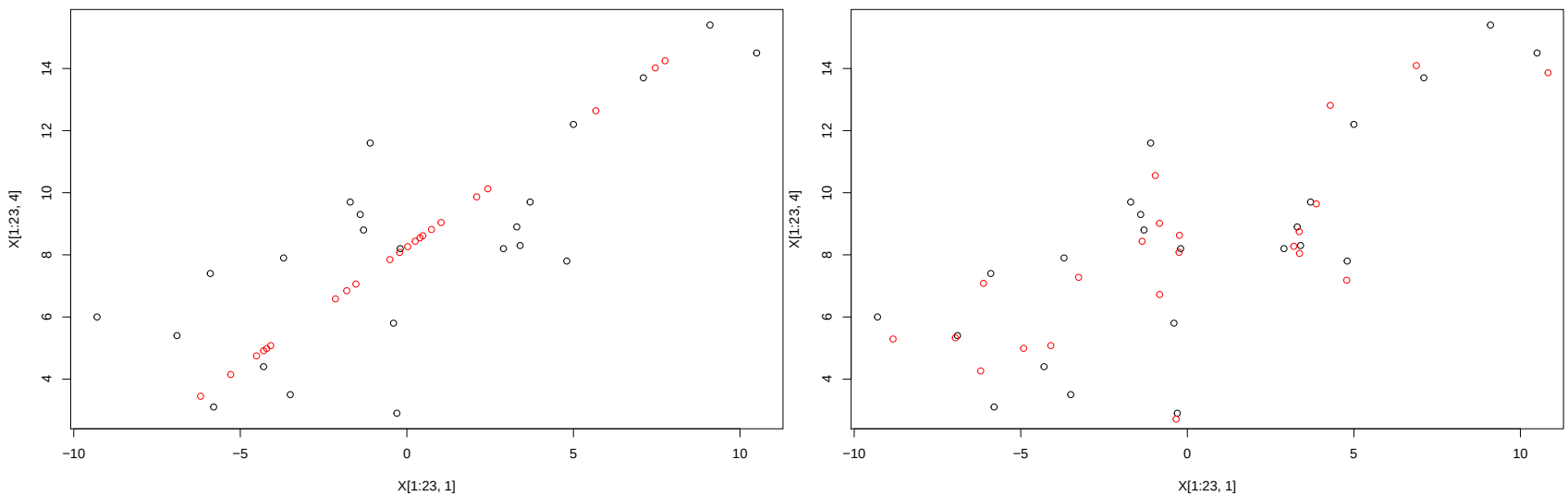
Si on considère les s composantes principales avec $s < r$, on peut reconstruire une approximation de la matrice X, notée $\tilde{X}$ qui correspond aux composantes principales projetées dans l'espace de départ

$$X \approx \tilde{X} = \sum_{k=1}^{s} c_{ik}\, u_k^\top .$$

Formule de reconstitution

On considère un jeu de données composé de 100 observations décrites par 4 variables.

On compare les données originales et leurs reconstructions basées sur le premier axe factoriel (gauche) et sur les deux premiers axes factoriels.



Approches naïves

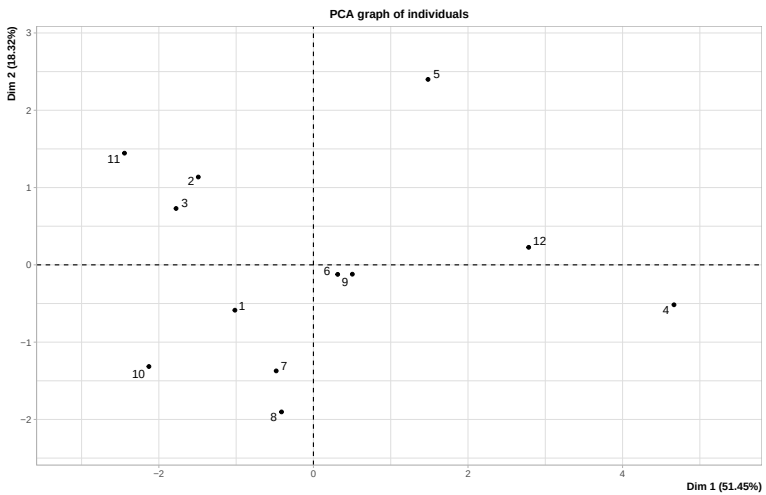
```
library(missMDA)
data(orange)
orange[1:6,1:4]
```

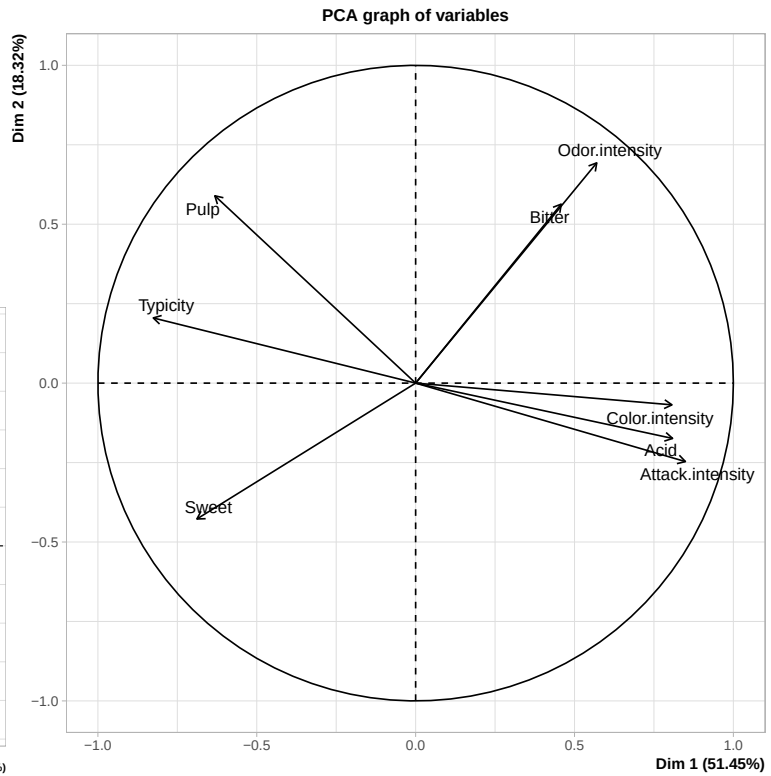
```
  Color.intensity Odor.intensity Attack.intensity    Sweet
1        4.791667       5.291667                NA       NA
2        4.583333       6.041667          4.416667 5.458333
3        4.708333       5.333333                NA       NA
4        6.583333       6.000000          7.416667 4.166667
5              NA       6.166667          5.333333 4.083333
6        6.333333       5.000000          5.375000 5.000000
```

Approches naïves

- Suppression des individus ayant des valeurs manquantes
- Imputation par la moyenne

```
require(FactoMineR)
out <- PCA(orange, graph=F)
plot(out, choix="ind")
plot(out, choix="var")
```





ACP itérative

Idées :

- Si deux variables X et Y sont fortement corrélées : on peut imputer une valeur manquante sur Y grâce à la valeur prise par l'individu sur X.
- Si deux individus ont des valeurs proches sur toutes les variables observées, on peut imputer une valeur manquante d'un individu par la valeur prise par l'autre individu pour la même variable.

Le but de l'ACP itérative est de prendre en compte à la fois les ressemblances globales entre les individus et les liaisons entre les variables.

ACP itérative

- Initialisation : imputation arbitraire
- Tant que la convergence n'est pas atteinte :
  1. ACP sur le tableau complété à partir de S dimension conservées
  2. Données manquantes imputées par ACP
  3. Statistiques mise à jour (moyennes et écarts-types)

En pratique, il est nécessaire de régulariser l'ACP pour éviter un problème de surajustement causé par une surestimation des liaisons entre variables provoquée par les valeurs manquantes.
Attention, cette méthode ne peut s'appliquer que pour un mécanisme MCAR ou MAR.

ACP itérative avec R

```
require(missMDA)
data(orange)
nb <- estim_ncpPCA(orange, scale=TRUE) # choix de S
comp <- imputePCA(orange, ncp=nb$ncp, scale = TRUE)
head(orange[,1:4])
```

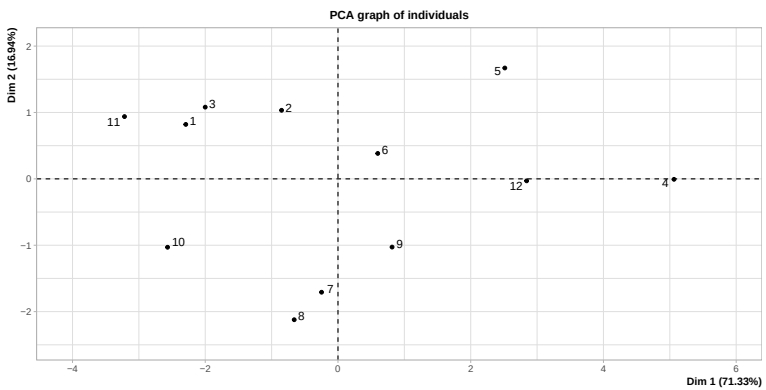
```
  Color.intensity Odor.intensity Attack.intensity    Sweet
1        4.791667       5.291667                NA       NA
2        4.583333       6.041667          4.416667 5.458333
3        4.708333       5.333333                NA       NA
4        6.583333       6.000000          7.416667 4.166667
5              NA       6.166667          5.333333 4.083333
6        6.333333       5.000000          5.375000 5.000000
```

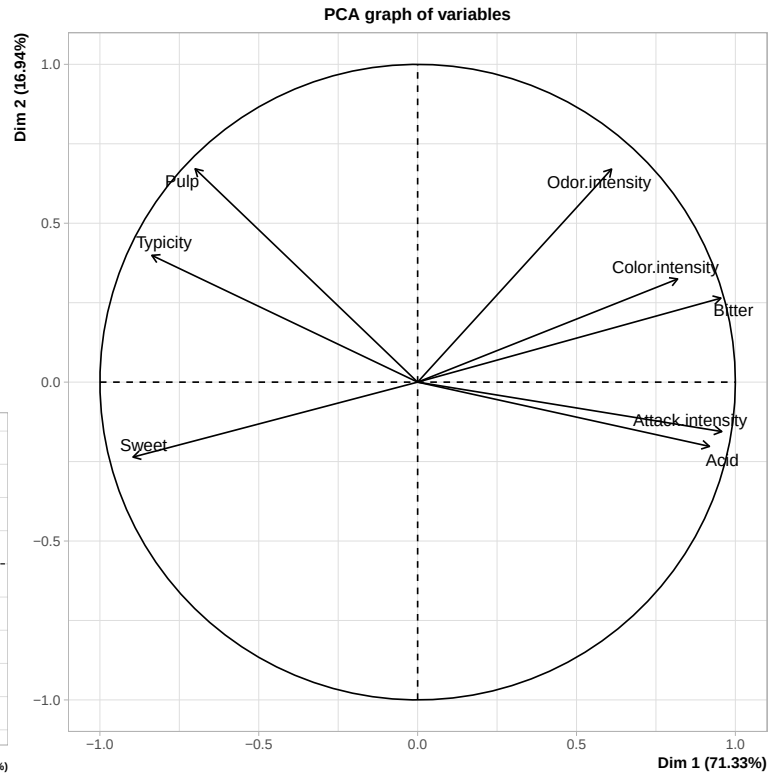
```
head(comp$completeObs[,1:4])
```

```
  Color.intensity Odor.intensity Attack.intensity    Sweet
1        4.791667       5.291667          4.077034 5.527352
2        4.583333       6.041667          4.416667 5.458333
3        4.708333       5.333333          4.158054 5.442936
4        6.583333       6.000000          7.416667 4.166667
5        6.271605       6.166667          5.333333 4.083333
6        6.333333       5.000000          5.375000 5.000000
```

ACP itérative avec R

```
res.pca <- PCA(comp$completeObs)
```





ACP itérative : remarques

Par nature, cette méthode renforce les liaisons entre les variables.

On fera attention si on effectue une ACP sur les données imputées : l'inertie des premiers axes est surestimée.

Il n'y a pas d'information sur l'incertitude de l'imputation. Ici une donnée imputée ressemble à une vraie donnée ! ! ! !

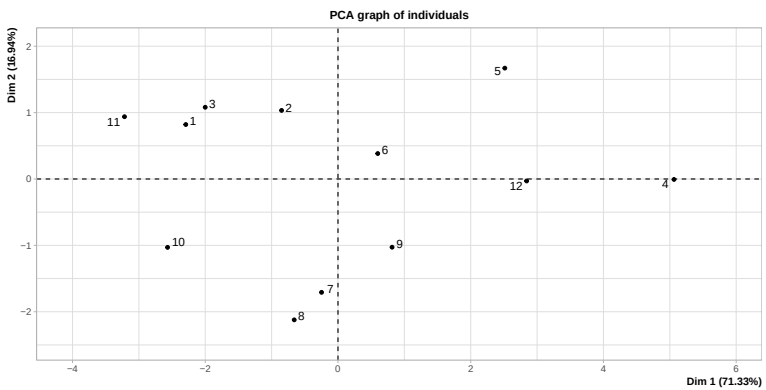
Une valeur unique ne peut pas refléter la variabilité de la prédiction.

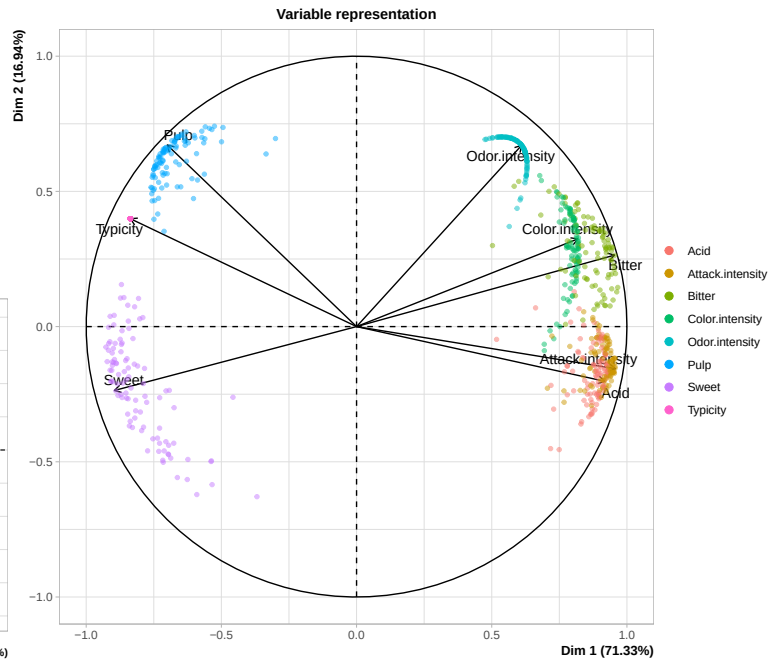
On a besoin d'avoir une information sur l'incertidue liée à l'imputation.

On utilise alors l'imputation multiple (génération de plusieurs valeurs plausibles pour chaque valeur manquante).

Visualisation de l'incertitude liée à l'imputation

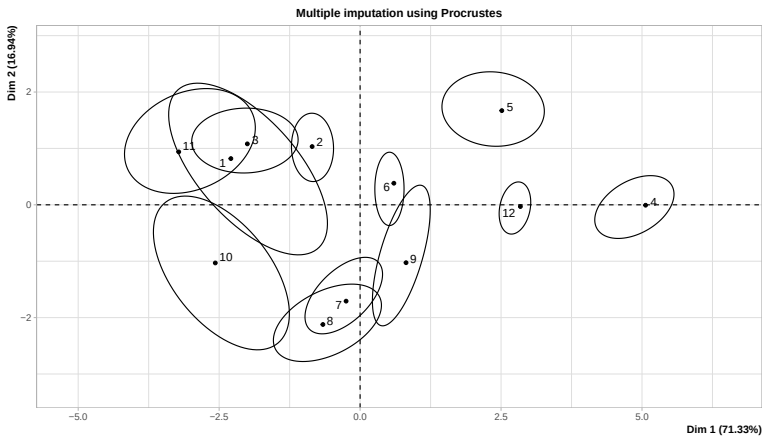
```
mi <- MIPCA(orange, scale=TRUE, ncp=2)
plot(mi, choice="ind.supp")
plot(mi, choice="var")
```

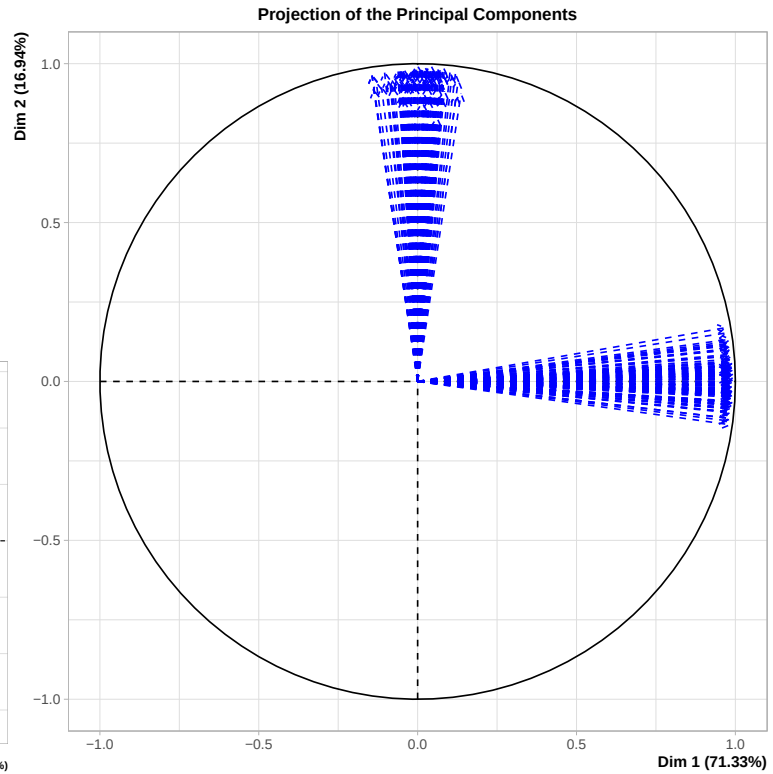




Visualisation de l'incertitude liée à l'imputation

```
plot(mi, choice="ind.proc")
plot(mi, choice="dim")
```





Terminologie

Description des schémas des données manquantes

Description des mécanismes de données manquantes

Ignorabilité et modélisation de la loi de l'ensemble des données

Méthodes d'analyse pour données manquantes

Analyse en Composantes Principales et imputation

Conclusion

Conclusion

Des données présentant des valeurs manquantes sont caractérisées par

- Un schéma.
- Un mécanisme.

Lorsque le mécanisme n'est pas ignorable, il faut modéliser la distribution de $\boldsymbol{R} \mid \mathbf{X}_i$.

Trois familles de méthodes d'estimation

- Méthodes basées sur la vraisemblance (Schafer [1997] et McLachlan and Krishnan [2007]).
- Méthodes basées sur des pondérations (Tsiatis [2007] et Efromovich [2018]).
- Méthodes basées sur des imputations (van Buuren [2018]).

Les références principales sont

- Présentation générale Little and Rubin [2019] et Molenberghs et al. [2014].
- Données longitudinales Daniels and Hogan [2008].
- Données d'enquêtes Rubin [1987] et Särndal and Lundström [2005].

Conclusion

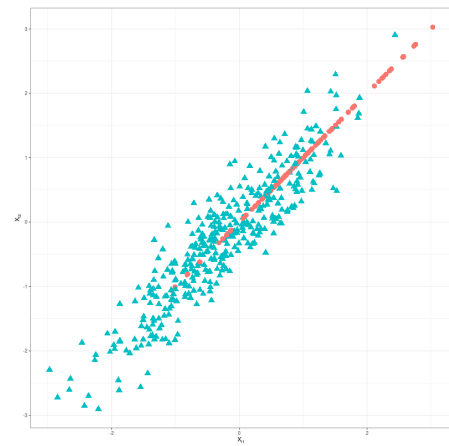
Les méthodes factorielles permettent une imputation (simple ou multiple) pour un mécanisme MCAR ou MAR.

Ces méthodes sont implémentées dans le package R `missMDA` (basé sur `FactoMineR`). Une documentation est disponible `http://factominer.free.fr/index_fr.html` et des tutoriels sont disponibles `https://husson.github.io/`

Attention, l'imputation d'un jeu de données est rarement un objectif final...

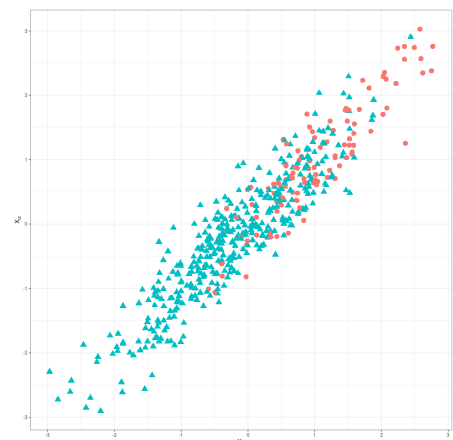
ACP itérative :

- Estimation de la matrice complétée par une ACP itérative régularisée
- Imputation des données manquantes



Multiples imputations :

- Génération de plusieurs matrices complétée en utilisant la règle de Bayes
- Imputation à partir de l'ACP en considérant un bruit Gaussien.



Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278) :200–203.

Chi, J. T., Chi, E. C., and Baraniuk, R. G. (2016). k-pod : A method for k-means clustering of missing data. *The American Statistician*, 70(1) :91–99.

Daniels, M. J. and Hogan, J. W. (2008). *Missing data in longitudinal studies : Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, Boca Raton.

Day, S. (1999). *Dictionary for clinical trials*. John Wiley & Sons, Hoboken.

Du Roy de Chaumaray, M. and Marbac, M. (2020). Clustering data with nonignorable missingness using semi-parametric mixture models. *arXiv preprint arXiv :2009.07662*.

Dunson, D. B. (2007). Bayesian methods for latent trait modelling of longitudinal data. *Statistical Methods in Medical Research*, 16(5) :399–415.

Dunson, D. B. and Perreault, S. D. (2001). Factor analytic models of clustered multivariate data with informative censoring. *Biometrics*, 57(1) :302–308.

Efromovich, S. (2018). *Missing and modified data in nonparametric estimation : with R examples*. CRC Press, Boca Raton.

Hafez, M. S., Moustaki, I., and Kuha, J. (2015). Analysis of multivariate longitudinal data subject to nonrandom dropout. *Structural Equation Modeling : A Multidisciplinary Journal*, 22(2) :193–201.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In Berg, S. V., editor, *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. National Bureau of Economic Research, Cambridge.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260) :663–685.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404) :1198–1202.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431) :1112–1121.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, 3^{e} edition.

Little, R. J., Rubin, D. B., and Zangeneh, S. Z. (2017). Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *Journal of the American Statistical Association*, 112(517) :314–320.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, 2e edition.

Miao, W., Ding, P., and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516) :1673–1683.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(2) :371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press, Boca Raton.

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media, Berlin.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Survey*. John Wiley & Sons, New-York.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons, Hoboken.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press, Boca Raton.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media, Berlin.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press, Boca Raton.