Gestion des données manquantes en clustering

Présenté par :

Fadela SADOU ZOULEYA
Mouhamadou Lamine NDAO

Encadré par :

Ndèye Niang
Vincent Audigier

29 Octobre 2021

Clustering

Données $\mathbf{Z} = z_{ij}$ $\begin{matrix} 1 & i & n \\ 1 & j & p \end{matrix}$ ensemble de données numériques

P partition en K classes où chaque i appartient à une seule classe C_i 2 f1, ..., Kg. Le centroïde de la classe k est noté c_k 2 $M_{1\ p}$

Objectif Identifer C_i pour chaque individu i à partir de $^1\mathbf{z}_i{}^{\circ}{}_{1\ 1\ i\ n^{\circ}}$

Approches existantes

basées sur distances

- k-means
- fuzzy C-means
- hierarchical clustering
- pam

basées sur modèles

- gaussian mixture models
- mixture of multivariate t-distributions

Gestion des données manquantes en clustering

Contraintes, $\mathbf{Z}$ souvent incomplet... $\mathbf{z}_i = {}^1\mathbf{z}_i^{obs}, \mathbf{z}_i^{miss}$o

Méthodes "Ad-hoc"

Analyse des cas complets

Imputation simple

Méthodes avancées

Imputation Multiple (IM)

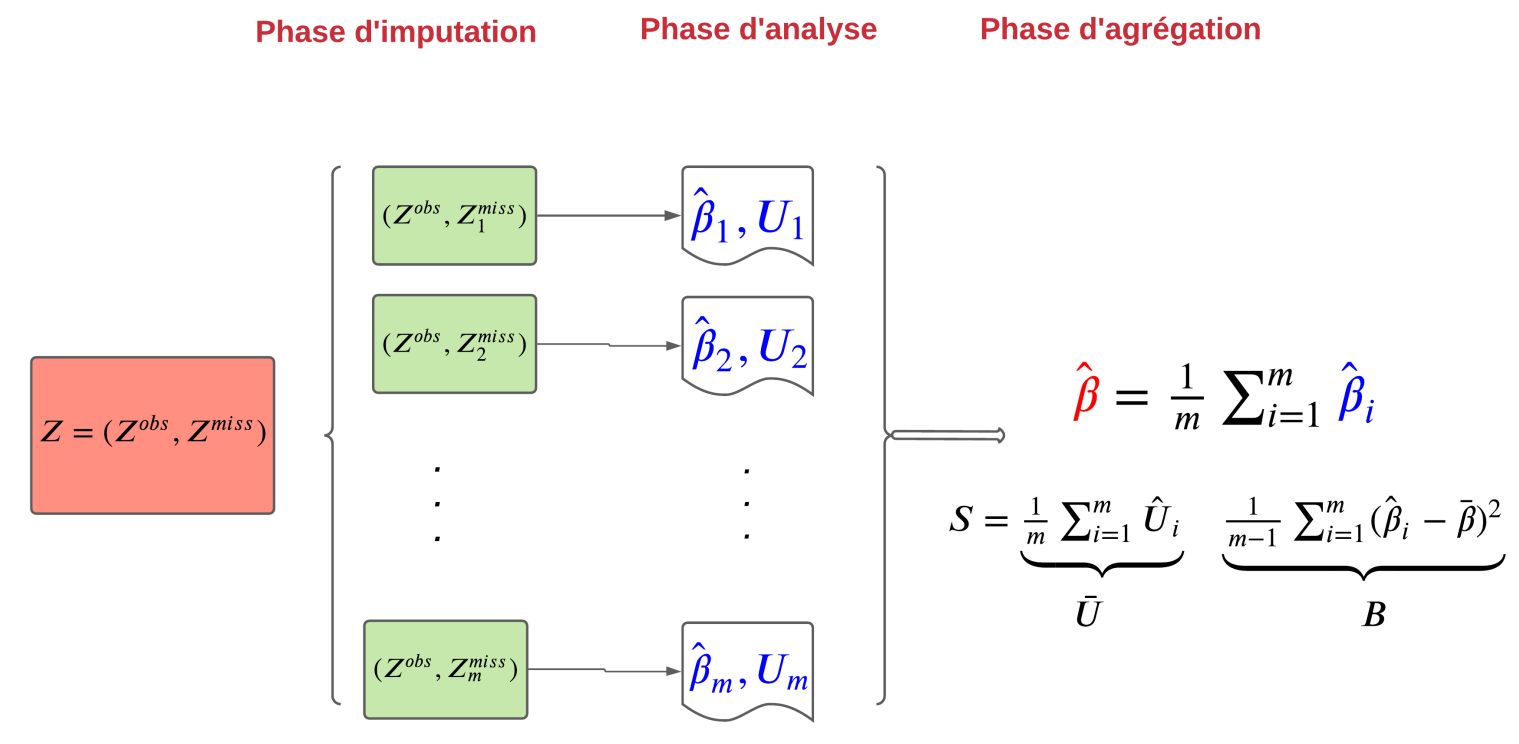
Méthodes directes (fuzzy C-means, k-Pod, ...)

Plan

Plan

Imputation multiple (Rubin, 1987) : cas de la régression

Principe



Estimation unique de β et sa variance : $S = \bar{U} \, , \, B$

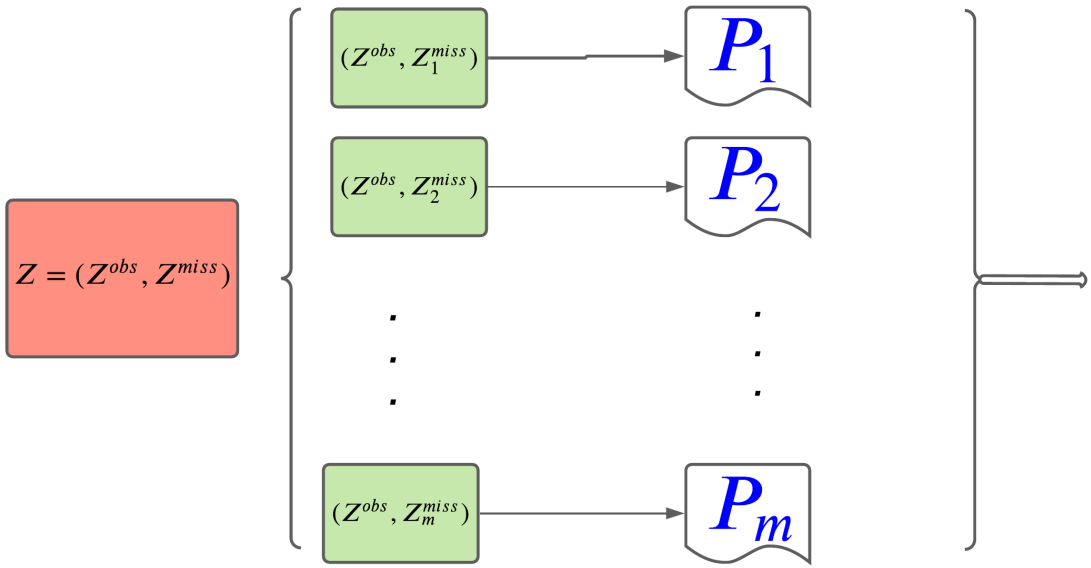
Qualité de $\hat{\beta}$ dépend de l'indépendance des tableaux imputés conditionnellement aux valeurs observées

IM pour le clustering (Audigier and Niang, 2020)

Phase d'imputation
Prise en compte de la structure en groupes (JM-DP)

Phase d'analyse
Appliquer un clustering (kmeans)

Phase d'agrégation
Consensus de partitions (extension des règles de Rubin)



$$S = \bar{U}^{boot} + B$$

Phase d'imputation : JM-DP (Kim et al., 2014)

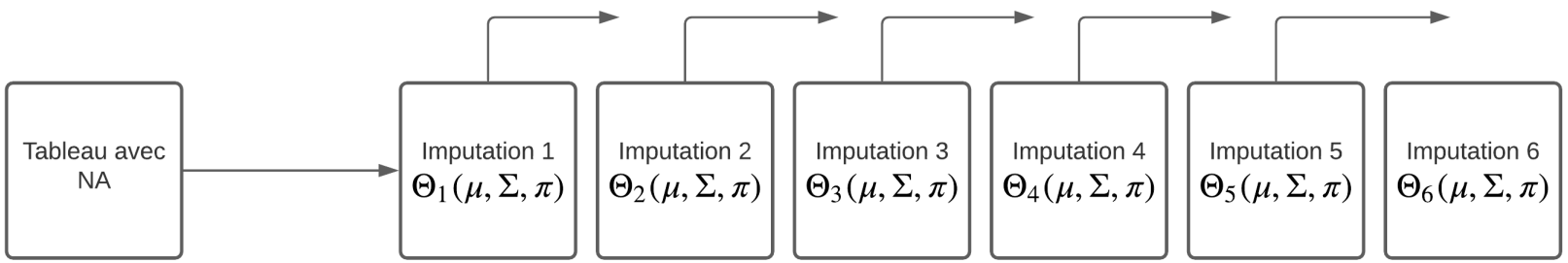
Approche d'imputation

On suppose que $Z_i | C_i, \mu, \quad N(Z_i | \mu_{C_i}, \ C_i)$

Principe :

1. Alterner le tirage des paramètres dans leur distribution a posteriori et l'imputation des NA selon les paramètres simulés.
2. Répéter jusqu'à convergence vers la distribution a posteriori des paramètres $(\mu, \ , \pi)$.

Imputation multiple : Répéter à nouveau (1) et (2) de façon à récupérer m réalisations des paramètres issus de la loi a posteriori.

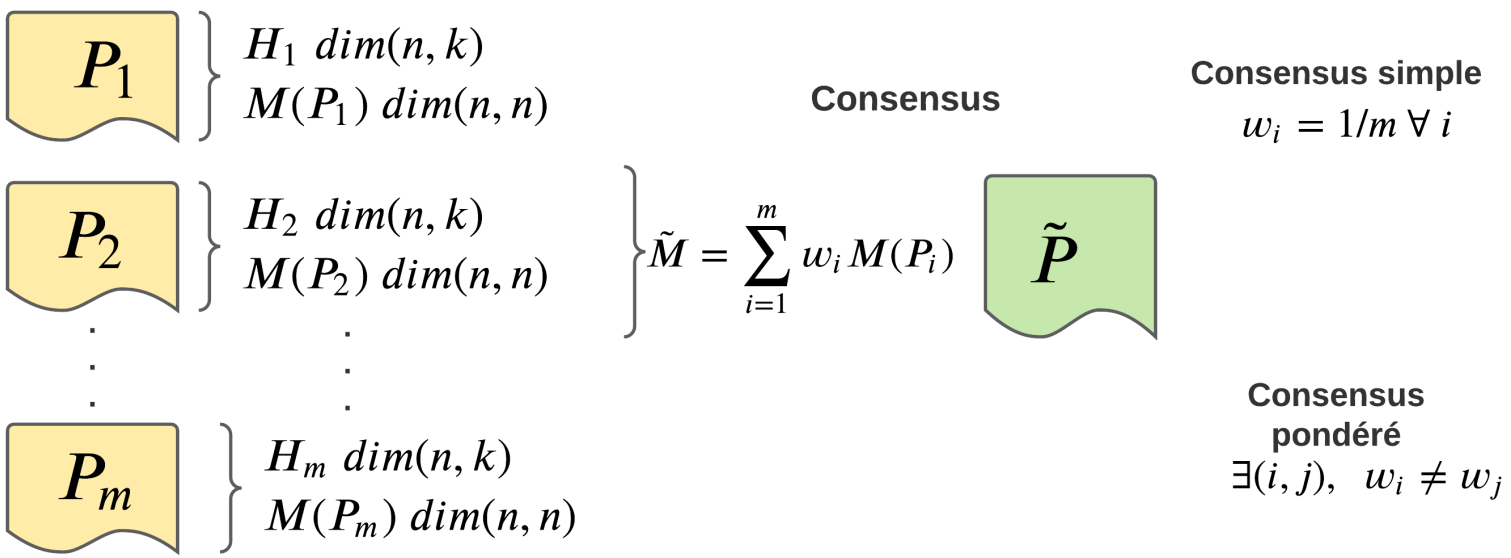


Phase d'agrégation : consensus de partitions

Consensus de partitions

Objectif : Identifier une partition compromis d'un ensemble de partitions obtenues sur le même ensemble d'observations.

Principe :



Approche simple : NMF (Tao Li and al., 2007)

Formulation :

$$\min_{H} \mathrm{k}\tilde{M} \quad HH^{0}\mathrm{k}^{2} \tag{1}$$

$H = \mathrm{f}0,1\mathrm{g}^{1nk^{\mathrm{o}}}$: matrice d'indicatrices
$M = \mathrm{f}0,1\mathrm{g}^{1nn^{\mathrm{o}}}$: matrice de connectivité.
En posant $\tilde{H} = H\,{}^{1}H^{0}H^{\mathrm{o}}\,{}^{1\bullet 2}$ et $D = \mathrm{diag}\,{}^{1}H^{0}H^{\mathrm{o}}$ (1) devient :

$$\min_{\tilde{H}^{0}\tilde{H}=I,\tilde{H}\ 0} \mathrm{k}\tilde{M} \quad \tilde{H}D\tilde{H}^{0}\mathrm{k}^{2} \tag{2}$$

$$\tilde{H} \quad \tilde{H}\ \mathrm{s}\frac{}{\frac{{}^{1}\tilde{M}^{0}\tilde{H}D^{\mathrm{o}}}{{}^{1}\tilde{H}\tilde{H}^{0}\tilde{M}^{0}\tilde{H}D^{\mathrm{o}}}} \tag{3}$$

$$D \quad D\ \mathrm{s}\frac{}{\frac{{}^{1}\tilde{H}^{0}\tilde{M}\tilde{H}^{\mathrm{o}}}{{}^{1}\tilde{H}^{0}\tilde{H}D\tilde{H}^{0}\tilde{H}^{\mathrm{o}}}} \tag{4}$$

Approche pondérée : Weighted-NMF (Tao Li and al., 2008)

Introduire w tel que :

$$w = {}^{1}w_1, w_2, \ldots, w_m{}^{o0}, \quad w_t \quad 0, \quad \mathrm{k}w\mathrm{k}_1 = \tilde{\mathbf{m}} \sum_{t=1} w_t = 1 \tag{5}$$

On définit $\tilde{M} = \acute{\mathrm{I}}_{t=1}^{m} w_t M\, {}^{1}P_t{}^{o}$ et le problème (1) devient :

$$\min_{w,\tilde{H}} \quad \tilde{M} \quad \tilde{H}\tilde{H}^{0\ 2} \tag{6}$$

1. Chercher $\tilde{H}$ en fixant w en utilisant NMF
2. Chercher w en fixant $\tilde{H}$ en résolvant $\min w^0 A w \quad 2b^0$, cte

Problématique

Les travaux sur l'agrégation des partitions après IM n'effectue que du consensus simple

Problème : par construction, les tableaux imputés ne sont jamais indépendants.

Idée : utiliser le consensus pondéré pour corriger les redondances des partitions.

Simulations

Données

30 expériences ont été simulées. Pour une expérience, nous avons :

Données

$n = 100$ observations, $p = 10$ variables

$K = 2$ clusters : $n_w = 50$, w 2 f1, 2g

moyennes μ_w, w 2 f1, 2g : $\mu_1 = {}^1 0, 0, 0, 0, 0, 0, 0, 0, 0, 0^o$

$\mu_2 = {}^1 0, 0, 0, 0, 0, 2, 2, 2, 2, 2^o$

Matrice variance covariance $= \begin{pmatrix} I_5 & 0 \\ 0 & \end{pmatrix}$

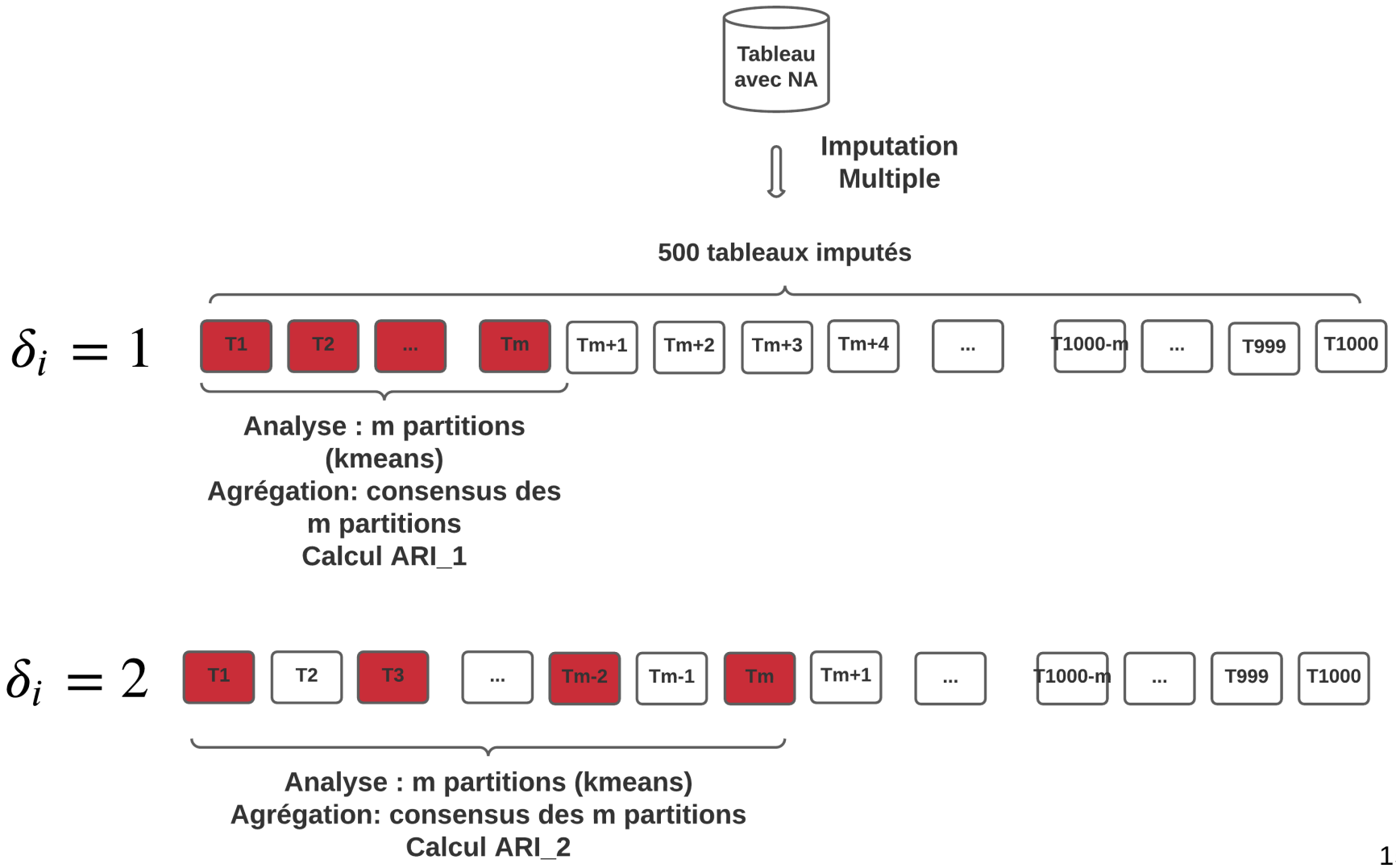
$$\begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{pmatrix}$$

30% de données manquantes selon un mécanisme MAR.

Plan

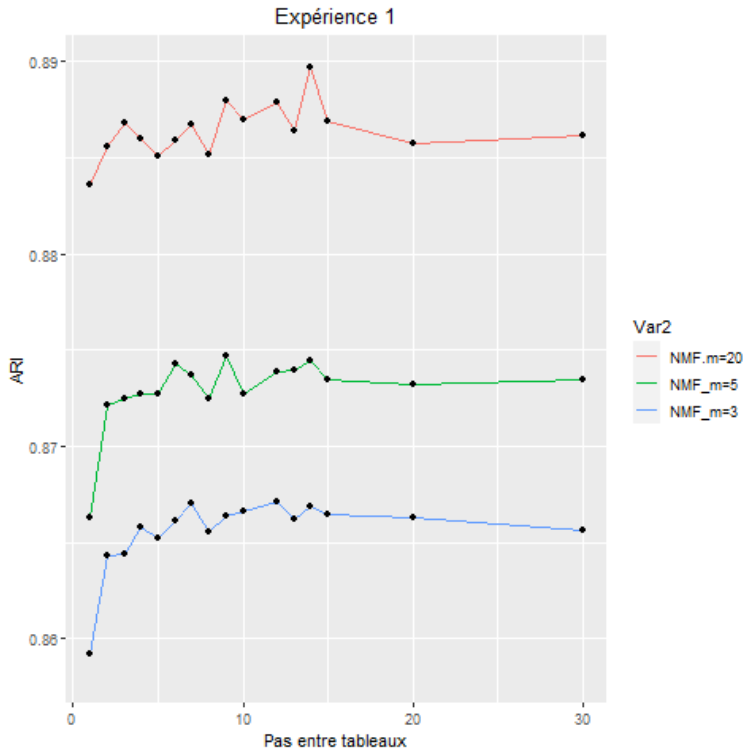
Plan d'expérimentation

m 2 f3, 5, 20g : nombre de tableaux

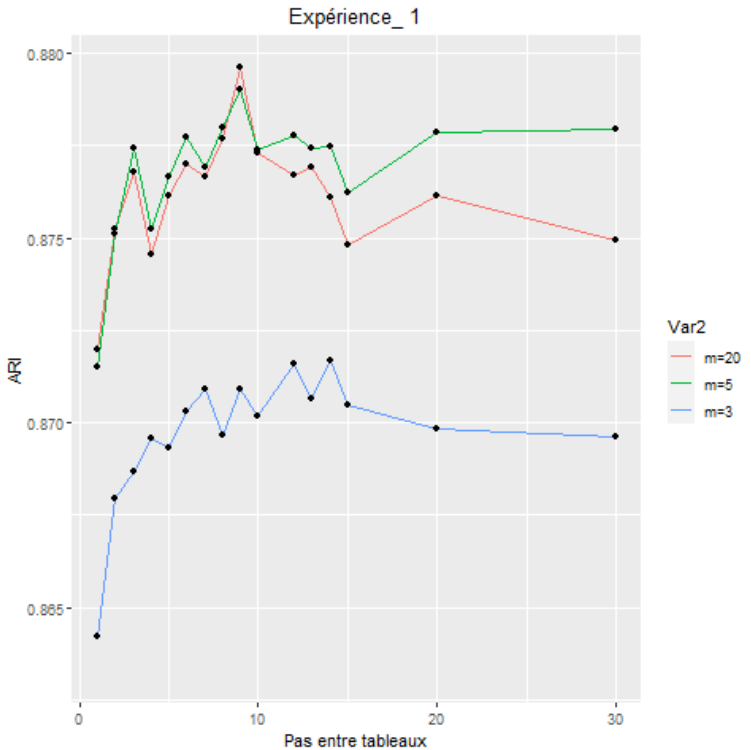


Résultats

Résultats NMF



Résultats WNMF



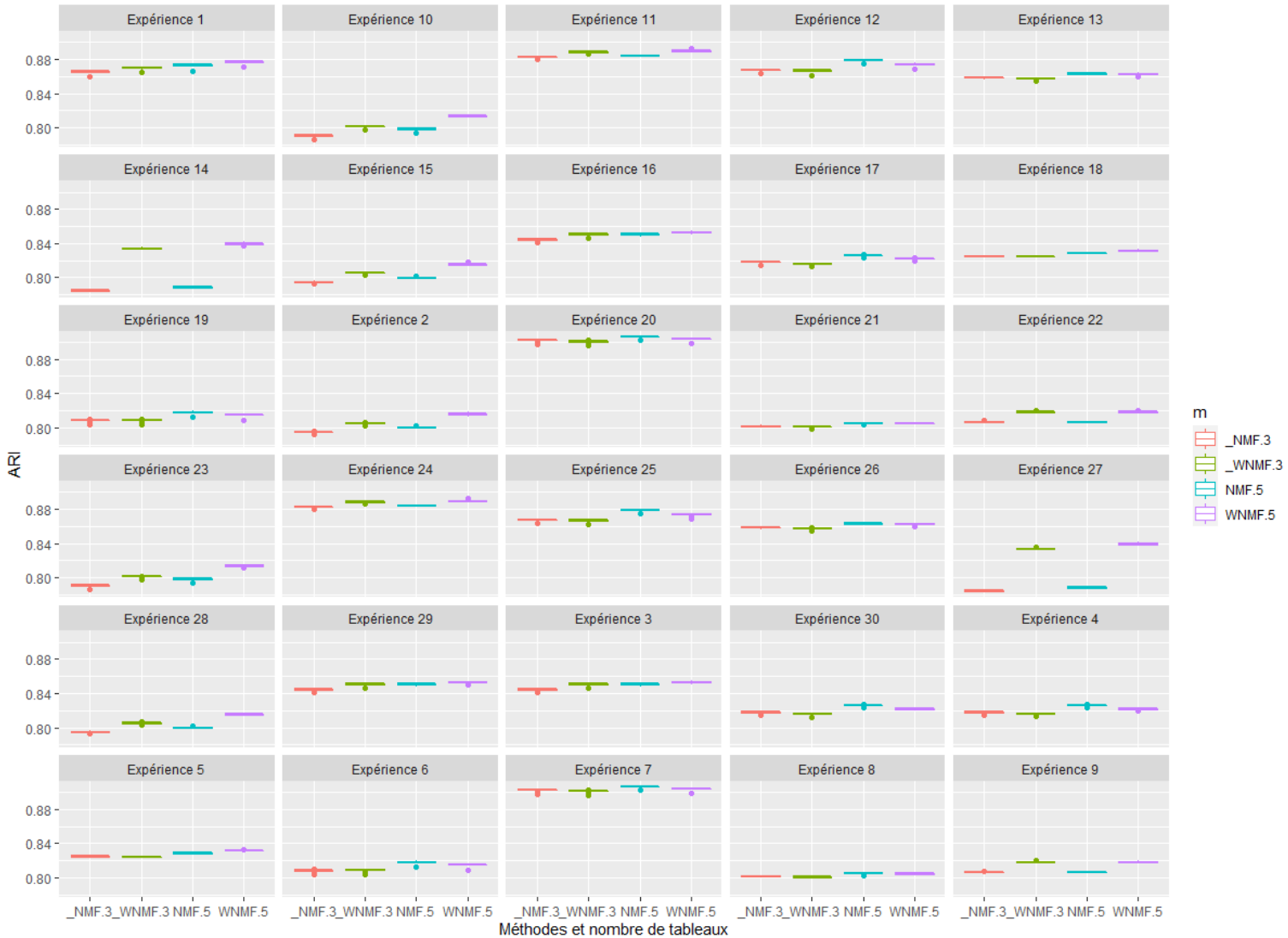
Résultats

Moins sensible à l'indépendance entre tableaux.

Le nombre de tableaux considéré améliore la qualité $\tilde{P}$.

Résultats

Comparaison des approches NMF et WNMF



Synthèse

Clustering après IM est beaucoup moins sensible à l'indépendance

En clustering après IM, en prenant un nombre relativement élevé de tableaux, il n'est pas nécessaire de s'assurer de l'indépendance entre les tableaux.

Pas de di□érences significatives, en termes de performances, entre les approches NMF et WNMF après IM.

Perspectives

L'apport du consensus pondéré dans une stratégie de stacking.

Évaluation de la qualité d'une partition consensus après IM

Plan

K-POD (Chi et al., 2016)

k-means

$$min_{c_1,\ldots,c_K,C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} d(Z_i, c_k)^2$$

( )

$$min_{H,C} \|Z - HC\|_F^2$$

k-POD

$$min_{c_1,\ldots,c_K,C_1,\ldots C_K} \sum_{k=1}^{K} \sum_{ij,i \in C_k \text{ et } ij \neq} (Z_{ij} - c_{kj})^2$$

( )

$$min_{H,C} \|P(Z) - P(HC)\|_F^2$$

Pas de solution explicite

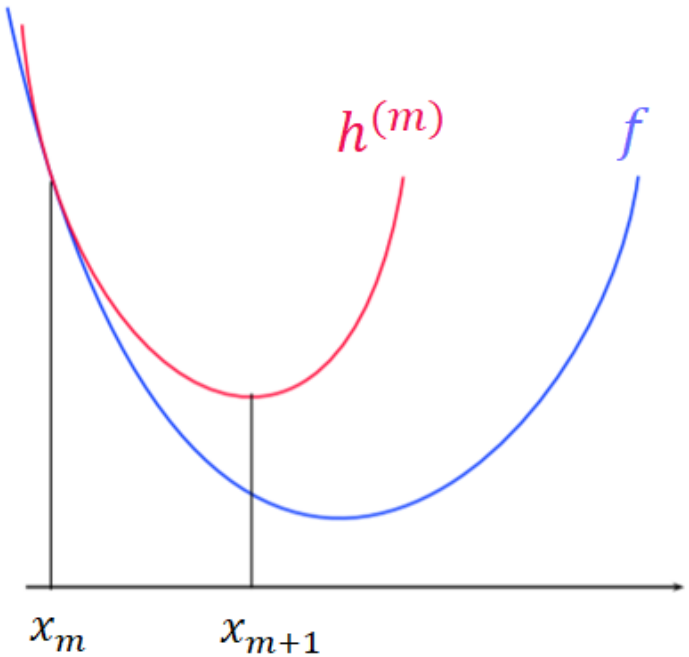
$H = [h_{ik}] \in M_{n,K}$! $h_{ik} = \begin{cases} 1 & \text{si } i \in C_k \\ 0 & \text{sinon} \end{cases}$

$\{1,\ldots,n\} \quad \{1,\ldots,p\}$! sous-ensemble des indices des valeurs observées

P ! $\tilde{P}(Z)_{ij} = \begin{cases} z_{ij} & \text{si } (i,j) \in \\ 0 & \text{si } (i,j) \in \end{cases}$ c .

K-POD (Chi et al., 2016)

Algorithme de majoration-minimisation



La fonction de perte $f : x \; ! \; f\,{}^{1}x^{o}$ est majorée à l'itération m par la fonction $h^{1}m^{o}$.

K-POD (Chi et al., 2016)

Fonction de perte $f\,{}^{1}H, C^{o}$

$$\text{jj } P \;\; {}^{1}Z^{o} \quad P \;\; {}^{1}HC^{o} \text{ jj}_F^2$$

Fonction majorante $g^{\,{}^{1}q^{o}}$ en ${}^{1}H\,{}^{1}q^{o}, C\,{}^{1}q^{o}\,o$

$$\text{jj } P \;\; {}^{1}Z^{o} \quad P \;\; {}^{1}HC^{o} \text{ jj}_F^2 \;\text{¸ jj } P \;\; {}_{C}\,{}^{1}Z^{o} \quad P \;\; {}_{C}\,{}^{1}H\,{}^{1}q^{o}C\,{}^{1}q^{o}\,o \text{ jj}_F^2$$

=

$$\text{jj } Z^{\,{}^{1}q^{o}} \quad HC \text{ jj}_F^2$$

K-POD (Chi et al., 2016)

Pour le nombre K de clusters fixé :

Initialiser les valeurs manquantes et obtenir la base complétée Z^{1^0}, puis calculer la partition initiale ${}^1H^{1^0}, C^{1^0}{}^{\mathrm{o}}$ par k-means ;

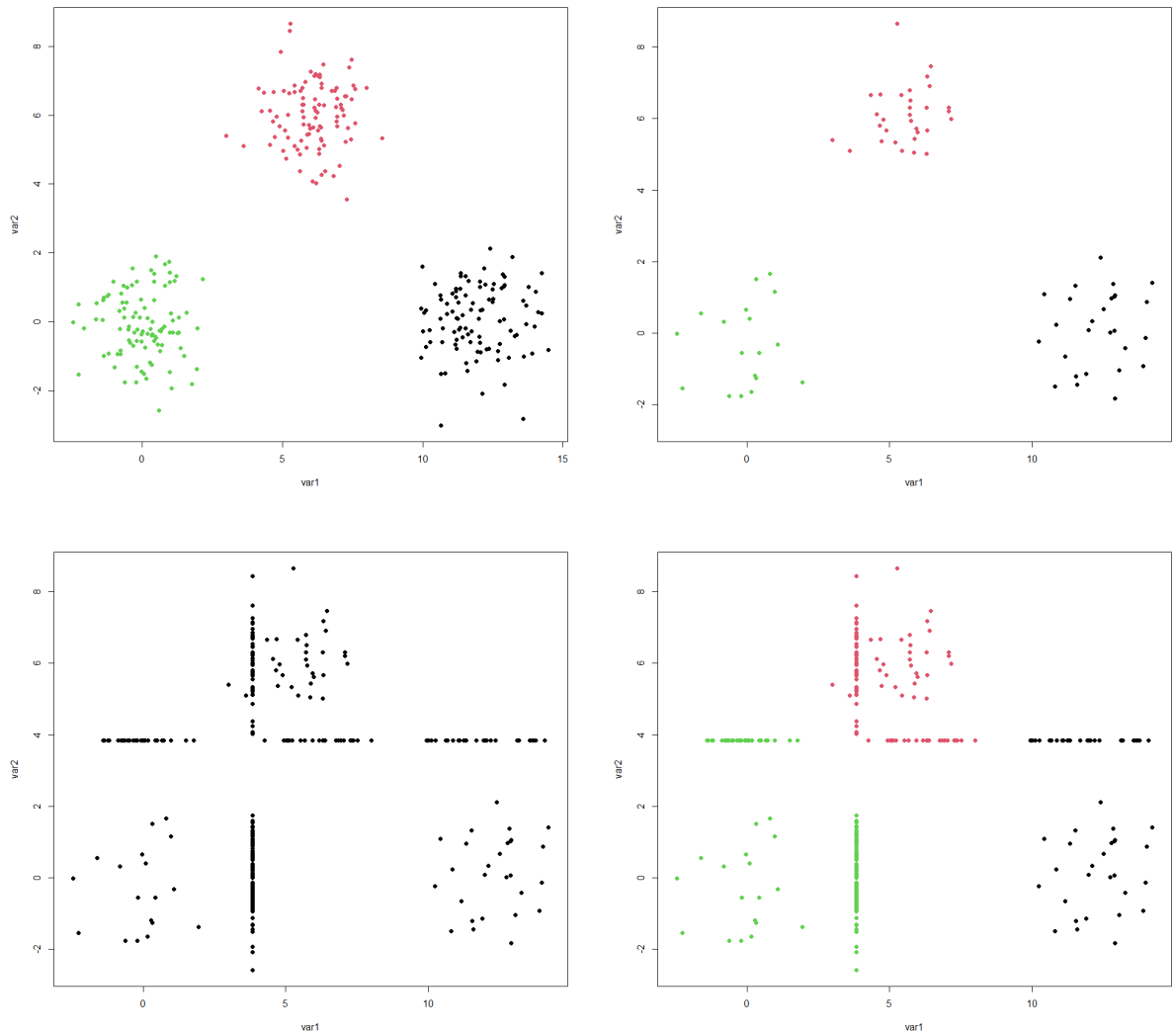
Répéter pour les itérations $q = 0, 1, 2, \ldots$ jusqu'à convergence :

remplacer les valeurs manquantes par celles des centroïdes de leur groupe respectif pour obtenir Z^{1^q} ;

estimer le minimum ${}^1H^{1^q,1^0}, C^{1^q,1^0}{}^{\mathrm{o}}$ de la fonction majorante $g^{1^q}({}^1H, C | H^{1^q}, C^{1^q}{}^{\mathrm{o}})$ par k-means sur Z^{1^q}.

K-POD (Chi et al., 2016)

Initialisation des valeurs manquantes



Ignorable-GMM (Marbac et al., 2019)

Notation O_i $1, \ldots, p$ l'ensemble des indices des variables pour lesquelles l'individu i est observé.

Si l'individu i appartient au cluster C_k, alors :

GMM :

$$Z_i | c_i = k \quad g_k(.|\theta_k) = N(.; \mu_k, \Sigma_k)$$

Ignorable-GMM :

$$Z_i | c_i = k \quad N(.; \mu_k, \Sigma_k) = \prod_{j \in O_i} N(z_{ij}; \mu_{kj}, \sigma_{kj}^2)$$

Log-vraisemblance ignorable-GMM :

$$l(Z, \theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \prod_{j \in O_i} N(z_{ij}; \mu_{kj}, \sigma_{kj}^2)$$

Fuzzy c-means

Degré d'appartenance/probabilité d'appartenir à chaque cluster

$$\underset{c_1,\ldots,c_K,\gamma_{11},\ldots,\gamma_{nK}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_{ik}^{m} \, \| x_i - c_k \|_2^2$$

Pour le nombre de clusters K et le paramètre de flou α fixés :

Attribuer aléatoirement des coefficients γ_{ik} à chaque individu ;

Répéter jusqu'à convergence :

Calculer les centroïdes $c_{kj} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{\alpha} . x_{ij}}{\sum_{i=1}^{n} \gamma_{ik}^{\alpha}}$

Recalculer les coefficients $\gamma_{ik} = \frac{d(x_i, c_k)^{\frac{1}{\alpha - 1}}}{\sum_{l=1}^{K} d(x_i, c_l)^{\frac{1}{\alpha - 1}}}$

Fuzzy c-means : Optimal Completion Strategy (Hathaway et Bezdek, 2001)

Pour le nombre de clusters K et le paramètre de flou α fixés :

Initialiser les centroïdes $c_{k,k=1,\dots,K}$ et initialiser les valeurs manquantes par celles du centroïde le plus proche au sens de la distance locale ;

Répéter pour les itérations $q = 1, 2, \dots$ jusqu'à convergence :

Calculer les coefficients $y_{ik}^{(q)} = \frac{d(x_i, c_k^{(q-1)})^{\frac{1}{\alpha-1}}}{\sum_{l=1}^{K} d(x_i, c_l^{(q-1)})^{\frac{1}{\alpha-1}}}$

Calculer les nouveaux centroïdes de chacun des K clusters

$$c_{kj}^{(q)} = \frac{\sum_{i=1}^{n} y_{ik}^{\alpha} . x_{ij}}{\sum_{i=1}^{n} y_{ik}^{\alpha}}$$

Calculer les valeurs manquantes $x_{ij}^{(q,1)} = \frac{\sum_{k=1}^{K} y_{ik}^{(q-1)\alpha} . c_{kj}^{(q)}}{\sum_{k=1}^{K} y_{ik}^{(q-1)\alpha}}$

Simulations

Configurations de données simulées

Configuration de référence (modèle I)

$n = 750$ observations, $p = 8$ variables

$K = 3$ clusters : $n_w = 250, w\ 2\ \{1, 2, 3\}$

moyennes $\mu_w, w\ 2\ \{1, 2, 3\}$) : $\mu_1 = {}^{1}0, 0, 0, 0, \quad , \quad , 0, \quad {}^{2o}$

$\mu_2 = {}^{1}0, 0, 0, 0, \quad , \quad , \quad , 0^{o}$

$\mu_3 = {}^{1}0, 0, 0, 0, \quad , \quad , \quad , \quad {}^{2o}$

Matrices de variance-covariance

$$_{w,w=1,2,3} = \quad {}^{1}\rho^{o} = \begin{pmatrix} I_4 & 0 \\ 0 & \begin{matrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{matrix} \end{pmatrix} \text{ avec } \rho = 0.3.$$

Simulations

Pourcentages de valeurs manquantes : $\tau = 10\%, 25\%, 40\%$

Mécanismes de valeurs manquantes :

MCAR [a] : $Prob(r_{ij} = 0) = \tau$

MAR [b] 1 : $Prob(r_{ij} = 0) = \quad (a_\tau \ , \ x_{i1})$

MAR 2 : $Prob(r_{ij} = 0) = \quad (a_\tau \ , \ x_{i8})$

---

a. Missing Completely At Random
b. Missing At Random

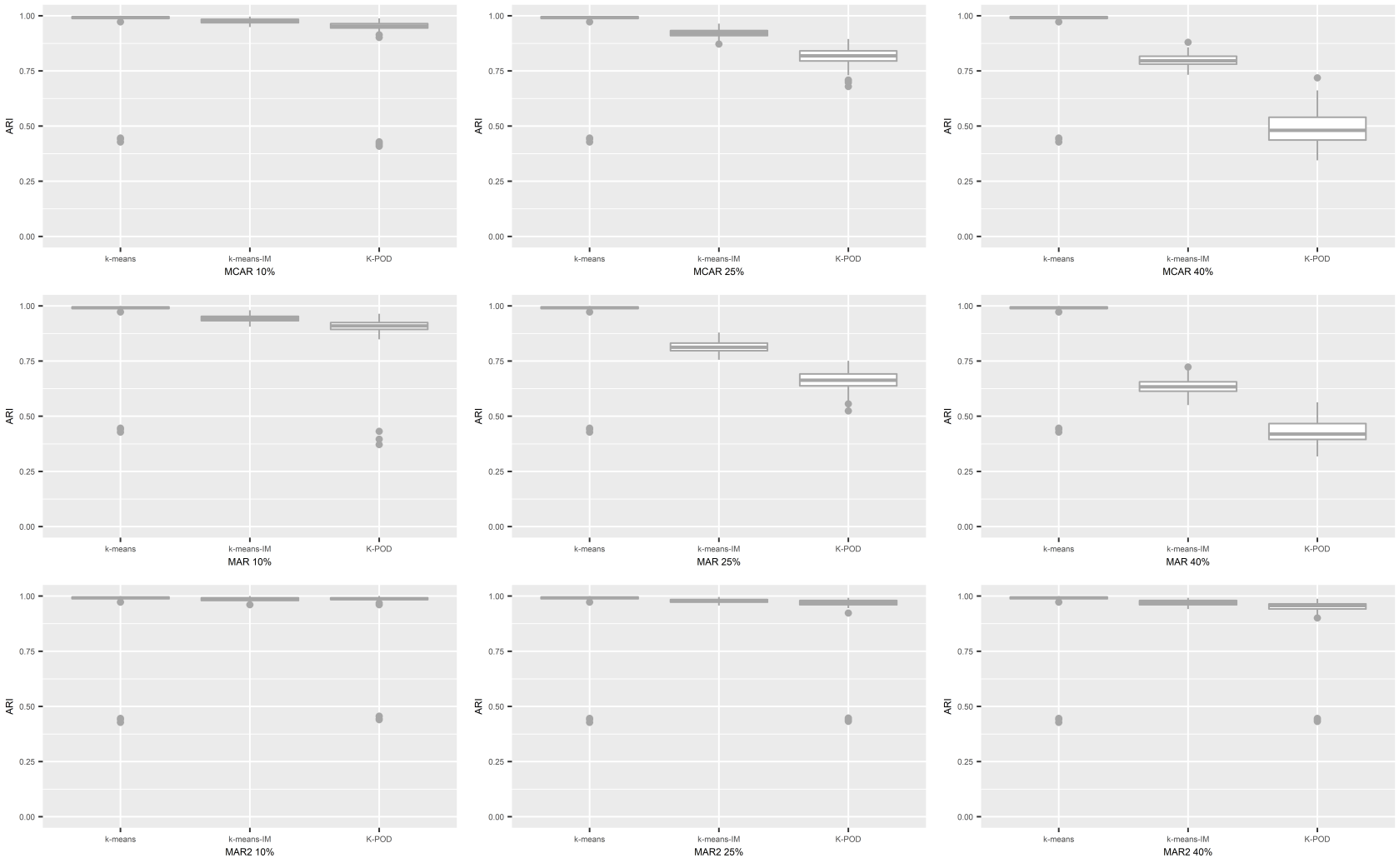
la fonction de répartition de la loi normale standard

τ une constante contrôlant la proportion de valeurs manquantes attendue

$R = (r_{ij})_{1 \ i \ n, 1 \ j \ p}$ le dispositif des valeurs manquantes, $r_{ij} = 1$ si x_{ij} est observé et $r_{ij} = 0$ si x_{ij} est manquant

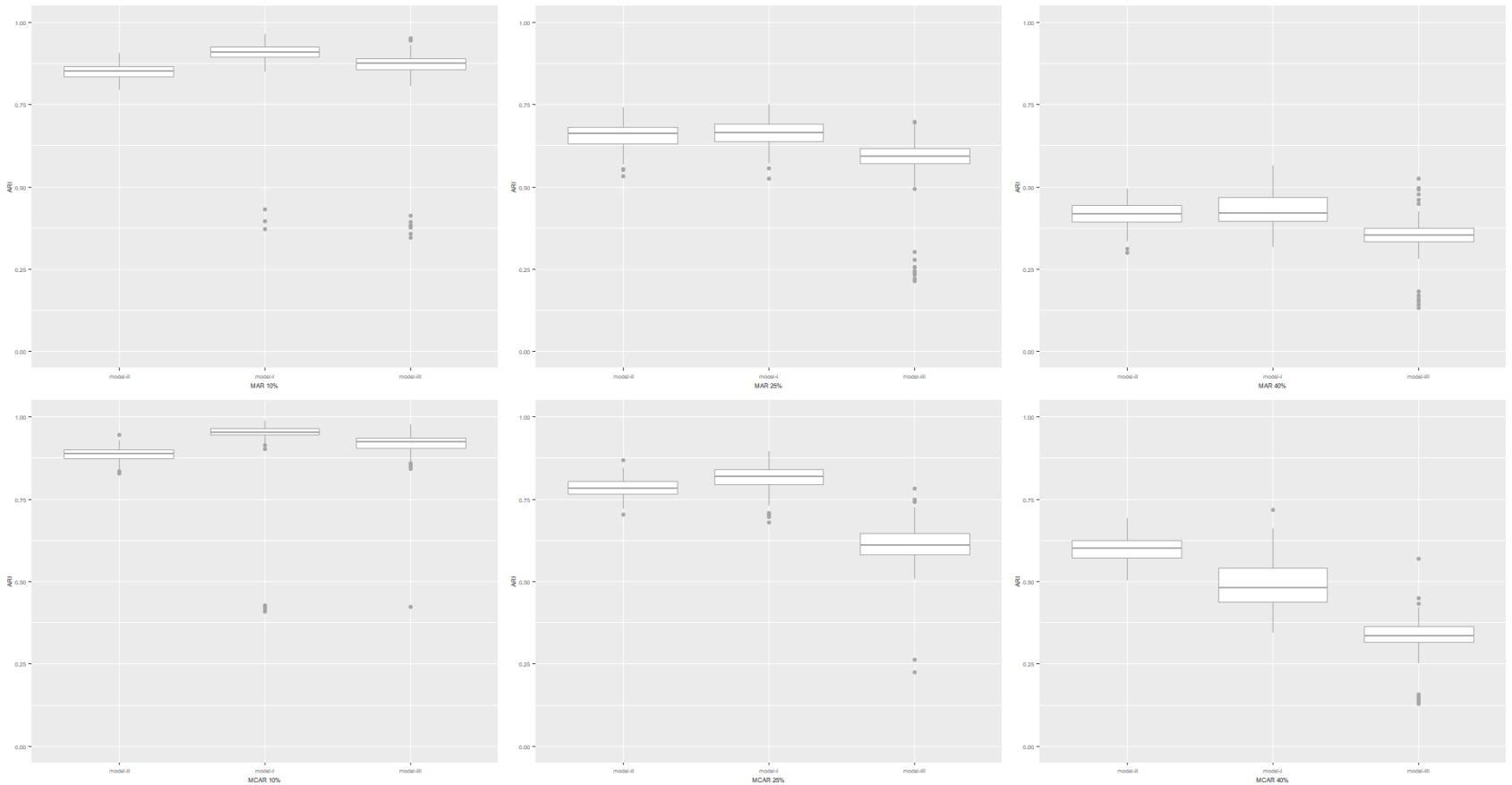
Résultats

K-means



Résultats

K-POD



Conclusion

Le consensus pondéré n'est pas meilleur que le consensus simple pour l'imputation multiple en clustering ;

L'indépendance des tableaux imputés n'est pas très importante pour effectuer une imputation multiple en clustering ;

L'imputation multiple est meilleure que les méthodes directes de clustering sur données incomplètes.