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 $\mathbf{Z} = (z_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ ensemble de données numériques

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

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

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**... $\mathbf{z}_i = (\mathbf{z}_i^{obs}, \mathbf{z}_i^{miss})$

Méthodes "Ad-hoc"

- Analyse des cas complets
- Imputation simple

Méthodes avancées

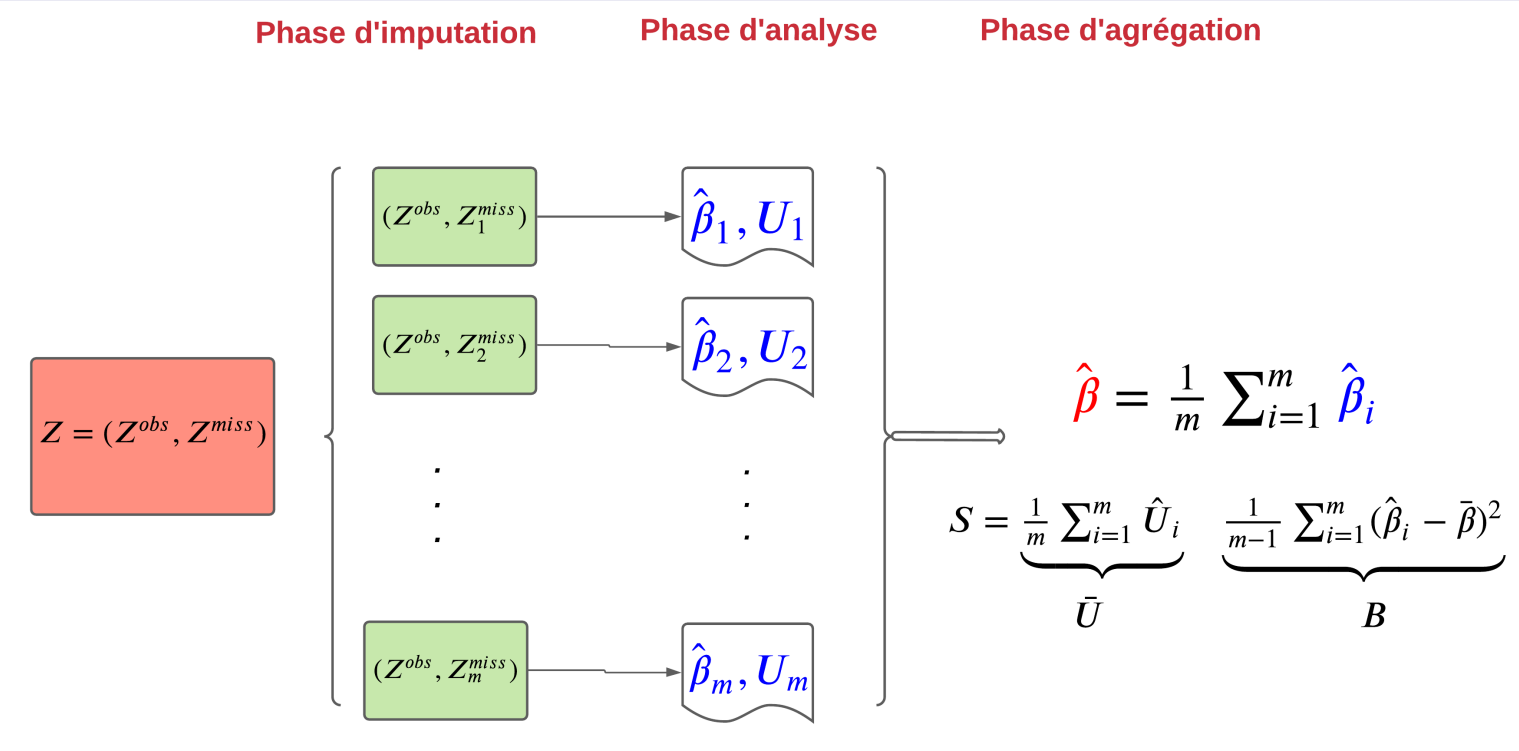
- Imputation Multiple (IM)
- Méthodes directes (fuzzy C-means, k-Pod, ...)

Plan

Plan

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

Principe



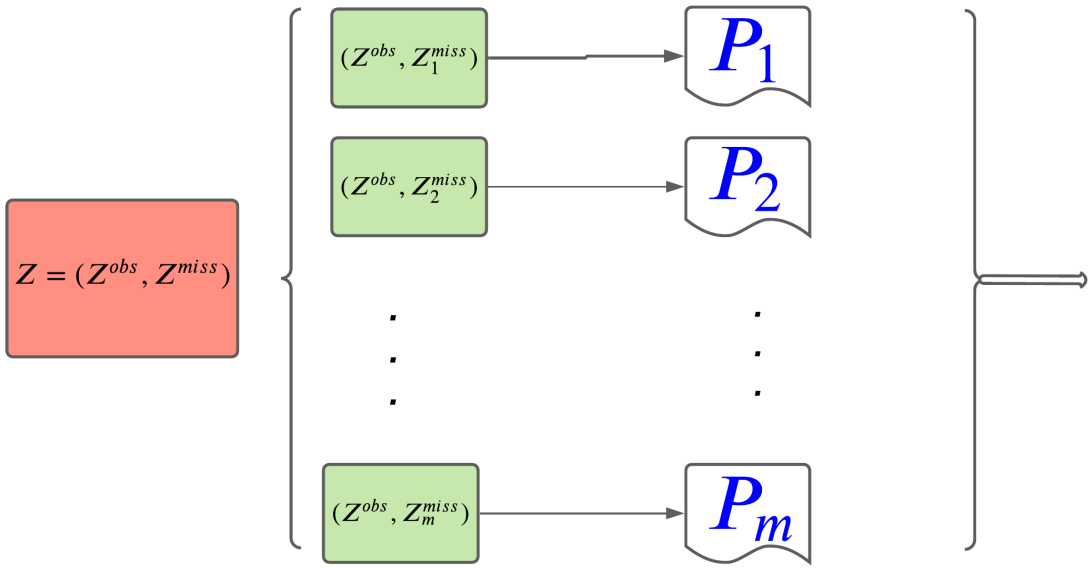
- Estimation unique de β et sa variance : $S = \bar{U} + B$
- Qualité de $\hat{\beta}$ dépend de l'indépendance des tableaux imputés conditionnellement aux valeurs observées

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)

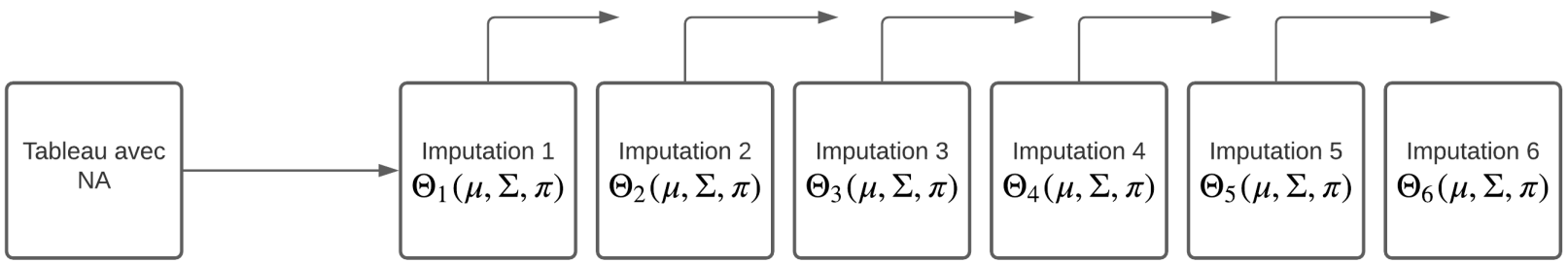
Approche d'imputation

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

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 $\Theta(\mu, \Sigma, \pi)$.

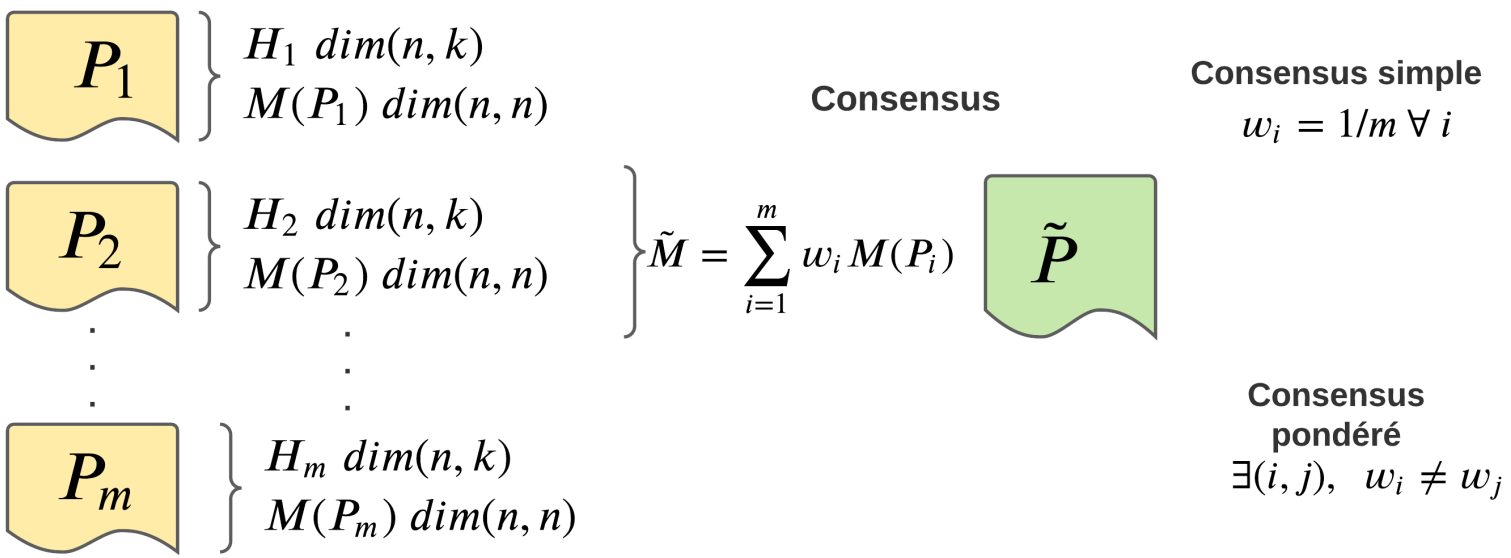
Imputation multiple : Répéter à nouveau (1) et (2) de façon à récupérer m réalisations des paramètres issus de la loi a posteriori.



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 :

$$\min_{H} \|\tilde{M} - HH^{'}\|^2 \tag{1}$$

- $H = \{0,1\}^{(nk)}$: matrice d'indicatrices
  $M = \{0,1\}^{(nn)}$: matrice de connectivité.
- En posant $\tilde{H} = H\,(H'H)^{-1/2}$ et $D = \text{diag}\,(H'H)$ (1) devient :

$$\min_{\tilde{H}'\tilde{H}=I,\tilde{H}\geq 0} \|\tilde{M} - \tilde{H}D\tilde{H}^{'}\|^2 \tag{2}$$

$$\tilde{H} \leftarrow \tilde{H}\sqrt{\frac{(\tilde{M}'\tilde{H}D)}{(\tilde{H}\tilde{H}'\tilde{M}'\tilde{H}D)}} \tag{3}$$

$$D \leftarrow D\sqrt{\frac{(\tilde{H}'\tilde{M}\tilde{H})}{(\tilde{H}'\tilde{H}D\tilde{H}'\tilde{H})}} \tag{4}$$

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

- Introduire w tel que :

$$w = (w_1, w_2, \ldots, w_m)', \quad w_t \geq 0, \quad \|w\|_1 = \sum_{t=1}^{m} w_t = 1 \tag{5}$$

- On définit $\tilde{M} = \sum_{t=1}^{m} w_t M(P_t)$ et le problème (1) devient :

$$\min_{w,\tilde{H}} \left\| \tilde{M} - \tilde{H}\tilde{H}' \right\|^2 \tag{6}$$

1. Chercher $\tilde{H}$ en fixant w en utilisant NMF
2. Chercher w en fixant $\tilde{H}$ en résolvant $\min w'Aw - 2b' + cte$

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 :

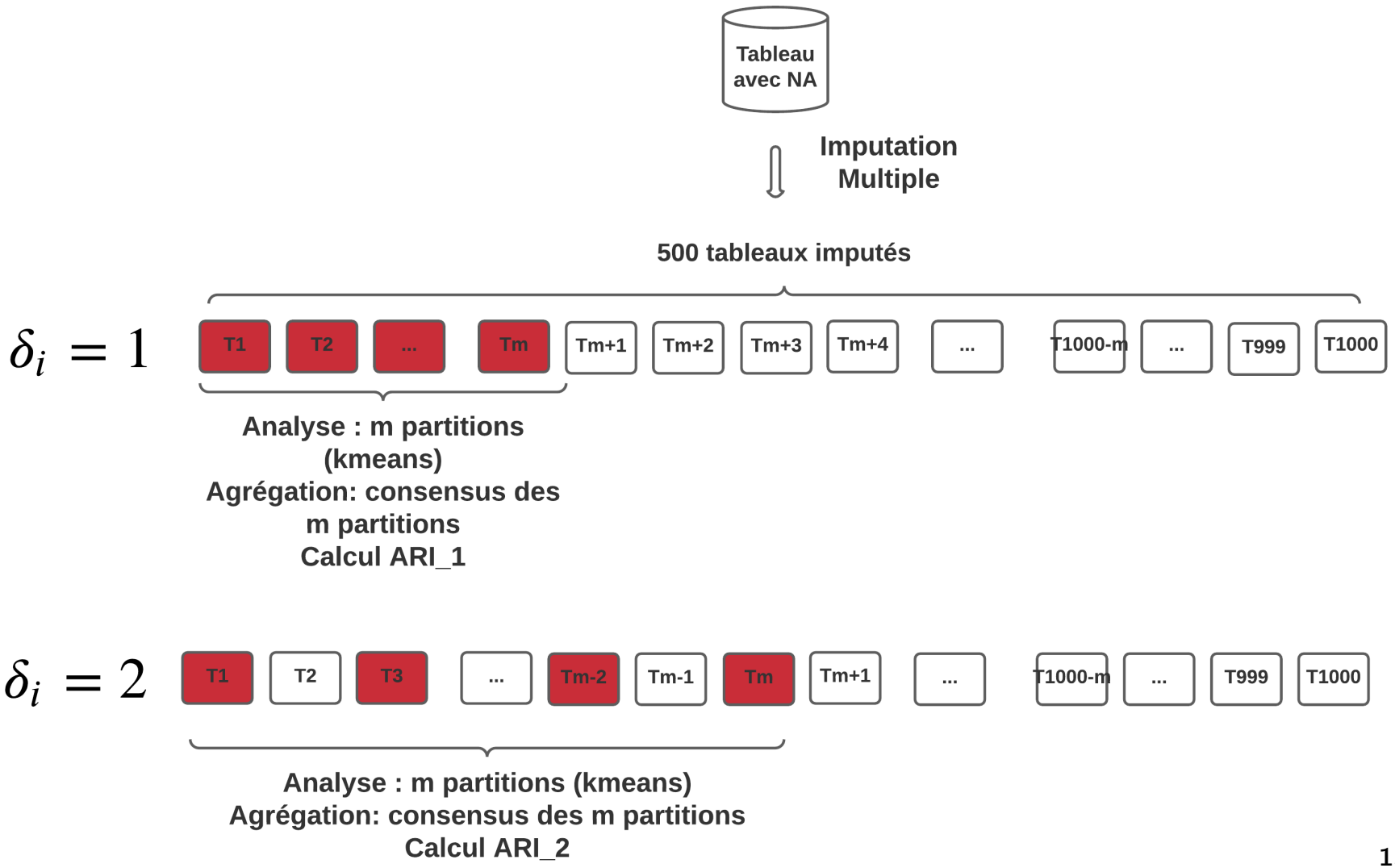
Données

- $n = 100$ observations, $p = 10$ variables
- $K = 2$ clusters : $n_w = 50$, $w \in \{1, 2\}$
- moyennes μ_w, $w \in \{1, 2\}$: $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
  $\mu_2 = (0, 0, 0, 0, 0, 2, 2, 2, 2, 2)$
- Matrice variance covariance $\Sigma = \begin{pmatrix} I_5 & 0 \\ 0 & \begin{matrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{matrix} \end{pmatrix}$
- **30%** de données manquantes selon un mécanisme MAR.

Plan

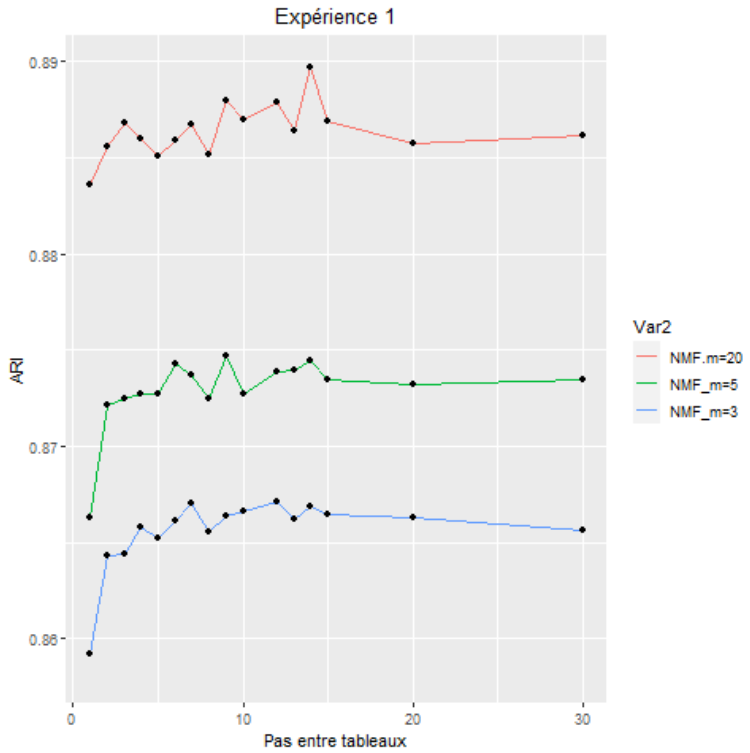
Plan d'expérimentation

$m \in \{3, 5, 20\}$: nombre de tableaux

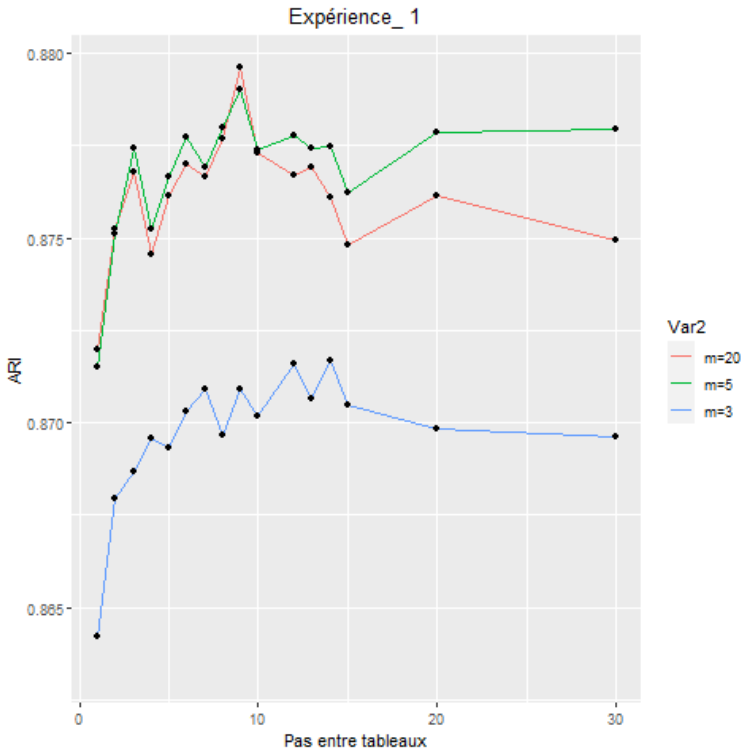


Résultats

Résultats NMF



Résultats WNMF

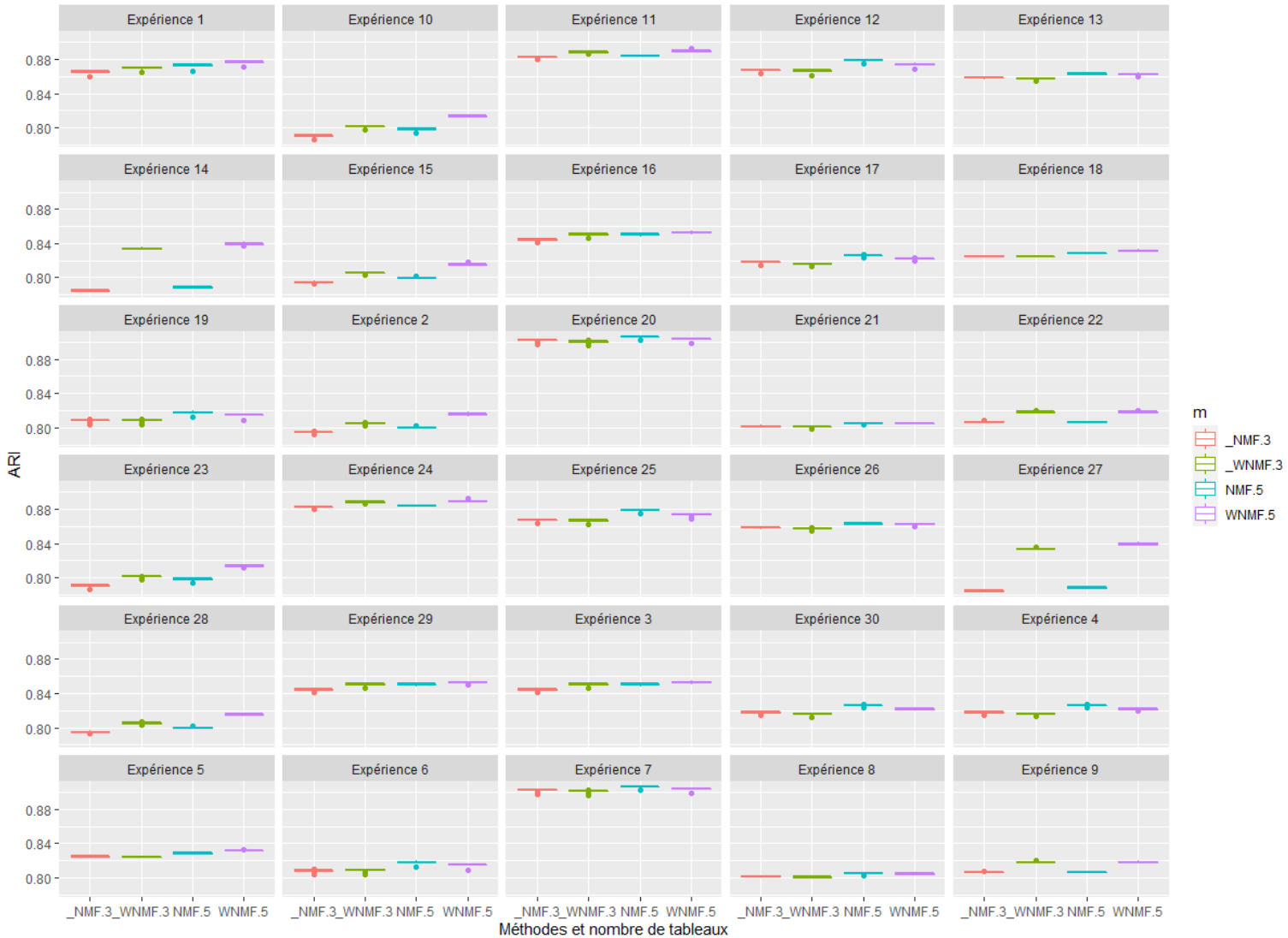


Résultats

- 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 **différences 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

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

$$\Longleftrightarrow$$

$$min_{H,C} \mid\mid Z - HC \mid\mid_F^2$$

k-POD

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

$$\Longleftrightarrow$$

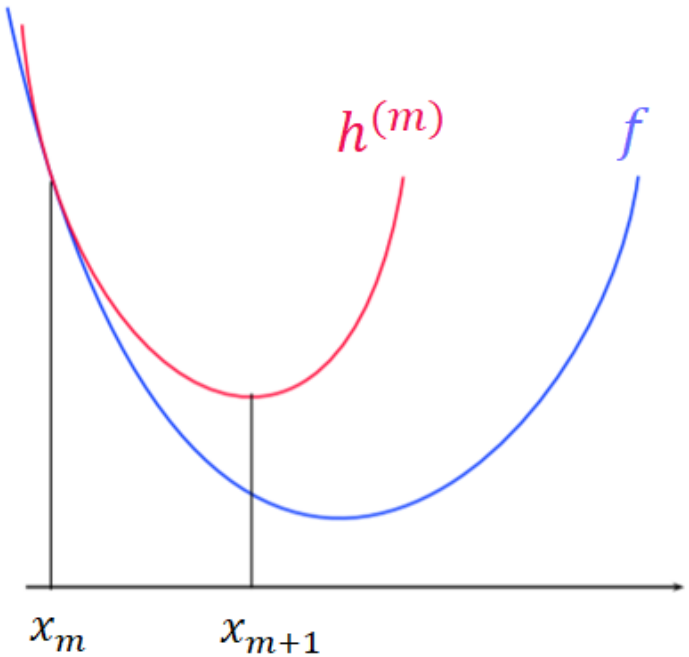
$$min_{H,C} \mid\mid P_\Omega(Z) - P_\Omega(HC) \mid\mid_F^2$$

Pas de solution explicite

- H=[h$_{ik}$] $\in M_{n \times K} \rightarrow h_{ik} = \begin{cases} 1 & si \quad i \in C_k \\ 0 & sinon \end{cases}$
- $\Omega \subset \{1,\ldots,n\} \times \{1,\ldots,p\} \rightarrow$ sous-ensemble des indices des valeurs observées
- $P_\Omega \rightarrow [P_\Omega(Z)]_{ij} = \begin{cases} z_{ij} & si \quad (i,j) \in \Omega \\ 0 & si \quad (i,j) \in \Omega^c \end{cases}$.

K-POD (Chi et al., 2016)

Algorithme de majoration-minimisation



La fonction de perte $f : x \longrightarrow f(x)$ est majorée à l'itération m par la fonction $h^{(m)}$.

K-POD (Chi et al., 2016)

Fonction de perte $f(H, C)$

$$|| P_\Omega(Z) - P_\Omega(HC) ||_F^2$$

Fonction majorante $g^{(q)}$ **en** $(H^{(q)}, C^{(q)})$

$$|| P_\Omega(Z) - P_\Omega(HC) ||_F^2 + || P_{\Omega^C}(Z) - P_{\Omega^C}(H^{(q)} C^{(q)}) ||_F^2$$

$$=$$

$$|| Z^{(q)} - HC ||_F^2$$

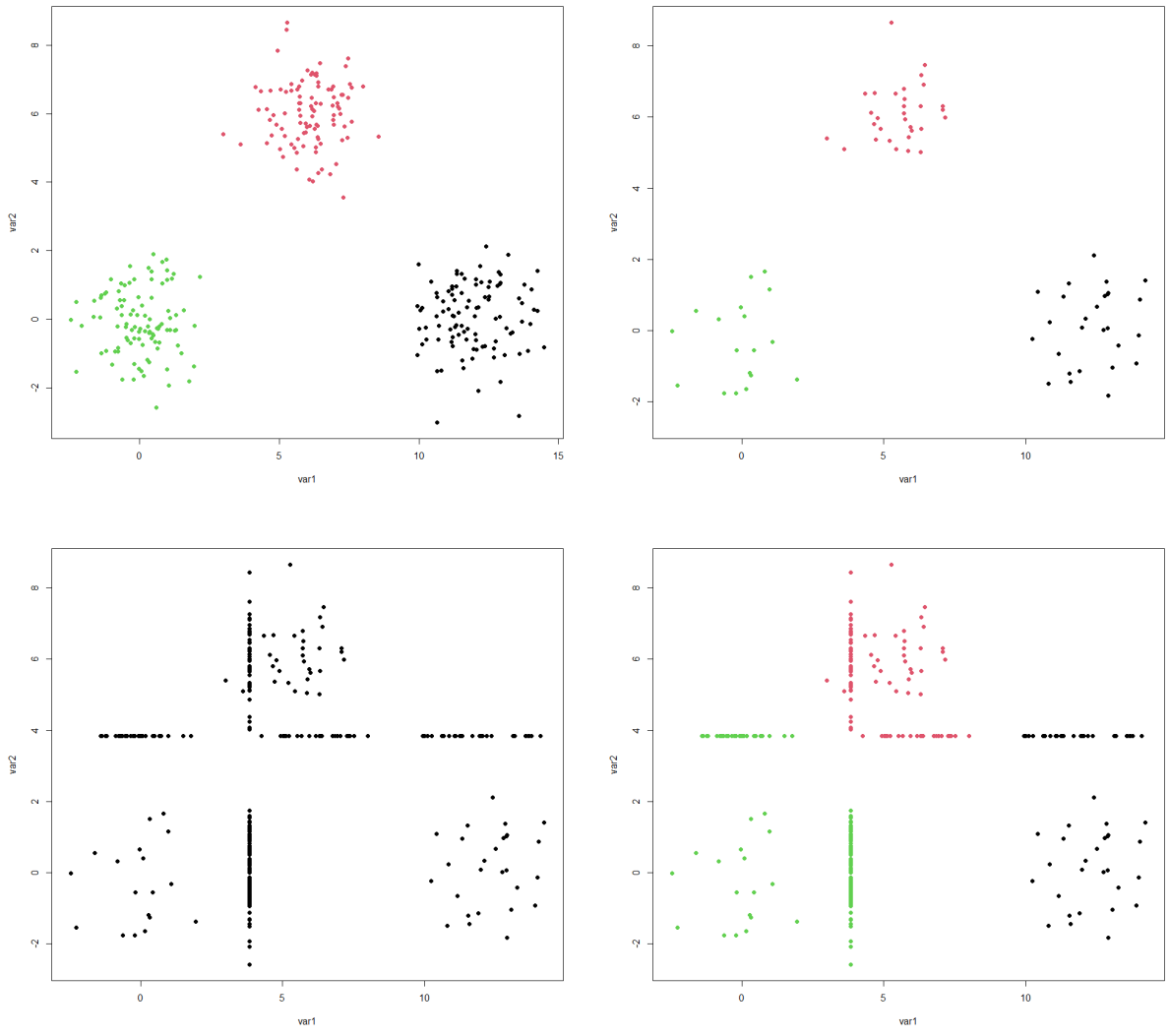
K-POD (Chi et al., 2016)

Pour le nombre K de clusters fixé :

- **Initialiser** les valeurs manquantes et obtenir la base complétée $Z^{(0)}$, puis calculer la **partition initiale** $(H^{(0)}, C^{(0)})$ par **k-means** ;
- **Répéter** pour les itérations $q = 0, 1, 2, ...$ jusqu'à convergence :
  - remplacer les valeurs manquantes par celles des centroïdes de leur groupe respectif pour obtenir $Z^{(q)}$;
  - estimer le minimum $(H^{(q+1)}, C^{(q+1)})$ de la fonction majorante $g^{(q)}(H, C|H^{(q)}, C^{(q)})$ par k-means sur $Z^{(q)}$.

K-POD (Chi et al., 2016)

Initialisation des valeurs manquantes



Ignorable-GMM (Marbac et al., 2019)

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

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

- GMM :

$$Z_i|c_i = k \sim g_k(.|\theta_k) = \mathcal{N}(.; \mu_k, \Sigma_k)$$

- Ignorable-GMM :

$$Z_i|c_i = k \sim \mathcal{N}(.; \mu_k, \Sigma_k) = \prod_{j \in O_i} \mathcal{N}(z_{ij}; \mu_{kj}, \Sigma_{kj})$$

- Log-vraisemblance ignorable-GMM :

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

Fuzzy c-means

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

$$argmin_{c_1,\ldots,c_K,\gamma_{11},\ldots,\gamma_{nK}} \sum_{k=1}^{K}\sum_{i=1}^{n} \gamma_{ik}^{m} \mid\mid x_i - c_k \mid\mid_2^2$$

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 $\gamma_{ik}^{(q)} = \frac{d(X_i, c_k^{(q-1)})^{\frac{1}{\alpha-1}}}{\sum_{l=1}^{K} d(X_i, c_l^{(q-1)})^{\frac{1}{\alpha-1}}}$
  - Calculer les nouveaux centroïdes de chacun des K clusters

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

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

Simulations

Configurations de données simulées

Configuration de référence (**modèle I**)

- $n = 750$ observations, $p = 8$ variables
- $K = 3$ clusters : $n_w = 250, w \in \{1, 2, 3\}$
- moyennes $\mu_w, w \in \{1, 2, 3\})$: $\mu_1 = (0, 0, 0, 0, \Delta, \Delta, 0, \Delta^2)$
  $\mu_2 = (0, 0, 0, 0, -\Delta, -\Delta, -\Delta, 0)$
  $\mu_3 = (0, 0, 0, 0, -\Delta, \Delta, \Delta, -\Delta^2)$
- Matrices de variance-covariance

$$\Sigma_{w,w=1,2,3} = \Sigma(\rho) = \begin{pmatrix} I_4 & 0 \\ 0 & \begin{matrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{matrix} \end{pmatrix} \text{ avec } \rho = 0.3.$$

Simulations

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

- MCAR [a] : $Prob(r_{ij} = 0) = \tau$
- MAR [b] 1 : $Prob(r_{ij} = 0) = \Phi(a_\tau + x_{i1})$
- MAR 2 : $Prob(r_{ij} = 0) = \Phi(a_\tau + x_{i8})$

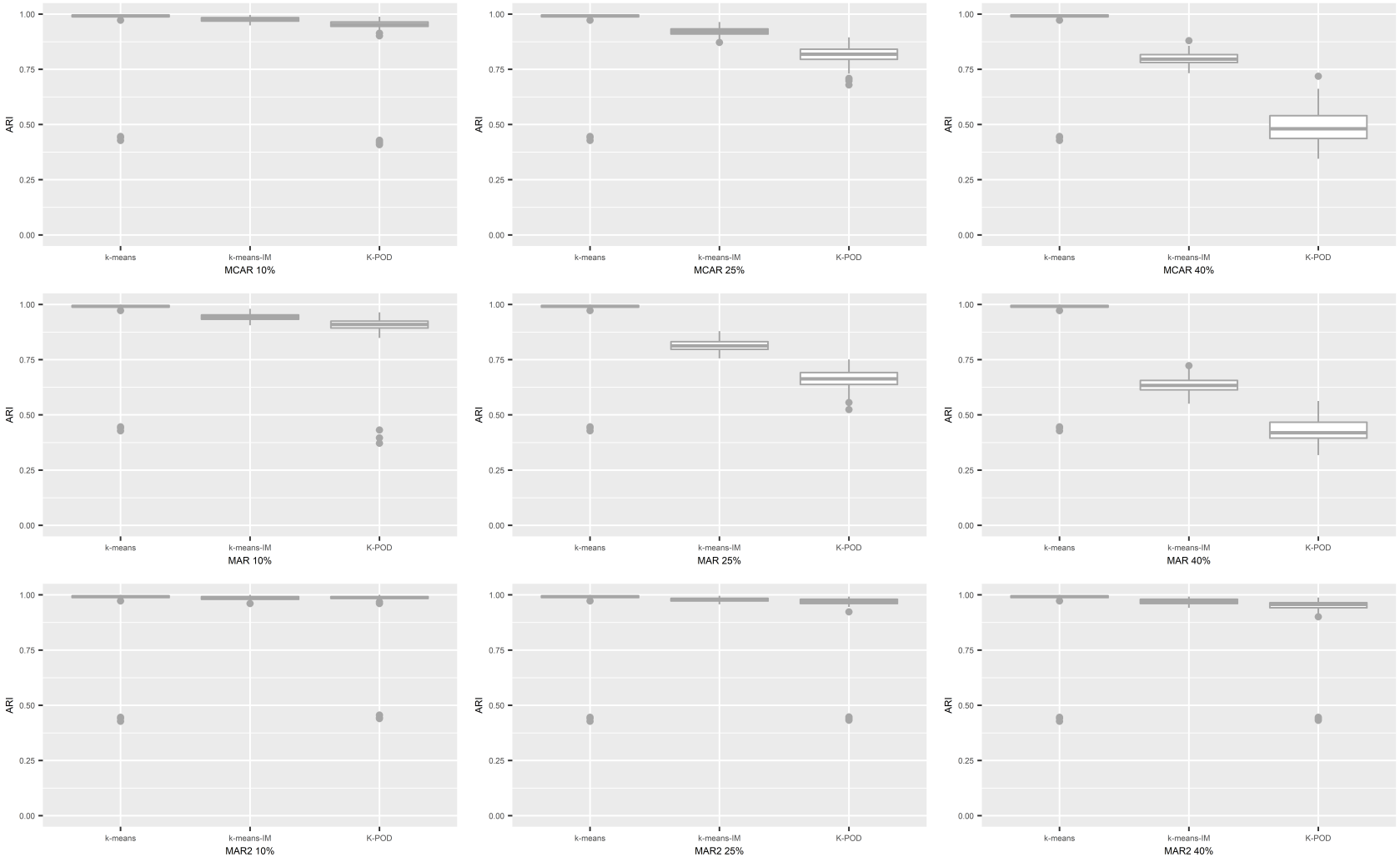
---

a. Missing Completely At Random
b. Missing At Random

- Φ la fonction de répartition de la loi normale standard
- τ une constante contrôlant la proportion de valeurs manquantes attendue
- $R = (r_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ le dispositif des valeurs manquantes, $r_{ij} = 1$ si x_{ij} est observé et $r_{ij} = 0$ si x_{ij} est manquant

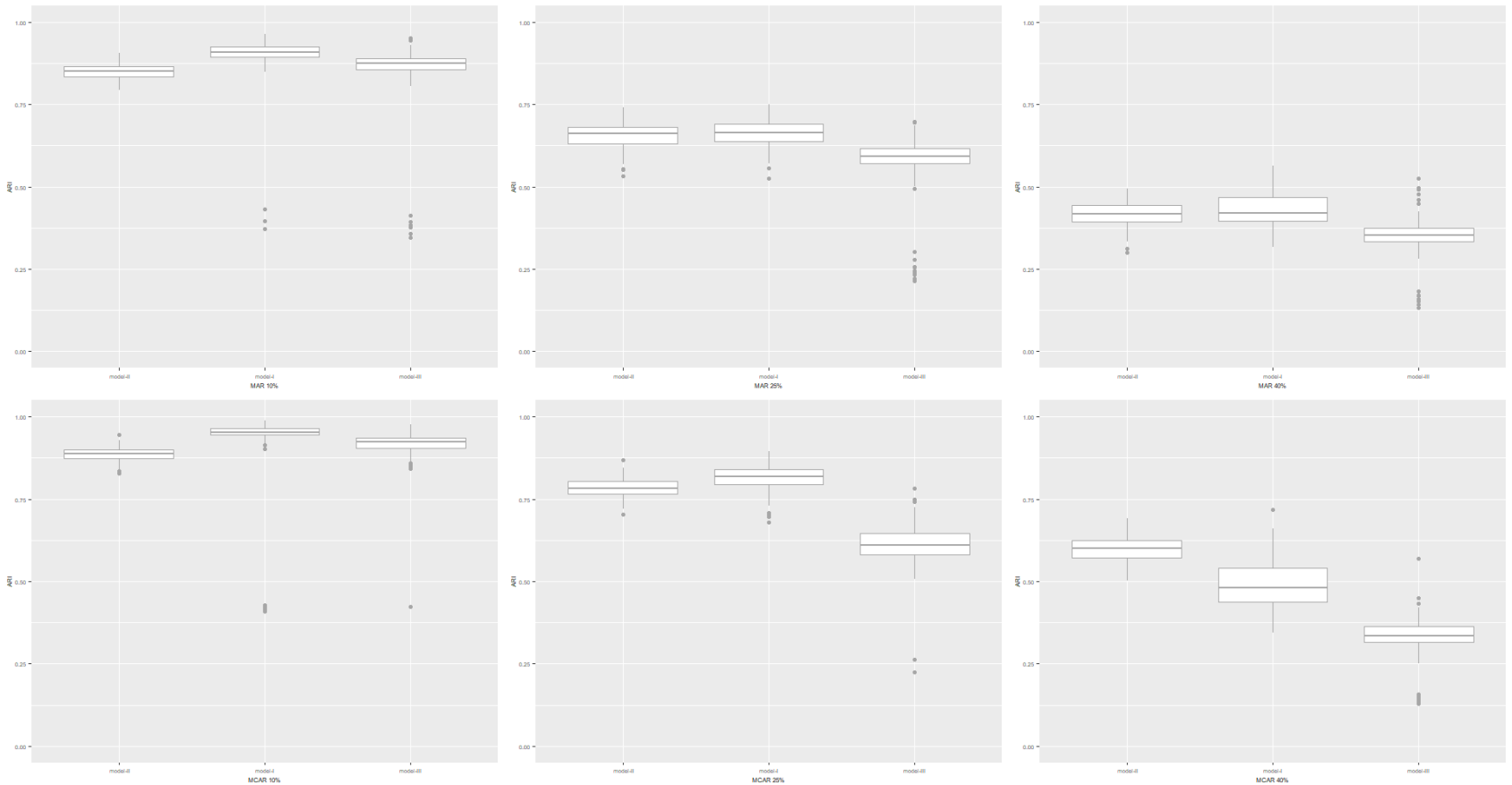
Résultats

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.