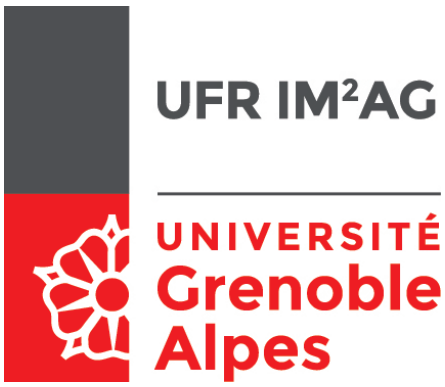
Rapport de Stage de M2

Soheïl Mecheloukh,

Méthodes d'imputation des données manquantes avec des méthodes de faible rang,



Encadrantes :

- Emilie Devijver,
- Adeline Leclercq Samson,
- Aude Sportisse,

Remerciements

Je tenais à remercier mes encadrantes qui m'ont accompagnés durant tout le stage : Emilie Devijver et Adeline Leclercq Samson et tout particulièrement Aude Sportisse qui m'a beaucoup aidé durant nos entretiens hebdomadaires, qui était toujours disponible si j'avais des interrogations et qui m'a bien motivé durant le stage !

Je tenais aussi à remercier l'équipe Aptikal du LIG, qui a su m'accueillir très rapidement, je me suis tout de suite senti à ma place.

Je tenais enfin à remercier particulièrement mes collègues du bureau 318 Romain Alves, Corentin Masson et Théotime Le Goff, qui ont rendu mon stage très agrèable et très humain, qui m'ont aidé à comprendre ce que c'était la recherche et qui ont toujours été de bon conseil pour m'apprendre le métier !

Contents

Notation	**Description**
n	Nombre d'individus ($i = 1, \ldots, n$)
R	Nombre de blocs/modalités de variables ($r = 1, \ldots, R$)
m_r	Nombre de variables dans le bloc r avec $r \in (1, \ldots, R)$
$m = \sum_{r=1}^{R} m_r$	Nombre total de variables toutes modalités confondues
$X \in \mathbb{R}^{n \times m}$	Matrice complète, sans données manquantes (non observée)
X^{NA}	Matrice incomplète, contenant les données manquantes
$X_{\text{obs}(M)}, X_{\text{obs}}$	Ensemble des valeurs observées dans X^{NA}, dépendant du masque M
$X_{\text{miss}(M)}, X_{\text{miss}}$	Ensemble des valeurs manquantes dans X^{NA}, dépendant du masque M
X_{ij}	Valeur (potentiellement manquante) de la variable j pour l'individu i
$M \in \{0, 1\}^{n \times m}$	Matrice binaire de présence de données : $M_{ij} = 1$ si X_{ij} est manquant, 0 sinon
K	Nombre total de clusters
$N_i \in \{1, \ldots, K\}$	Cluster latent auquel appartient l'individu i

Table 1: Résumé des notations utilisées

Présentation du stage et du rapport

J'ai fait mon stage de fin de M2 au LIG du 17 mai au 1er août 2025. Mon stage portait sur l'imputation des données manquantes à l'aide de méthodes de faible rang. J'ai été encadré durant la totalité de mon stage par Aude Sportisse, Emilie Devijver et Adeline Leclercq Samson. C'était un stage de recherche où j'ai souvent alterné entre la bibliographie, la reproduction de méthode et les simulations.

J'ai commencé mon stage par la lecture de l'ouvrage de Little and Rubin [2019] pour pouvoir me mettre à jour sur les données manquantes, que ce soit les notations, les principales méthodes d'imputation et l'algorithme Espérance-Maximisation (EM).

L'objectif initial du stage était de trouver une nouvelle méthode d'imputation des données manquantes à l'aide de méthodes de faible rang. Cet objectif a évolué durant le stage pour des raisons de temps et j'ai trouvé un article sur arXiv (Yang and Li [2025]) durant le stage qui proposait cette méthode.

Le rapport de stage est un résumé de ce que j'ai compris et réalisé, des notions de bases jusqu'aux preuves de certains résultats. Le plan du rapport est le suivant : la première partie est une présentation des données manquantes. La deuxième partie présente brièvement les méthodes d'imputation. La troisième partie fait l'inventaire des outils qui ont été utilisés pendant mon stage, l'algorithme EM et un rappel sur les méthodes de faible rang. La quatrième partie porte sur le coeur du sujet: une méthode utilisant l'algorithme EM pour un dérivé de l'analyse canonique des corrélations (CCA).

1 Introduction aux données manquantes

1.1 Définition d'une donnée manquante

Voici comment une donnée manquante est défini dans Little and Rubin [2019].

> *"Missing data are unobserved values that would be meaningful for the analysis if observed; in other words, a missing value hides a meaningful value."*

Une valeur manquante représente une valeur non définie dans la matrice d'observations. Ci-dessous, un exemple d'une matrice 3×3 sans données manquantes est présentée à côté d'une matrice contenant des valeurs manquantes.

Sans données manquantes:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Avec données manquantes:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & \text{Na} & 6 \\ 7 & 8 & \text{Na} \end{bmatrix}$$

Une façon naïve et pourtant encore très répandue pour gérer les données manquantes et de travailler avec uniquement les individus (les lignes) où toutes les variables sont observées mais cette méthode présente de nombreux inconvénients, l'un des principaux étant la perte d'information.

L'exemple de Zhu et al. [2022] montre comment la perte d'information se manifeste dans un jeu de données avec d variables et un ratio de valeurs manquantes de 1%. Les résultats sont les suivants :

- Pour $d = 5$, 95% des lignes sont complètes.
- Pour $d = 300$, seulement 5% des lignes sont complètes.

L'un des autres problèmes est que cela introduit un biais dans l'analyse si la sous-population des lignes complètes n'est pas nécessairement représentative de la population globale. On reviendra à ce problème en section sur les mécanismes des données manquantes.

1.2 Notation

Nous allons dans cette sous partie présenter la notation qui sera utilisée pour le reste du document, cette notation est inspirée de Little and Rubin [2019].

Soit X la matrice des données complète, de dimension $n \times p$, qui n'est pas observée. Soit X^{NA} la matrice des données incomplète. Soit $X_{\text{obs(M)}}$ (resp. $X_{\text{miss(M)}}$) ou X_{obs} (resp. X_{miss}) l'ensemble des valeurs observées (resp. manquantes) dans X^{NA}, qui dépend de M. Soit $M \in \{0,1\}^{n \times p}$ le masque, qui indique où se trouvent les valeurs manquantes dans X^{NA}. Plus précisément, pour tout i, pour tout j, on a :

$$M_{ij} = \begin{cases} 1 & \text{si } X_{ij}^{NA} \text{ est manquante,} \\ 0 & \text{sinon.} \end{cases}$$

Les notations du document sont résumées dans le tableau 1.

Exemple Soit la matrice X suivante :

$$X = \begin{pmatrix} 4 & 5 & 8 \end{pmatrix}$$

La matrice X^{NA} avec les valeurs manquantes est :

$$X^{\text{NA}} = \begin{pmatrix} 4 & \text{NA} & 8 \end{pmatrix}$$

La matrice M, indiquant où sont les données manquantes, est :

$$M = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

La matrice des données observées $X_{\text{obs}(M)}$ est donnée par :

$$X_{\text{obs}(M)} = \begin{pmatrix} 4 & 8 \end{pmatrix}$$

La matrice des données manquantes $X_{\text{miss}(M)}$ est donnée par :

$$X_{\text{miss}(M)} = \begin{pmatrix} 5 \end{pmatrix}$$

$X_{\text{miss}(M)}$ ne sera jamais connu en pratique ni observé.

1.3 Mécanismes des données manquantes

Il existe trois mécanismes qui expliquent la présence des données manquantes, définis par Rubin [1976]. C'est ce que nous allons voir dans cette partie.

Missing Completely at Random (MCAR) Cela signifie que le masque M indiquant si la donnée est présente ou manquante ne dépend pas des données, donc X et M sont indépendants, ce qui implique:

$$p(M \mid X) = p(M)$$

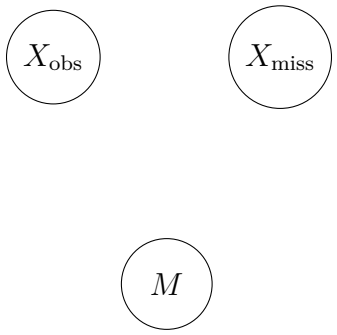


Figure 1: Schéma causal de MCAR

Exemple de données manquantes MCAR :

1. **Exemple 1 :** Dans une étude longitudinale, des mesures sont manquantes sur certains participants à chaque vague de l'étude, mais cela est dû à des absences aléatoires, les absences ne sont pas liées aux variables observées ou non observées.

2. **Exemple 2 :** Des erreurs de mesure peuvent se produire de manière aléatoire dans un appareil de collecte de données. Si une sonde se déconnecte parfois sans raison systématique, des valeurs manquantes se produisent, indépendamment des données.

Missing at Random (MAR) Le mécanisme de données manquantes dépend uniquement des variables observées et non des valeurs manquantes elles-mêmes, c'est-à-dire:

$$p(M \mid X) = p(M \mid X_{\text{obs(M)}}).$$

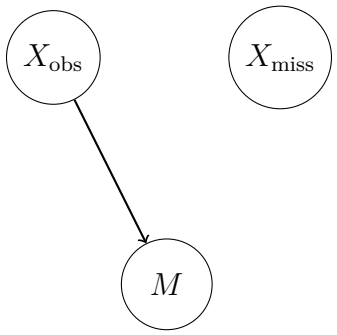


Figure 2: Schéma causal de MAR

Exemple de données manquantes MAR :

1. **Exemple 1 :** Lors d'un sondage, on pose une question sur le salaire et on remarque que les personnes agées ne répondent pas à la question, donc le manque de réponse dépend d'une autre variable : l'âge.

2. **Exemple 2 :** Dans une étude longitudinale, les données sont manquantes sur certains participants à chaque vague de l'étude en raison de facteurs externes (par exemple les vacances), donc à chaque même période de l'année on aura plus de manque.

Missing not at Random (MNAR) Le mécanisme de données manquantes dépend à la fois des variables observées et des variables manquantes:

$$p(M \mid X) = p(M \mid X_{\text{obs(M)}}, X_{\text{miss(M)}}).$$

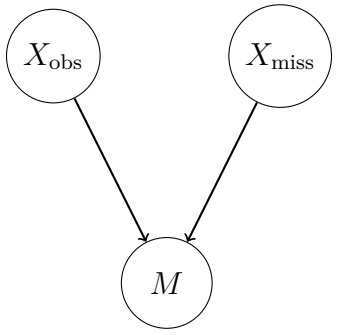


Figure 3: Schéma causal de NMAR

Exemple de donnée manquante MNAR :

1. **Exemple 1:** Lors d'un sondage, on pose une question sur le salaire et on remarque que les personnes très aisées ou très pauvres répondent moins à la question, donc le manque dépend de la réponse à la question elle-même (manque auto-masqué).

2. **Exemple 2:** Dans une étude clinique, les mesures d'un biomarqueur sont manquantes pour certains patients, car ceux-ci appartiennent à un sous-groupe biologique non identifié qui interfère avec la méthode de dosage. Le mécanisme de données manquantes dépend donc d'une variable non observée.

1.4 Ignorabilité des mécanismes

Paramétrisation des distributions. On suppose que les données complètes suivent une distribution paramétrée par $\theta \in \Omega_\theta$, notée $p(X;\theta)$, et que le mécanisme de données manquantes est modélisé par une distribution conditionnelle $p(M \mid X;\phi)$, où $\phi \in \Omega_\phi$ est un second ensemble de paramètres.

Un point central dans l'analyse des données manquantes est que certains mécanismes sont dits *ignorables* : cela signifie que l'on peut estimer les paramètres θ sans avoir à modéliser explicitement le mécanisme des données manquantes (i.e., sans connaître ϕ).

Définition (Ignorabilité). Le mécanisme des données manquantes est dit *ignorable* pour l'estimation de θ si :

1. Il est **MAR** (*Missing At Random*) :

$$p(M \mid X;\phi) = p\Big(M \mid X_{\text{obs(M)}};\phi\Big)$$

2. Les paramètres θ et ϕ sont **distincts** : il n'y a pas de dépendance a priori entre eux.

Ignorabilité de MAR. La vraisemblance conjointe des données observées et du masque des données manquantes s'écrit :

$$L(\theta,\phi) = p(X_{\text{obs(M)}}, M;\theta,\phi) = \int p(X_{\text{obs}}, X_{\text{miss}};\theta)\cdot p(M \mid X_{\text{obs}}, X_{\text{miss}};\phi)\,\mathrm{d}X_{\text{miss}}.$$

Sous l'hypothèse MAR, on a :

$$p(M \mid X_{\text{obs}}, X_{\text{miss}};\phi) = p(M \mid X_{\text{obs}};\phi)$$

d'où :

$$L(\theta,\phi) = \left[\int p(X_{\text{obs}}, X_{\text{miss}};\theta)\,\mathrm{d}X_{\text{miss}}\right]\cdot p(M \mid X_{\text{obs}};\phi).$$

La première partie dépend uniquement de θ, tandis que la seconde dépend uniquement de ϕ. Ainsi, la vraisemblance des données observées devient :

$$L(\theta) \propto \int p(X_{\text{obs}}, X_{\text{miss}};\theta)\,\mathrm{d}X_{\text{miss}} = p(X_{\text{obs(M)}};\theta)$$

Cette expression ne dépend pas de ϕ, ce qui prouve que le mécanisme est **ignorable** pour l'inférence sur θ.

Conséquences selon le type de mécanisme.

Mécanisme	**Forme de $p(M \mid X)$**	**Ignorabilité**	**Modélisation de la vraisemblance**
MCAR	$p(M)$	Ignorable	$L(\theta) \propto p\Big(X_{\text{obs}}; \theta\Big)$
MAR	$p\Big(M \mid X_{\text{obs}}\Big)$	Ignorable	$L(\theta) \propto p\Big(X_{\text{obs}}; \theta\Big)$
MNAR	$p\Big(M \mid X_{\text{obs}}, X_{\text{miss}}\Big)$	Non-ignorable	$L(\theta) \propto \int p\Big(X_{\text{obs}}, X_{\text{miss}}; \theta\Big)\, p(M \mid X; \phi)\, \mathrm{d}X_{\text{miss}}$

Table 2: Classification des mécanismes de données manquantes, leur ignorabilité, et les conséquences sur la modélisation de la vraisemblance.

1.5 L'imputations des données manquantes

Les analyses en cas complet ignorent les lignes avec des valeurs manquantes pour estimer les distributions marginales ou les covariances. Cette approche est problématique car elle ne tient pas compte de la possible corrélation entre les variables observées et manquantes. Par exemple, si une variable Y_j (comme la taille) est manquante, mais qu'une autre variable Y_k (comme le poids) est fortement corrélée avec Y_j, il est possible d'estimer Y_j en fonction de Y_k.

L'imputation est une méthode pour traiter les données manquantes qui consistent à remplacer la donnée manquante NA par une valeur selon une distribution basée sur X_{obs}.

Les Principales Méthodes d'Imputation Simple

- **Imputation par la Moyenne** : Cette méthode consiste à remplacer les valeurs manquantes par la moyenne des unités observées.

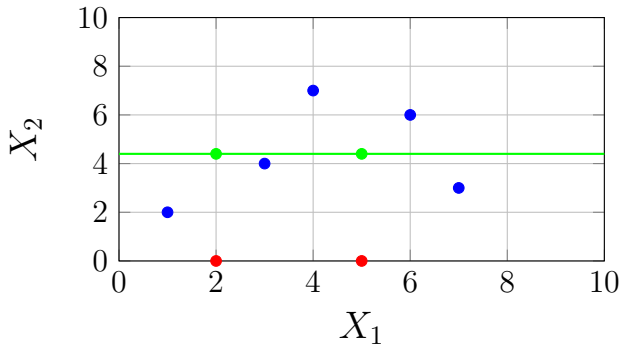


Figure 4: Illustration de l'imputation par la moyenne sur la variable X_2. Les points rouges sont manquants, les points verts sont imputés avec la moyenne des valeurs observées (ligne verte).

- **Imputation par Régression** : Cette méthode remplace les valeurs manquantes par les prédictions obtenues via une régression entre la variable manquante et d'autres variables observées. Les points imputés appartiennent à la droite de régression.

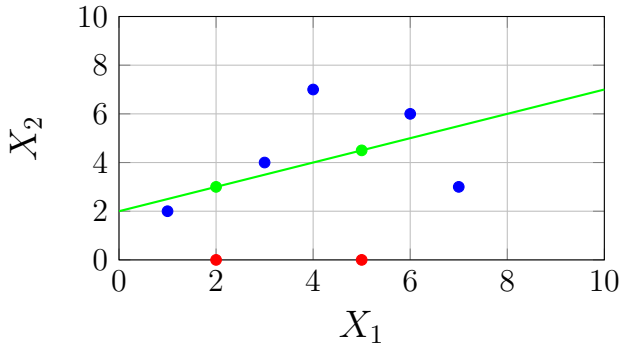


Figure 5: Imputation par régression : les valeurs manquantes sont imputées en projetant les observations sur la droite de régression estimée à partir des données complètes.

- **Imputation par Régression Stochastique** : Similaire à l'imputation par régression, cette méthode ajoute un résidu aléatoire tiré d'une distribution centrée autour de la droite de régression. Ce résidu reflète la variance observée des points autour de cette droite, ce qui permet de capturer l'incertitude associée aux prédictions. Ainsi, on évite de sous-estimer la variabilité des données imputées et on limite les biais potentiels liés à une imputation trop déterministe.

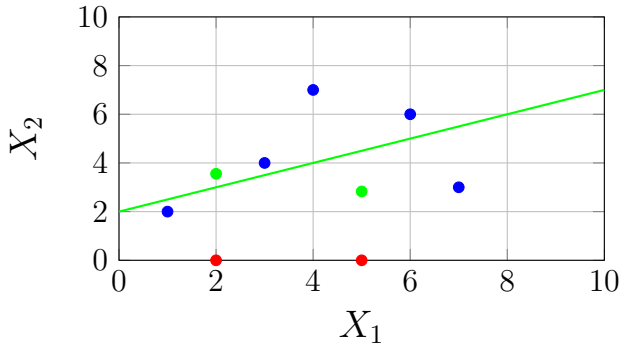


Figure 6: Imputation par régression stochastique : les valeurs manquantes sont imputées à partir de la droite de régression, avec une composante aléatoire simulée selon l'erreur résiduelle.

Il existe également l'imputation multiple, une approche qui consiste à générer plusieurs versions complètes du jeu de données en imputant plusieurs fois les valeurs manquantes selon un modèle probabiliste, afin de mieux refléter l'incertitude liée à l'imputation. Le détail de cette méthode est présenté en annexe C.

2 Méthodes de référence et outils analytiques

2.1 Les méthodes de faible rang

Soit une matrice $\Theta \in \mathbb{R}^{n \times p}$. On dit que Θ est de *faible rang* (ou *low-rank*) si son rang r est strictement inférieur à $\min(n, p)$, c'est-à-dire que ses lignes (ou ses colonnes) sont linéairement dépendantes, et que l'information contenue dans Θ peut être représentée à l'aide d'un nombre réduit de composantes principales.

Une matrice Θ de rang r peut être factorisée comme suit : $\Theta = UV^\top$, où $U \in \mathbb{R}^{n\times r}$ et $V \in \mathbb{R}^{p\times r}$. Cette décomposition indique que la matrice originale peut être reconstruite à partir de deux matrices de dimensions réduites. Les matrices de faible rang présentent plusieurs avantages importants dans le cadre de l'analyse de données, notamment lorsque celles-ci sont incomplètes ou bruitées :

- **Réduction de dimension** : Une matrice de faible rang peut être représentée par deux matrices de dimensions plus petites. Cela permet de réduire la complexité du modèle tout en conservant l'essentiel de l'information.
- **Modélisation des relations entre variables** : En supposant que la structure sous-jacente des données est de faible rang, il est possible d'établir un lien fort entre les variables, ce qui permet de mieux comprendre leur interdépendance et de capturer des structures latentes dans les données.

Pour ces deux principales raisons, utiliser des matrices de faible rang devient très intéressant dans la gestion des données manquantes.

Obtention en pratique : En pratique, une approximation de faible rang d'une matrice peut être obtenue via la *décomposition en valeurs singulières* (SVD). Soit une matrice $\Theta \in \mathbb{R}^{n\times p}$, sa SVD s'écrit :

$$\Theta = U\Sigma V^\top,$$

où $U \in \mathbb{R}^{n\times n}$ et $V \in \mathbb{R}^{p\times p}$ sont des matrices orthogonales, et $\Sigma \in \mathbb{R}^{n\times p}$ est une matrice diagonale contenant les valeurs singulières $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

On peut alors construire une *approximation de faible rang* Θ_r en ne conservant que les r plus grandes valeurs singulières :

$$\Theta_r = \sum_{i=1}^{r} \sigma_i u_i v_i^\top = U_r \Sigma_r V_r^\top,$$

où U_r, Σ_r et V_r sont les matrices tronquées correspondant aux r premières composantes. Cette approximation minimise l'erreur quadratique entre Θ et Θ_r, et constitue la meilleure approximation de rang r au sens de la norme de Frobenius.
Voici certaines méthodes célèbres qui utilisent des représentations ou des approximations de matrices de faible rang pour faire de l'analyse de donnée :

- **Analyse en Composantes Principales (ACP)** : Méthode de réduction de dimension qui cherche une approximation de faible rang de la matrice de données centrée, en conservant les directions principales de variation.
- **Méthodes dérivées de l'ACP** : Plusieurs extensions de l'ACP reposent également sur l'hypothèse de faible rang, telles que :
  - *ACP probabiliste (PPCA)* : version probabiliste de l'ACP, introduisant une modélisation bayésienne.
  - *ACP par noyau (Kernel PCA)* : extension non linéaire de l'ACP via des fonctions noyaux, capturant des structures de faible rang dans un espace de dimension plus élevée.

- **Analyse Factorielle des Correspondances (AFC)** : Adaptée aux tableaux de contingence (variables qualitatives), cette méthode applique une pondération aux lignes et colonnes et projette tous les points (lignes et colonnes) dans un même espace factoriel.
- **Analyse Factorielle de Données Mixtes (AFDM)** : Méthode qui permet de traiter conjointement des variables quantitatives et qualitatives dans un même tableau.

2.2 L'algorithme EM

Dans cette partie, je vais présenter l'algorithme EM et un cas simple d'utilisation. L'algorithme **EM** a été introduit par Dempster et al. [1977]. Il s'agit d'une méthode itérative largement utilisée pour l'estimation des paramètres dans des modèles statistiques impliquant des données incomplètes ou latentes. L'algorithme vise à maximiser la vraisemblance lorsque certaines données sont non observables.

Principe général dans le cadre des données manquantes Soit X les données, X_{obs} les données observés, X_{miss} les données latentes (ou manquantes), et $\theta^{(t)}$ le vecteur des paramètres du modèle à l'étape t. L'objectif est de maximiser la vraisemblance marginale $p(X_{\text{obs}} \mid \theta)$. L'algorithme EM procède en deux étapes principales, répétées jusqu'à convergence :

1. **Étape E (Expectation)** : Calculer l'espérance de la log-vraisemblance complète, conditionnellement aux données observées et à l'estimation courante des paramètres :
$$Q(\theta; \theta^{(t)}) = \mathbb{E}_{X_{miss}|X_{obs},\theta^{(t)}} \left[\log L(\theta; X) \mid X_{obs}, \theta^{(t)} \right]$$
2. **Étape M (Maximization)** : Maximiser cette espérance pour obtenir une nouvelle estimation des paramètres :
$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta; \theta^{(t)})$$

Convergence L'algorithme EM garantit que la vraisemblance augmente à chaque itération. Toutefois, il peut converger vers un maximum local plutôt qu'un maximum global, ce qui dépend fortement de l'initialisation.

Pour limiter ce risque, il est courant d'utiliser des initialisations intelligentes, comme une initialisation par ACP sur les données complétées par la moyenne. Une deuxième possibilité est d'opter pour une initialisation aléatoire répétée plusieurs fois pour choisir la meilleure solution parmi plusieurs runs.Une troisième possibilité est de faire une initialisation basée sur une imputation simple (par la moyenne ou une régression), suivie d'une estimation des paramètres du modèle.

Ces stratégies permettent généralement d'obtenir une meilleure convergence et de limiter l'impact des maximums locaux.

Exemple de cas d'utilisation Voici un exemple de cas d'utilisation que j'ai réalisé pour bien comprendre le fonctionnement de l'EM : Considérons une variable aléatoire bivariée $X = (X_1, X_2)$ suivant une loi normale multivariée centrée réduite :

$$X \sim \mathcal{N}(\mu, \Sigma), \quad \text{où} \quad \mu = \begin{pmatrix} -5 \\ 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Supposons que seules des valeurs de X_2 sont manquantes dans une certaines proportion p dans l'échantillon observé. On utilise alors l'algorithme EM pour estimer les paramètres μ et Σ. Voir en annexe les calculs A.

EM et méthode de faible rang Il existe plein de variante de l'algorithme EM notamment avec des méthodes de faible rang, une variante qui m'a beaucoup servi à le comprendre est celle avec l'Analyse par Composante Principal (ACP). Elle est implémenté dans le package R MissMDA qui a été concu par Julie Josse et François Husson, la présentation de ce méthode est tiré de leurs articles sur ce package : Josse and Husson [2016], c'est une méthode qui implémente l'ACP de façon itérative à l'aide de l'algorithme EM, je mets le détail de ce package en annexe G.

3 CCA, PCCA et GPCCA

3.1 Présentation de la CCA

L'Analyse en Corrélation Canonique (CCA) est une technique d'analyse multivariée visant à identifier les relations linéaires maximales entre deux « vues » d'un même phénomène, chacune représentée par un ensemble de variables aléatoires. Elle a été introduite par Hotelling [1936].

Modèle et notations

On dispose de deux matrices de données centrées :

$$X_1 \in \mathbb{R}^{n \times m_1}, \quad X_2 \in \mathbb{R}^{n \times m_2},$$

où chaque ligne correspond à un individu (n observations) et chaque colonne à une variable (m_1 et m_2 variables respectivement).

Objectif de la CCA

On cherche deux vecteurs de projection

$$\mathbf{a} \in \mathbb{R}^{m_1}, \quad \mathbf{b} \in \mathbb{R}^{m_2},$$

tels que les projections

$$u = X_1\mathbf{a}, \quad v = X_2\mathbf{b}$$

maximisent la corrélation de Pearson :

$$\rho = \frac{\mathrm{Cov}(u, v)}{\sqrt{\mathrm{Var}(u)\,\mathrm{Var}(v)}} = \frac{\mathbf{a}^\top \tilde{\Sigma}_{12}\,\mathbf{b}}{\sqrt{\mathbf{a}^\top \tilde{\Sigma}_{11}\,\mathbf{a}}\,\sqrt{\mathbf{b}^\top \tilde{\Sigma}_{22}\,\mathbf{b}}},$$

où

$$\tilde{\Sigma}_{11} = \tfrac{1}{n-1} X_1^\top X_1, \quad \tilde{\Sigma}_{12} = \tfrac{1}{n-1} X_2^\top X_2, \quad \tilde{\Sigma}_{22} = \tfrac{1}{n-1} X_1^\top X_2.$$

avec $\tilde{\Sigma}_{11}$ et $\tilde{\Sigma}_{22}$ les matrices de covariance respectives de X_1 et X_2, et $\tilde{\Sigma}_{12}$ la matrice de covariance croisée entre X_1 et X_2.

Les étapes de la CCA sont à retrouver en annexe : voir annexe E

3.2 Présentation de la PCCA

L'Analyse en Corrélation Canonique Probabiliste (PCCA) reformule la CCA classique comme un modèle gaussien à variables latentes, permettant d'estimer les paramètres par maximum de vraisemblance, ce résultat a été prouvé par Bach and Jordan [2005].

Modèle et notations

On suppose que les données centrées

$$X_1 \in \mathbb{R}^{n \times m_1}, \quad X_2 \in \mathbb{R}^{n \times m_2}$$

sont générées à partir d'une variable latente commune $z \in \mathbb{R}^{n \times d}$, avec :

$$\begin{cases} \mathbf{z}_i \sim \mathcal{N}(0, I_d), \\ \mathbf{x}_{1,i} \mid \mathbf{z}_i \sim \mathcal{N}(W_1 \mathbf{z}_i + \mu_1, \Psi_1), & W_1 \in \mathbb{R}^{m_1 \times d}, \quad \Psi_1 \succ 0, \\ \mathbf{x}_{2,i} \mid \mathbf{z}_i \sim \mathcal{N}(W_2 \mathbf{z}_i + \mu_2, \Psi_2), & W_2 \in \mathbb{R}^{m_2 \times d}, \quad \Psi_2 \succ 0, \end{cases}$$

pour $i = 1, \ldots, n$ et d le nombre de dimensions latentes. Avec :

- $W_1 \in \mathbb{R}^{m_1 \times d}$ et $W_2 \in \mathbb{R}^{m_2 \times d}$ sont les matrices de *chargements* (ou *loadings*) qui lient la variable latente z aux variables observées des matrices X_1 et X_2. Chaque colonne correspond à une direction latente.
- $\mu_1 \in \mathbb{R}^{m_1}$ et $\mu_2 \in \mathbb{R}^{m_2}$ sont les vecteurs de moyenne (centres) des variables observées de X_1 et X_2.
- $\Psi_1 \in \mathbb{R}^{m_1 \times m_1}$ et $\Psi_2 \in \mathbb{R}^{m_2 \times m_2}$ sont les matrices de covariance de X_1 et X_2, modélisant le bruit ou la variance non expliquée par la variable latente z. Elles sont supposées symétriques définies positives (ou semi-définies positives).

Objectif de la PCCA

Estimer les paramètres $(W_1, W_2, \mu_1, \mu_2, \Psi_1, \Psi_2)$ par maximum de vraisemblance, pour modéliser la dépendance entre X_1 et X_2 via la variable latente Z. Les étapes de la PCCA sont détaillées en annexe F.

EM PCCA

Voici les étapes de l'algorithme EM dans le cadre de la PCCA. :

$$W_{t+1} = \tilde{\Sigma} \Psi_t^{-1} W_t M_t \left(M_t + M_t W_t^\top \Psi_t^{-1} \tilde{\Sigma} \Psi_t^{-1} W_t M_t \right)^{-1}$$

$$\Psi_{t+1} = \begin{pmatrix} \left(\tilde{\Sigma} - \tilde{\Sigma} \Psi_t^{-1} W_t M_t W_{t+1}^\top \right)_{11} & 0 \\ 0 & \left(\tilde{\Sigma} - \tilde{\Sigma} \Psi_t^{-1} W_t M_t W_{t+1}^\top \right)_{22} \end{pmatrix}$$

J'ai retrouver les résultats de l'EM dans le cadre de la PCCA, car la preuve n'est pas donnée dans le papier de Bach and Jordan [2005]. Pour des raisons de clarté, la preuve a été mise en annexe B.

3.3 Présentation de la GPCCA

La *Generalized Probabilistic Canonical Correlation Analysis* (GPCCA), introduite par Yang and Li [2025] en 2025, est une généralisation de la PCCA, structurée en R groupes de variables au lieu de deux, de tailles respectives $m_1, \ldots, m_R$, totalisant $m = \sum_{r=1}^{R} m_r$ variables. On observe les données sous la forme d'une matrice $X^{\mathrm{NA}} \in \mathbb{R}^{n \times m}$, où n est le nombre d'individus, et certaines entrées peuvent être manquantes. Le masque de données manquantes est représenté par une matrice binaire $M \in \{0,1\}^{n \times m}$, où $M_{ij} = 1$ signifie que la valeur X_{ij} est manquante, et 0 sinon. GPCCA suppose que les données complètes X (non observées en pratique) sont générées à partir d'un espace latent de dimension réduite, via un modèle linéaire avec bruit gaussien.

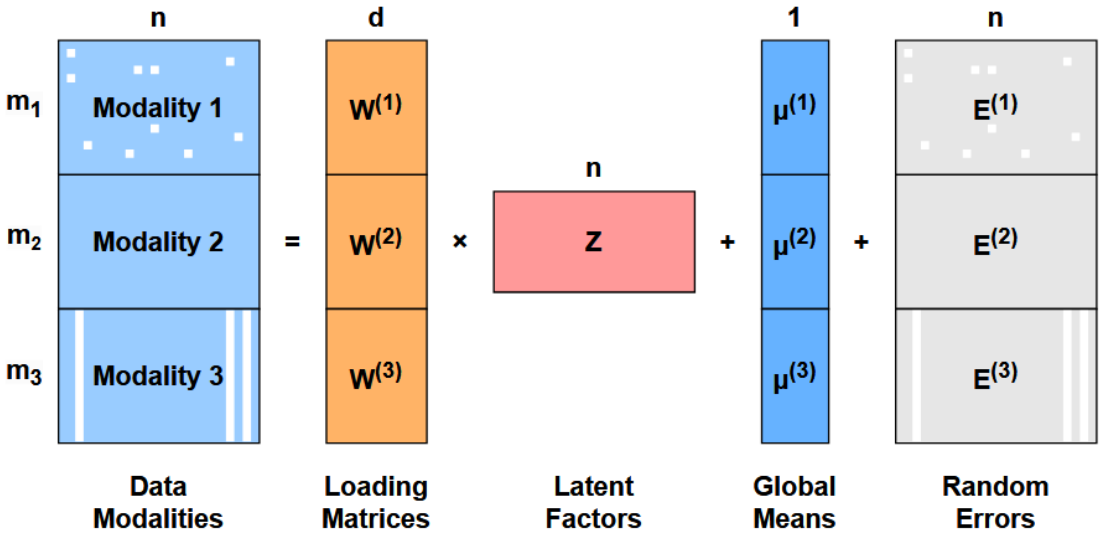


Figure 7: Illustration du modèle de la GPCCA

Composants du modèle GPCCA :

- $X \in \mathbb{R}^{n \times m}$: données complètes (non observées).
- $Z \in \mathbb{R}^{n \times d}$: variables latentes partagées, avec $d \ll m$.
- $W^r \in \mathbb{R}^{m_r \times d}$: matrices de projection spécifiques à chaque groupe $r \in \{1, \ldots, R\}$.
- $E^r \sim \mathcal{N}(0, \Sigma_r)$: bruit gaussien propre à chaque groupe.
- $X_r = ZW_r^\top + \epsilon_r$: génération des données du groupe r à partir de l'espace latent.

Algorithme EM-GPCCA : L'algorithme EM (Expectation-Maximization) permet d'estimer les paramètres du modèle malgré les données manquantes.

- **E-step** : Calcul de l'espérance de la log-vraisemblance complète par rapport à la distribution conditionnelle de Z donnée X_{obs} et les paramètres courants :

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z|X_{\mathrm{obs}},\theta^{(t)}} \left[\log p(X, Z \mid \theta)\right]$$

- **M-step** : Maximisation de $Q(\theta \mid \theta^{(t)})$ par rapport aux paramètres :

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$$

- **Prise en compte des données manquantes** : Les espérances sont calculées uniquement sur les valeurs observées, avec une intégration implicite sur les valeurs manquantes via la distribution conditionnelle du modèle.
- **Itération** : Répéter E et M jusqu'à convergence.

4 Simulation sur la GPCCA

L'objectif de cette simulation est d'évaluer l'influence de la dimension latente d dans la méthode GPCCA (Probabilistic Generalized Canonical Correlation Analysis)

On se place dans un contexte où la méthode est appliquée à des jeux de données comportant des valeurs manquantes générées selon deux schémas MNAR distincts. L'étude vise également à explorer dans quelle mesure ces schémas MNAR peuvent, en pratique, se rapprocher du cas MAR lorsque la dimension d est faible, c'est-à-dire lorsque la structure des données est simple et que les variables sont fortement corrélées entre elles. Pour cela, on considère les deux mécanismes MNAR suivants :

- un mécanisme dépendant du cluster d'appartenance N,
- un mécanisme auto-masqué dépendant directement de la valeur X_{ij}.

On va les comparer aux deux mécanismes MAR suivants :

- un mécanisme où on a une variable fixée pour X et le manque dépend de cette variable,
- un mécanisme où on a une variable fixée par jeu de données et le manque dans chaque jeu de données dépend de cette variable.

Dans tous les cas, si les blocs partagent une structure latente de faible dimension d, il devient possible d'approximer les mécanismes MNAR par des mécanismes MAR, dans la mesure où l'information manquante est reconstructible via les autres blocs. C'est le but de ce que nous voulons évaluer empiriquement. Dans quelle mesure une faible dimension d permet de compenser le caractère non ignorable du schéma de données manquantes ?

4.1 Génération des données complètes et des mécanismes

Génération des données complètes Les données complètes sont générées à l'aide de la fonction `generate_PGCCA_data`, qui produit un jeu de données multiblocs structuré autour de clusters latents. Chaque individu i reçoit un vecteur de variables X_i de taille m généré selon la distribution (`normal`), avec des moyennes dépendant de son cluster Ni. La sortie est un tableau $X \in \mathbb{R}^{n \times m}$ avec $m = \sum_r m_r$, accompagné de la colonne `cluster`.

Mécanismes simulés Deux types de mécanismes MNAR et deux mécanismes MAR sont implémentés pour générer des valeurs manquantes à partir des données complètes.

MNAR selon le cluster latent N

- La probabilité de bloc manquant dépend du cluster d'origine :

$$\mathbb{P}(M_{ij} = 1 \mid N_i = k) = p_k$$

- Chaque cluster k a un taux de valeurs manquantes p_k spécifié (ex. 0.1, 0.2, 0.4).
- Cela correspond à un mécanisme où certains groupes d'individus sont plus exposés au manque que d'autres.

MNAR auto-masqué (dépend de X_{ij})

- Chaque valeur est manquante avec une probabilité dépendant directement de sa valeur :

$$\mathbb{P}(M_{ij} = 1 \mid X_{ij}) = \frac{1}{1 + \exp(-aX_{ij} - b)}$$

- Il s'agit ici d'un manque auto-masqué, dans le cas du biais de déclaration par exemple.

MAR basé sur une variable observée unique

- Le mécanisme de données manquantes dépend d'une seule variable observée X_j, choisie arbitrairement (ici la première colonne du jeu de données).
- Pour chaque individu i, la probabilité que la variable X_{ij} soit observée est modélisée par :

$$\mathbb{P}(M_{ij} = 0 \mid X_j = x_j) = \frac{1}{1 + \exp(-ax_j - b)}$$

  où a est un paramètre de pente fixé, et b est choisi pour atteindre une proportion cible de données manquantes.

- Le paramètre b est déterminé par l'optimisation de:

$$b = \arg\min_b \left(\mathbb{E}[\mathbb{P}(M_{ij} = 0 \mid x_i)] - (1 - \text{prop\_missing})\right)^2$$

- Ce mécanisme est bien MAR, car la variable X_j utilisée pour contrôler le manque est toujours observée.
- Dans l'implémentation, toutes les variables sauf X_j (et certaines colonnes) peuvent être rendues manquantes.

MAR par modalité (une variable observée différente pour chaque bloc)

- Ici, chaque bloc de variables (modalité) est affecté par un mécanisme MAR propre, basé sur une variable observée spécifique à ce bloc.
- Soit R le nombre de blocs, et $X_j^{(r)}$ la variable de référence pour le bloc r.

- Pour chaque bloc r, on modélise la probabilité d'observation d'une variable X_{ij} (autre que $X_j^{(r)}$) par :

$$\mathbb{P}(M_{ij} = 0 \mid X_j^{(r)} = x_j^{(r)}) = \frac{1}{1 + \exp(-ax_j^{(r)} - b_r)}$$

  où b_r est ajusté séparément pour chaque bloc pour atteindre la proportion cible.

- Cela permet d'introduire des schémas de manque différents entre les modalités, tout en restant dans un cadre MAR (car $X_j^{(r)}$ est toujours observée).

- En pratique, la première variable de chaque bloc est utilisée comme $FX_j^{(r)}$, et les autres colonnes du bloc peuvent devenir manquantes.

4.2 Etapes de la simulation

Les étapes de la simulation sont les suivantes:

1. Générer un jeu de données complet avec la fonction `generate_PGCCA_data`, Le jeu de données est composé de trois blocs de 3, 4 et 5 variables et de $n = 2000$ individus répartis en 4 clusters de taille n_k.

2. Simuler 10 fois les schémas de données manquantes selon les quatre mécanismes distincts pour chaque proportion de données manquantes (15%, 30%, 60%) :

   - **MNAR** :
     - `MNAR_cluster (Clust)` : les valeurs manquantes dépendent du cluster latent.
     - `auto-masqué (Self)` : les valeurs manquantes dépendent de la valeur elle-même.
   - **MAR** (références pour comparaison) :
     - `MAR_single_variable (MAR1)` : une seule variable est entièrement observée.
     - `MAR_per_modality (MARR)` : une variable par bloc est entièrement observée (soit 3 au total).

3. Pour chaque dataset incomplet ainsi obtenu, appliquer la méthode `em_gpcca` en faisant varier la dimension latente $d \in \{(1, 2, 4, 6, 9, 12)\}$.

4. Pour chaque combinaison (schéma, proportion de données manquantes, d) :

   - Imputer les données manquantes avec la moyenne conditionnelle donnée par le modèle GPCCA.
   - Calculer la RMSE d'imputation par rapport aux vraies valeurs du jeu de données complet.
   - Appliquer un clustering (via `Mclust`) sur les variables latentes inférées dans le but de prédire les clusters auxquels appartiennent les individus. La performance de cette prédiction est ensuite évaluée en comparant l'appartenance aux clusters prédite à la vraie appartenance aux clusters N_i à l'aide de l'Indice de Rand Ajusté (ARI).

5. Comparer les performances (RMSE, ARI) des mécanismes **MNAR** à celles des mécanismes **MAR**, en particulier pour étudier :

   - si la dimension d fait que les performances se ressemblent sous MNAR (Self et Clust) et sous MAR (MAR1 et MARR).
   - l'influence de la proportion de données manquantes sur ce possible rapprochement.

Rappel méthodologique Dans cette section, nous évaluons deux types de performances : la capacité d'imputation des données manquantes, mesurée par la Root Mean Square Error (RMSE), et la qualité du regroupement en clusters, mesurée par l'Adjusted Rand Index (ARI).

Root Mean Square Error (RMSE). La RMSE permet de mesurer l'écart quadratique moyen entre les valeurs réelles X_{ij} et les valeurs prédites $\hat{X}_{ij}$ pour les données manquantes. Elle est définie comme suit :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \hat{X}_{ij})^2} \tag{1}$$

Plus la RMSE est faible, meilleure est la qualité de l'imputation.

Adjusted Rand Index (ARI). L'ARI permet de quantifier la similarité entre deux partitions : la partition correspondant aux vrais clusters et celle obtenue après clustering. Contrairement à l'indice de Rand simple, l'ARI est corrigé pour tenir compte du hasard. Il est défini par :

$$\text{ARI(a,b)} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \tag{2}$$

où :

- n_{ij} est le nombre d'éléments communs au cluster i de la partition réelle et au cluster j de la partition prédite ;
- $a_i = \sum_j n_{ij}$ est le nombre total d'éléments dans le cluster i de la partition réelle ;
- $b_j = \sum_i n_{ij}$ est le nombre total d'éléments dans le cluster j de la partition prédite.

L'ARI est compris entre -1 (pire accord) et 1 (accord parfait), avec 0 le score qu'on aurait obtenu en classifiant la partition aléatoirement.

4.3 Limites et difficultés rencontrées

Il y a plusieurs problèmes pour comparer les résultats, le principal est de savoir comment contrôler le pourcentage de données manquantes selon les schémas.

D'abord, expliquons comment le pourcentage de données manquantes peut être contrôlé pour chaque schéma. Dans notre simulaiton, prenons l'exemple lorsque le taux est fixé à 10%.

- Pour le cas MNAR auto-masqué (Self), la probabilité de manque dépend seulement de la valeur X_{ij} de la case elle-même. Il suffit donc dans la fonction logit de fixer a et de trouver le b optimal associé au pourcentage de manque.

- Pour le cas MNAR avec le clustering (Clust), il suffit de contrôler la somme des $p_k * n_k$ soit la probabilité de chaque cluster d'être manquant en fonction du nombre d'individu par cluster. Le cas trivial est de prendre la même probabilité pour chaque cluster. Dans cette simulation, on n'empêche que les cluster n'aient pas tous la même probabilité d'être manquants. Pour cela, on tire un poids aléatoire w associé aux clusters. Ensuite, on le normalise et au lieu de considérer p_k comme étant égal à x/n_k avec $x = n * 0.1$, on considère que $x = n * 0.1 * w$. On répète ce tirage de poids tant que l'écart type de p_k est inférieur à 0.01.

- Pour le cas MAR où uniquement une variable est fixée (MAR1), et le manque dépend de cette variable, on utilise aussi une logit avec le même raisonnement que pour le MNAR auto-masqué sur la colonne fixée.

- Pour le cas MAR où une variable est fixé par modalité (MARR), on utilise aussi une logit avec le même raisonnement que pour le MNAR auto-masqué mais cette fois sur les trois colonnes fixées.

Le soucis est le suivant. Lorsque l'on fixe une variable par modalité (soit R variables totalement observées pour les R blocs), cela réduit mécaniquement la proportion globale de données manquantes, car une part plus importante des variables est exclue du processus de masquage. Cela introduit une différence dans le taux de données manquantes entre les différents schémas.

Pour garantir une comparaison équitable des performances entre les différentes méthodes d'introduction de valeurs manquantes (MNAR auto-masqué, MAR avec une variable, MAR avec R variables, MNAR cluster), nous choisissons de fixer les mêmes R variables (une par modalité) comme entièrement observées dans chacun des quatre schémas.

Ce choix n'est pas réaliste dans une situation pratique (où la structure du manque peut varier selon le mécanisme sous-jacent), mais il permet ici de maintenir un pourcentage global de données manquantes comparable entre les différents scénarios, assurant ainsi une évaluation cohérente des performances des méthodes.

4.4 Résultats

La figure sur L'ARI est zoomée sur les dimensions latentes d de 4 à 12 en annexe K.

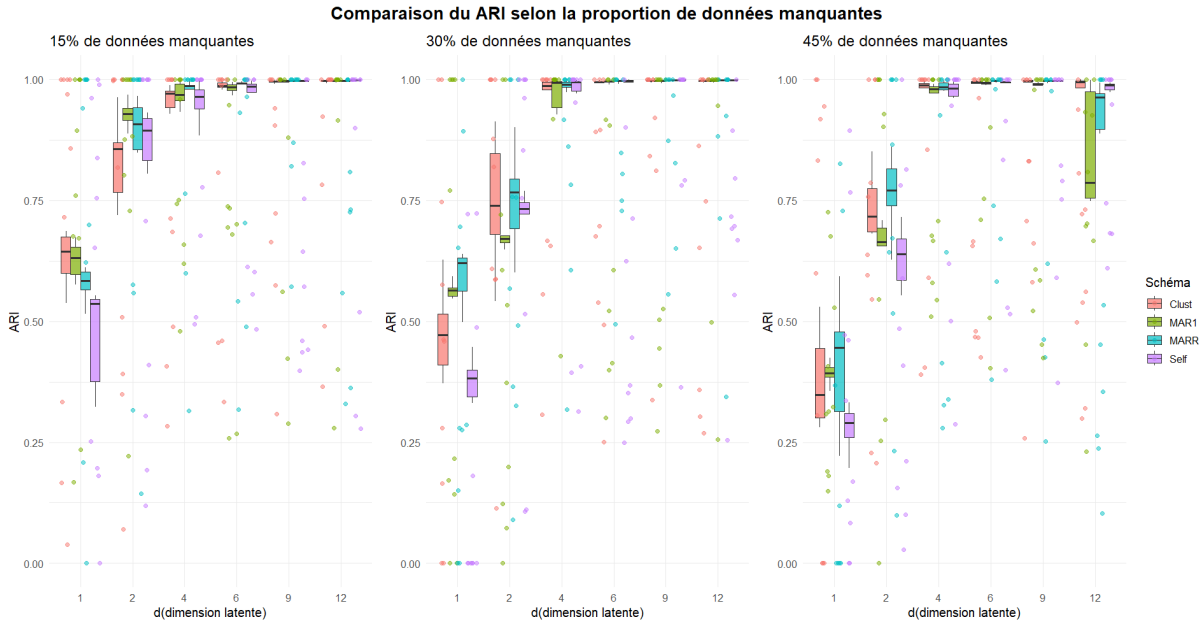


Figure 8: Indice de Rand Ajusté (ARI) selon les schémas et les dimensions latentes.

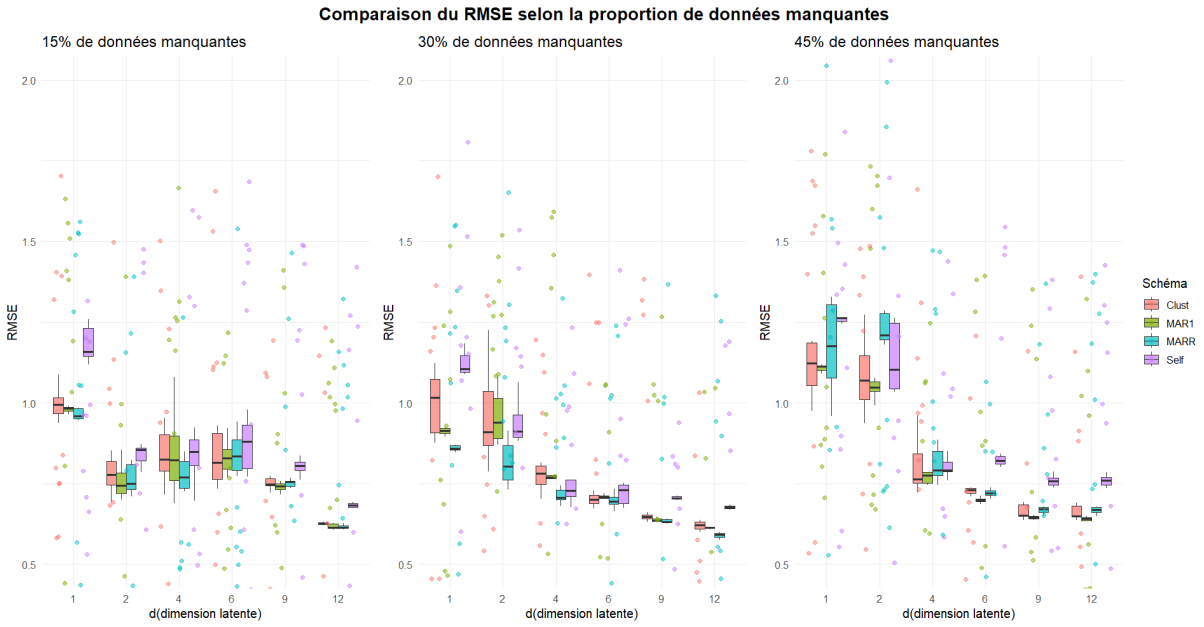


Figure 9: Erreur quadratique moyenne (RMSE) selon les schémas et les dimensions latentes.

4.5 Comparaison des performances selon le pourcentage de données manquantes

Pour la performance RMSE, on voit que plus le pourcentage de données manquantes est fort plus la RMSE est élevé. De façon plus précise on a les résultats suivants:

- **Pour le taux à 15% :** On peut voir que les meilleures RMSE sont atteintes sur les schémas MAR1, MARR et Clust avec des dimensions latentes d élevées ($d = 12$).

Les pires RMSE, sont atteintes surtout avec des dimensions faible ($d = 1$ ou $d = 2$) et principalement le schéma Self. On peut le voir sur les tableaux en annexe H.

- **Pour le taux à 30% :** On peut voir que les 10 meilleures RMSE sont atteintes avec des dimensions latentes élevées ($d = 12$) et avec les schémas MARR, Clust et MAR1. Les pires performances elles sont atteintes avec des dimensions faible ($d = 1$ ou $d = 2$) et le schéma Self et Clust principalement. On peut le voir sur les tableaux en annexe I.

- **Pour le taux à 45% :** Ici on peut voir contrairement aux autres que les meilleures RMSE sont atteintes avec des dimensions latentes élevées ($d = 9$ ou 12) et sur les schémas uniquement MAR1 et Clust. Les pires sont atteintes sur les schémas Self, Clust et MARR avec des dimensions faible ($d = 1, d = 2$ ou $d = 4$). On peut le voir sur les tableaux en annexe J.

Pour la performance ARI, on voit que le manque influe aussi sur la qualité du clustering même si c'est moins visible que pour la RMSE. Plus précisément, on a les résultats suivants:

- **Pour le taux à 15% :** Les meilleures performances de clustering sont atteintes avec avec des dimensions latentes élevées ($d = 9$ ou 12) et sur les schémas Self, MARR, MAR1 et Clust donc on a quelque chose d'assez équilibré. Les pires performances de clustering sont atteintes avec des dimensions faible (uniquement $d = 1$). Sur le schéma Self et MARR principalement. On peut le voir sur les tableaux en annexe K.

- **Pour le taux à 30% :** Les meilleures performances de clustering sont atteintes avec avec des dimensions latentes élevées ($d = 9$ ou 12) et sur les schémas Self,MARR, principalement donc on a quelque chose de moins équilibré qu'à 15% de manque. Les pires performances de clustering sont atteintes avec des dimensions faible ($d = 1$). Sur le schéma Self et Clust principalement. On peut le voir sur les tableaux en annexe L.

- **Pour le taux à 45% :** Les meilleures performances de clustering sont atteintes avec avec des dimensions latentes élevées ($d = 9$ ou 12) et sur les schémas Clust ,MARR principalement. Les pires performances de clustering sont atteintes avec des dimensions faible ($d = 1$). Sur le schéma Self et Clust principalement. On peut le voir sur les tableaux en annexe M.

De manière générale, plus le pourcentage de données manquantes augmente, plus la performance de reconstruction (RMSE) et la qualité du clustering (ARI) tendent à se dégrader. Toutefois, les résultats empiriques montrent que même sans modélisation explicite du mécanisme de masquage, certaines configurations MNAR n'entraînent pas nécessairement de pertes de performance pour l'`EM-GPCCA`. En particulier, le schéma Clust, où le masquage est structuré par cluster, conservent des performances proches des mécanismes MAR, notamment lorsque la dimension latente d est suffisante. À l'inverse, le schéma auto-masqué (Self) reste le plus difficile à traiter, ce qui reflète la complexité du biais introduit. Enfin, on observe que dans certains cas à 45% de données manquantes, les performances de clustering sous MNAR peuvent surpasser celles obtenues sous MAR, suggérant que le modèle parvient parfois à exploiter la structure implicite du masque.

4.6 Discussions et perspectives

Voici plusieurs pistes d'amélioration ou d'extension pour les expériences menées:

- **Augmenter la dimension des données simulées:** les résultats pourraient être plus représentatifs en travaillant avec des données de plus grande dimension au lieu d'une matrice de donnée avec uniquement 12 variables (3,4,5).
- **Introduire un autre schéma MNAR:** dans les expériences actuelles, seuls deux mécanisme MNAR ont été utilisés. Celui basé sur les clusters est moins complexe. Etant données les performances de reconstruction avec l'ARI, il est assez raisonnable de penser que le mécanisme est simple.
- **Varier davantage les proportions de données manquantes:** on aurait pu explorer des proportions intermédiaires (par exemple 20% ou 35%). Cela permettrait de mieux analyser l'évolution des performances selon le taux de données manquantes, pour avoir des résultats plus lisses.
- **Répéter les expériences sur des jeux de données réels:** cela permettrait de compléter les résultats obtenus par simulation avec une validation empirique sur des cas concrets.

5 Conclusion

En conclusion de ce stage, je dirais que c'était un excellent moyen de découvrir le monde de la recherche, de part le sujet très ouvert dans la mesure où le projet s'est cadré au fur et à mesure du stage. Le cadre joue aussi beaucoup, ayant côtoyer beaucoup de doctorants, que ce soit dans mon bureau ou durant les pauses du midi. C'était un environnement stimulant où j'ai pu poser mes questions, observer et découvrir le métier de chercheur, ce qui m'a beaucoup motivé et conforté dans l'idée de trouver une thèse.

References

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. *na*, 2005.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006. ISBN 978-0-387-31073-2.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL https://doi.org/10.18637/jss.v045.i03.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *"NA"*, 39(1):1–22, 1977. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176344552. URL https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-Jackknife/10.1214/aos/1176344552.full. Publisher: Institute of Mathematical Statistics.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi: 10.2307/2333955.

Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *NA*, 70:1–31, 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i01. URL https://doi.org/10.18637/jss.v070.i01.

Julie Josse, François Husson, and Jérôme Pagès. Gestion des données manquantes en analyse en composantes principales. *NA*, 150(2):28–51, 2009. ISSN 2102-6238. URL https://www.numdam.org/item/JSFS_2009__150_2_28_0/.

Roderick Little and Donald Rubin. *Statistical Analysis With MissingData*. Wiley, April 2019. ISBN 9780470526798. doi: 10.1002/9781119482260. URL https://doi.org/10.1002/9781119482260.fmatter.

Donald B. Rubin. Inference and missing data. *NA*, 63(3):581–592, 1976. ISSN 0006-3444. doi: 10.2307/2335739. URL https://www.jstor.org/stable/2335739. Publisher: [Oxford University Press, Biometrika Trust].

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. ISBN 0-471-60225-6.

Tianjian Yang and Wei Vivian Li. Generalized probabilistic canonical correlation analysis for multi-modal data integration with full or partial observations, 2025. URL http://arxiv.org/abs/2504.11610.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022.

A Exercice EM bivarié

A.1 Initialisation de l'algorithme EM

Avant de démarrer les itérations de l'algorithme EM, il est nécessaire de spécifier des valeurs initiales pour les paramètres du modèle. Dans le cas Gaussien, les paramètres à initialiser sont pour l=0:

- Le vecteur des moyennes $\mu^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)})^T$
- La matrice de covariance $\Sigma^{(0)} = \begin{pmatrix} \sigma_{11}^{(0)} & \sigma_{12}^{(0)} \\ \sigma_{21}^{(0)} & \sigma_{22}^{(0)} \end{pmatrix}$

On a plusieurs choix pour l'initialisation :

- **Moyennes et covariances empiriques** calculées sur les observations complètes seulement.
- **Imputation simple** des données manquantes (par la moyenne, par exemple), suivie du calcul des estimateurs classiques. C'est ce que nous allons faire dans notre cas.
- **Choix arbitraire** mais plausible, basé sur une connaissance a priori du problème.

Ces valeurs initiales constituent $\theta^{(0)}$ et serviront d'initialisation pour les itérations E-M suivantes.

A.2 Étape E de l'algorithme EM

L'objectif de l'étape E est de calculer l'espérance de la log-vraisemblance de $\theta = (\mu, \Sigma)$ et X sachant les données observés X_{obs} : avec θ inconnu et en conditionnant la loi des données manquantes sur $\theta^{(0)}$.

$$Q(\theta; \theta^{(0)}) = \mathbb{E}_{X_{miss(M)}|X_{obs(M)},\theta^{(0)}} \left[\log L(\theta; X) \mid X_{obs(M)}, \theta^{(0)} \right] =$$

$$\int \log L(\theta; X_{\text{obs}(M)}, X_{\text{miss}(M)}) \cdot p(X_{\text{miss}(M)} \mid X_{\text{obs}(M)}, \theta^{(0)}) \, dX_{\text{miss}(M)}$$

où L est la vraisemblance du jeu de données complet, et l'espérance est prise par rapport à la loi conditionnelle des données manquantes, connaissant les données observées et les paramètres initiaux $\theta^{(0)}$.

A.2.1 Vraisemblance complète

On suppose que chaque observation $x_i = (x_{i1}, x_{i2})^\top$ suit une loi normale bivariée, les x_i sont indépendant et identiquement distribué :

$x_i \sim \mathcal{N}(\mu, \Sigma)$ avec $\theta = (\mu, \Sigma)$
La densité jointe associée est alors :

$$f(x_i; \theta) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^\top (\Sigma)^{-1}(x_i - \mu) \right)$$

Par indépendance des x_i, la vraisemblance complète s'écrit comme un produit :

$L(\theta; X) \;=\; \prod_{i=1}^{n} f(x_i; \theta),$

$L(\theta; X) = \left(\frac{1}{2\pi|\Sigma|^{1/2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top(\Sigma)^{-1}(x_i-\mu)\right)$

A.2.2 Log-vraisemblance complète

On obtient donc la log-vraisemblance complète suivante :

$$\log L(\theta; X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top(\Sigma)^{-1}(x_i-\mu)$$

Le terme $-\frac{n}{2}\log(2\pi)$ étant constant, il peut être ignoré dans l'optimisation.

On rappelle que :

$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad x_i = \begin{pmatrix}x_{i1}\\ x_{i2}\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$$

Le déterminant de Σ est donné par :

$$|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$$

L'inverse de Σ s'écrit :

$$(\Sigma)^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{pmatrix}$$

On développe le terme quadratique :

$$(x_i-\mu)^\top(\Sigma)^{-1}(x_i-\mu) = \frac{1}{|\Sigma|}\left[\sigma_{22}(x_{i1}-\mu_1)^2 - 2\sigma_{12}(x_{i1}-\mu_1)(x_{i2}-\mu_2) + \sigma_{11}(x_{i2}-\mu_2)^2\right]$$

Ainsi, à une constante additive près, la log-vraisemblance s'écrit :

$$\log L(\theta; X) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2|\Sigma|}\sum_{i=1}^{n}\left[\sigma_{22}(x_{i1}-\mu_1)^2 - 2\sigma_{12}(x_{i1}-\mu_1)(x_{i2}-\mu_2) + \sigma_{11}(x_{i2}-\mu_2)^2\right]$$

Ainsi :

$$\begin{aligned} Q(\theta;\theta^{(0)}) &= \mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}\Big[-\tfrac{n}{2}\log|\Sigma| - \frac{1}{2|\Sigma|}\sum_{i=1}^{n}\Big(\sigma_{22}(x_{i1}-\mu_1)^2 \\ &\qquad - 2\,\sigma_{12}(x_{i1}-\mu_1)(x_{i2}-\mu_2) + \sigma_{11}(x_{i2}-\mu_2)^2\Big)\Big] \\ &= -\tfrac{n}{2}\log|\Sigma| - \frac{1}{2|\Sigma|}\sum_{i=1}^{n}\Big[\sigma_{22}(x_{i1}-\mu_1)^2 - 2\,\sigma_{12}(x_{i1}-\mu_1)\,\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}[x_{i2}-\mu_2] \\ &\qquad + \sigma_{11}\,\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}[(x_{i2}-\mu_2)^2]\Big]. \end{aligned}$$

Dans l'étape E, l'espérance est prise par rapport à la loi conditionnelle $X_{\text{miss}} \mid X_{\text{obs}}, \theta^{(0)}$.
x_{i1} est une quantité fixe (non aléatoire).
On rappelle que pour un individu i :

- S'il est **complet**, alors :

$$x_i^{\text{miss}} = \emptyset, \quad x_i^{\text{obs}} = (x_{i1}, x_{i2})$$

- Si la partie x_{i2} est **manquante**, alors :

$$x_i^{\text{miss}} = x_{i2}, \quad x_i^{\text{obs}} = x_{i1}$$

Par la linéarité de l'espérance conditionnelle, pour le terme de la forme $(x_{i1}-\mu_1)\,(x_{i2}-\mu_2)$, on a :

$$\begin{aligned}
\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}\Big[(x_{i1}-\mu_1)\,(x_{i2}-\mu_2)\Big] &= (x_{i1}-\mu_1)\,\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}\Big[x_{i2}-\mu_2\Big] \\
&= (x_{i1}-\mu_1)\,\mathbb{E}\Big[x_{i2}-\mu_2 \mid x_{i1},\theta^{(0)}\Big] \\
&= (x_{i1}-\mu_1)\,(\mathbb{E}\Big[x_{i2} \mid x_{i1},\theta^{(0)}\Big]-\mu_2) \\
&= (x_{i1}-\mu_1)\,\left(\int x_{i2}\cdot p(x_{i2}^{miss} \mid x_{i1},\theta^{(0)})\,dx_{i2}^{miss}-\mu_2\right)
\end{aligned}$$

où

$$\int x_{i2}\cdot p(x_{i2}^{miss} \mid x_{i1},\theta^{(0)})\,dx_{i2}^{miss} = \begin{cases} x_{i2} & \text{si } m_{i2}=0 \text{ (x_{i2} observé)} \\ \Big(\mathbb{E}[x_{i2} \mid x_{i1},\theta^{(0)}]\Big) & \text{si } m_{i2}=1 \text{ (x_{i2} manquant)} \end{cases}$$

Pour le terme $\sigma_{11}(x_{i2}-\mu_2)^2$, on a :

$$\begin{aligned}
\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}\Big[\sigma_{11}(x_{i2}-\mu_2)^2\Big] &= \sigma_{11}\,\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},\,\theta^{(0)}}\Big[(x_{i2}-\mu_2)^2\Big] \\
&= \sigma_{11}\,\mathbb{E}\Big[(x_{i2}-\mu_2)^2 \mid x_{i1},\theta^{(0)}\Big] \\
&= \sigma_{11}\,\Big(\mathbb{E}\Big[x_{i2}^2-2\mu_2 x_{i2} \mid x_{i1},\theta^{(0)}\Big]+\mu_2^2\Big) \\
&= \sigma_{11}\,\Big(\mathbb{E}[x_{i2}^2 \mid x_{i1},\theta^{(0)}]-2\mu_2\,\mathbb{E}[x_{i2} \mid x_{i1},\theta^{(0)}]+\mu_2^2\Big) \\
&= \sigma_{11}\,\left(\int x_{i2}^2\,p(x_{i2}^{miss} \mid x_{i1},\theta^{(0)})dx_{i2}^{miss}-2\mu_2\int x_{i2}\,p(x_{i2}^{miss} \mid x_{i1},\theta^{(0)})dx_{i2}^{miss}\right.
\end{aligned}$$

où

$$\int x_{i2}^k\cdot p(x_{i2}^{miss} \mid x_{i1},\theta^{(0)})\,dx_{i2}^{miss} = \begin{cases} x_{i2}^k & \text{si } m_{i2}=0 \text{ (x_{i2} observé)} \\ \mathbb{E}[x_{i2}^k \mid x_{i1},\theta^{(0)}] & \text{si } m_{i2}=1 \text{ (x_{i2} manquant)} \end{cases} \quad \text{avec } k=1 \text{ ou } 2.$$

A.2.3 Loi conditionnelle de $X_{i2} \mid X_{i1}$ en $\theta^{(0)}$

On rappelle que chaque observation $x_i = (x_{i1}, x_{i2})^\top \sim \mathcal{N}(\mu, \Sigma)$, avec :

$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$$

Alors, la loi conditionnelle de $X_{i2} \mid X_{i1}=x_{i1}$ en $\theta^{(0)}$ est donnée par :

$$X_{i2} \mid X_{i1}=x_{i1} \sim \mathcal{N}\left(\mu_2^{(0)}+\tfrac{\sigma_{21}^{(0)}}{\sigma_{11}^{(0)}}(x_{i1}-\mu_1^{(0)}),\ \sigma_{22}^{(0)}-\tfrac{{\sigma_{21}^{(0)}}^2}{\sigma_{11}^{(0)}}\right)$$

A.2.4 Moments conditionnels et imputation

On impute alors les valeurs de x_{i2} avec les moments conditionnels d'ordre 1 et 2:

- **Espérance conditionnelle de X_{i2} :**

$$\hat{x}_{i2}^{(0)} = \mathbb{E}[X_{i2} \mid X_{i1} = x_{i1}] = \begin{cases} x_{i2} & \text{si } m_{i2} = 0 \quad (x_{i2} \text{ observé}) \\ \mu_2^{(0)} + \dfrac{\sigma_{21}^{(0)}}{\sigma_{11}^{(0)}}(x_{i1} - \mu_1^{(0)}) & \text{si } m_{i2} = 1 \quad (x_{i2} \text{ manquant}) \end{cases}$$

- **Moment d'ordre 2 de X_{i2} :**

$$\mathbb{E}[X_{i2}^2 \mid X_{i1} = x_{i1}] = (\hat{x}_{i2}^{(0)})^2 + \begin{cases} 0 & \text{si } m_{i2} = 0 \quad (x_{i2} \text{ observé}) \\ \tau^{(0)^2} & \text{si } m_{i2} = 1 \quad (x_{i2} \text{ manquant}) \end{cases}$$

avec $\tau^{(0)^2} = \sigma_{22}^{(0)} - \dfrac{(\sigma_{21}^{(0)})^2}{\sigma_{11}^{(0)}}$.

On peut donc écrire, en tenant compte de la nature observée ou manquante de chaque x_{i2}, l'expression suivante :

$$Q(\theta, \theta^{(0)}) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2|\Sigma|}\sum_{i=1}^{n}\left[\sigma_{22}(x_{i1}-\mu_1)^2 - 2\sigma_{12}(x_{i1}-\mu_1)(\hat{x}_{i2}^{(0)}-\mu_2) + \sigma_{11}\left((\hat{x}_{i2}^{(0)}-\mu_2)^2 + \tau^{(0)^2}\right)\right]$$

Il s'agit de la "valeur espérée" de la log-vraisemblance complète à l'étape 1 en conditionnant X_{miss} selon X_{obs} avec le paramètres $\theta^{(0)}$.

A.3 Étape M de l'algorithme EM

L'étape M de l'algorithme EM consiste à maximiser la quantité :

$$Q(\theta, \theta^{(0)}) = \mathbb{E}\left[\log L(\theta; X) \mid X^{\text{obs}}, \theta^{(0)}\right]$$

par rapport aux paramètres $\theta^{(0)} = (\mu^{(0)}, \Sigma^{(0)})$.

On cherche donc :

$$\theta^{(1)} = \arg\max_{\theta} Q(\theta, \theta^{(0)})$$

où :

$$Q(\theta, \theta^{(0)}) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2|\Sigma|}\sum_{i=1}^{n}\left[\sigma_{22}(x_{i1}-\mu_1)^2 - 2\sigma_{12}(x_{i1}-\mu_1)(\hat{x}_{i2}^{(0)}-\mu_2) + \sigma_{11}\left((\hat{x}_{i2}^{(0)}-\mu_2)^2 + \hat{x}_{i2}^{(0)}\right)\right]$$

Les quantités $\hat{x}_{i2}^{(0)}$ et $\tau^{(0)^2}$ sont connues à cette étape car elles dépendent de $\theta^{(0)}$. Ainsi, Q est une fonction explicite de $\mu_1^{(0)}, \mu_2^{(0)}, \Sigma^{(0)}$, que l'on peut dériver pour obtenir les nouveaux estimateurs des paramètres :

- La nouvelle moyenne $\mu^{(1)}$ est obtenue en moyennant les données observées et les espérances conditionnelles :

$$\mu^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} x_{i1} \\ \hat{x}_{i2}^{(0)} \end{pmatrix}$$

- La nouvelle matrice de covariance $\Sigma^{(1)}$ est estimée à partir de la matrice empirique des résidus, en tenant compte de la variance conditionnelle $\tau^{(0)^2}$ pour les données manquantes :

$$\Sigma^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\begin{pmatrix} x_{i1} \\ \hat{x}_{i2}^{(0)} \end{pmatrix} - \mu^{(1)} \right) \left(\begin{pmatrix} x_{i1} \\ \hat{x}_{i2}^{(0)} \end{pmatrix} - \mu^{(1)} \right)^{\top} + \begin{pmatrix} 0 & 0 \\ 0 & \tau^{(0)^2} \end{pmatrix} \right]$$

Ces nouvelles valeurs $\mu^{(1)}$ et $\Sigma^{(1)}$ définissent le nouvel estimateur $\theta^{(1)}$. On peut alors réinjecter ces paramètres dans l'étape E suivante, et ainsi de suite, jusqu'à convergence.

Un critère de convergence peut être :

- **Critère sur la log-vraisemblance** : on arrête lorsque l'augmentation de la log-vraisemblance devient négligeable, c'est-à-dire :

$$\left| \log L(\theta^{(t+1)}; X) - \log L(\theta^{(t)}; X) \right| < \varepsilon$$

  pour un seuil $\varepsilon > 0$.

- **Critère sur les paramètres** : on peut aussi vérifier la stabilité des paramètres estimés :

$$\left\| \theta^{(t+1)} - \theta^{(t)} \right\| < \varepsilon$$

  où $\| \cdot \|$ désigne une norme (norme euclidienne ou norme Frobenius pour la matrice Σ).

Une fois l'un de ces critères satisfait, on considère que l'algorithme a convergé et on retourne les estimations finales $\theta^{(\hat{t})}$.

B Preuve EM PCCA

Preuve des résultats de Bach and Jordan [2005]

$$p(X, Z; \theta) \;=\; \log p(Z) \;+\; \log p(X \mid Z; \theta).$$

Le terme $\log p(Z)$**.**
Comme $Z = (z_1, \ldots, z_n$ et $z_i \sim \mathcal{N}(0, I_d)$ indépendants, on a

$$p(Z) \;=\; \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}} \exp\Big(-\tfrac{1}{2}\, z_i^\top z_i\Big).$$

Donc

$$\log p(Z) \;=\; \sum_{i=1}^{n}\Big[-\tfrac{d}{2}\log(2\pi) \;-\; \tfrac{1}{2}\, z_i^\top z_i\Big] \;=\; -\frac{nd}{2}\log(2\pi) \;-\; \frac{1}{2}\sum_{i=1}^{n} z_i^\top z_i.$$

Le terme $\log p(X \mid Z; \theta)$**.**
On rappelle que, conditionnellement à z_i,

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \sim \mathcal{N}\Big(W\, z_i + \mu,\; \Psi\Big) \quad \text{avec } W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix},\; \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\; \Psi = \begin{pmatrix} \Psi_1 & 0 \\ 0 & \Psi_2 \end{pmatrix}.$$

La densité conjointe de $X \mid Z$ s'écrit

$$p(X \mid Z; \theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{m/2}\,|\Psi|^{1/2}} \exp\Big(-\tfrac{1}{2}\,(x_i - Wz_i - \mu)^\top \Psi^{-1}(x_i - Wz_i - \mu)\Big),$$

où $m = m_1 + m_2$. En prenant le logarithme :

$$\log p(X \mid Z; \theta) = \sum_{i=1}^{n}\Big[-\tfrac{m}{2}\log(2\pi) \;-\; \tfrac{1}{2}\log|\Psi| \;-\; \tfrac{1}{2}\,(x_i - Wz_i - \mu)^\top \Psi^{-1}(x_i - Wz_i - \mu)\Big].$$

Soit, en regroupant les constantes :

$$\log p(X \mid Z; \theta) = -\frac{n\,m}{2}\log(2\pi) \;-\; \frac{n}{2}\log|\Psi| \;-\; \frac{1}{2}\sum_{i=1}^{n}(x_i - Wz_i - \mu)^\top \Psi^{-1}(x_i - Wz_i - \mu).$$

Log-vraisemblance complète.
En additionnant les deux parties, on obtient

$$\log p(X, Z; \theta) = -\frac{1}{2}\sum_{i=1}^{n} z_i^\top z_i - \frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}(x_i - Wz_i - \mu)^\top \Psi^{-1}(x_i - Wz_i - \mu) \quad (3)$$

Les termes en $\log(2\pi)$ peuvent être ignorés dans l'EM, puisqu'ils ne dépendent pas de θ.

Pour la fonction $Q(\theta, \theta^{(t)})$ on prend l'espérance conditionnelle sous la loi de $Z \mid X;\, \theta^{(t)}$.

$$p\Big(Z \mid X;\, \theta^{(t)}\Big) \;=\; \frac{p(Z)\; p\Big(X \mid Z;\, \theta^{(t)}\Big)}{p\Big(X;\, \theta^{(t)}\Big)} \;=\; \frac{p(Z)\; p\Big(X \mid Z;\, \theta^{(t)}\Big)}{\displaystyle\int p(Z)\; p\Big(X \mid Z;\, \theta^{(t)}\Big)\, \mathrm{d}Z}\,.$$

$$p\Big(Z \mid X;\, \theta^{(t)}\Big) \;\equiv\; \prod_{i=1}^{n} \mathcal{N}\Big(z_i \mid \xi_i,\, M^{(t)}\Big),$$

où l'on montre (formules de la loi normale conditionnelle Bishop [2006]) que

$$M^{(t)} \;=\; \Big(I_d + W^{(t)\top} \Psi^{(t)-1} W^{(t)}\Big)^{-1}, \qquad \xi_i \;=\; M^{(t)}\, W^{(t)\top}\, \Psi^{(t)-1}\, \Big(x_i - \mu^{(t)}\Big). \tag{4}$$

Pour chaque $i = 1, \ldots, n$, on définit alors(Bishop [2006]) :

$$\mathbb{E}[z_i \mid x_i; \theta^{(t)}] = \xi_i, \qquad \mathbb{E}[z_i z_i^\top \mid x_i; \theta^{(t)}] = M^{(t)} + \xi_i \xi_i^\top \tag{5}$$

On rappelle que :

$$\mathbb{E}\left[zz^\top\right] = Cov(z) + \mathbb{E}[z]\,\mathbb{E}[z]^\top$$

On utilise alors les identités suivantes(papier sur les formes quadratiques) :

- Pour tout vecteur z et toute matrice A :

$$z^\top A z = \mathrm{tr}(A z z^\top).$$

- En prenant l'espérance de chaque côté :

$$\mathbb{E}[z^\top A z] = \mathbb{E}[\mathrm{tr}(A z z^\top)] = \mathrm{tr}(A\, \mathbb{E}[z z^\top]).$$

- $$\mathbb{E}[z z^\top] = \mathrm{Cov}(z) + \mathbb{E}[z]\,\mathbb{E}[z]^\top,$$

  donc,

$$\mathbb{E}[z^\top A z] = \mathrm{tr}(A\, \mathrm{Cov}(z)) + \mathbb{E}[z]^\top A\, \mathbb{E}[z].$$

On peut alors écrire la fonction Q en remplaçant chaque z_i par$\mathbb{E}_{Z|X;\,\theta^{(t)}}[z_i]$:

- Pour chaque $i = 1, \ldots, n$, l'espérance conditionnelle est donnée par

$$\mathbb{E}\Big[z_i \mid x_i; \theta^{(t)}\Big] = \xi_i,$$

- Et la matrice des moments d'ordre 2 :

$$\mathbb{E}\Big[z_i z_i^\top \mid x_i; \theta^{(t)}\Big] = \mathrm{Cov}(z_i \mid x_i) + \mathbb{E}[z_i \mid x_i]\mathbb{E}[z_i \mid x_i]^\top = M^{(t)} + \xi_i \xi^{(i)\top}.$$

- On utilise donc :

$$\mathbb{E}\Big[(z_i)^\top A z_i \mid x_i\Big] = \mathrm{tr}\Big(A\, \mathbb{E}[z_i z_i^\top \mid x_i]\Big) = \mathrm{tr}\Big(A(M^{(t)} + \xi_i \xi^{(i)\top})\Big).$$

En remplaçant A par $W^{(t)^\top}\Psi^{-1}W^{(t)}$, on obtient :

$$\mathbb{E}\Big[(z_i)^\top W^{(t)^\top}\Psi^{-1}W^{(t)}\,z_i\Big] = \mathrm{tr}\Big(W^{(t)^\top}\Psi^{-1}W^{(t)}\,(M^{(t)} + \xi_i\xi^{(i)\top})\Big).$$

Or, en utilisant la propriété $\mathrm{tr}(B\,C) = \mathrm{tr}(C\,B)$, on écrit donc également :

$$\mathrm{tr}\Big(\Psi^{-1}\,W^{(t)}\,(M^{(t)} + \xi_i\xi^{(i)\top})\,W^{(t)^\top}\Big).$$

Ce qui donne une fonction $Q(\theta, \theta^{(t)})$ de la forme :

$$\begin{aligned}
Q(\theta,\theta^{(t)}) &= \mathbb{E}_{Z|X;\,\theta^{(t)}}\Big[\log p(X,Z;\theta)\Big] \\
&= -\frac{n}{2}\,\log|\Psi| \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\Big[(x_i-\mu)^\top\Psi^{-1}(x_i-\mu) - 2\,\xi^{(i)\top}W^{(t)^\top}\Psi^{-1}(x_i-\mu) \\
&\qquad\qquad + \mathrm{tr}\Big(\Psi^{-1}W^{(t)}(M^{(t)} + \xi_i\xi^{(i)\top})W^{(t)^\top}\Big)\Big] \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}(M^{(t)} + \xi_i\xi^{(i)\top}).
\end{aligned}$$

Étape M On maximise

$$Q(\theta,\theta^{(t)}) = \mathbb{E}_{Z|X;\theta^{(t)}}\Big[\log p(X,Z;\theta)\Big]$$

par rapport à $\theta = (\mu, W, \Psi)$. Pour alléger les notations, on pose pour chaque i :

$$S_i \;=\; \mathbb{E}[z_i z_i^\top \mid x_i;\theta^{(t)}] \;=\; M^{(t)} + \xi_i\xi^{(i)\top}.$$

De plus, notons

$$\widetilde{\Sigma} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu^{(t)})(x_i-\mu^{(t)})^\top, \quad A \;=\; \sum_{i=1}^{n}(x_i-\mu)\,\xi_i^\top, \quad B \;=\; \sum_{i=1}^{n}S_i.$$

Mise à jour de μ.

$$Q = -\frac{n}{2}\log|\Psi| \;-\; \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top\Psi^{-1}(x_i-\mu) + \text{const.}$$

Pour maximiser Q, on annule le gradient par rapport à μ :

$$\frac{\partial Q}{\partial\mu} = \frac{\partial}{\partial\mu}\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top\Psi^{-1}(x_i-\mu)\right) = \Psi^{-1}\sum_{i=1}^{n}(x_i-\mu).$$

En posant cette dérivée égale à zéro :

$$\Psi^{-1}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad\Longrightarrow\quad \sum_{i=1}^{n}(x_i-\mu) = 0 \quad\Longrightarrow\quad n\mu = \sum_{i=1}^{n}x_i \quad\Longrightarrow\quad \mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}x_i.$$

Ici on a toujours le même résultat pour μ, dans le cas de EM, vu qu'on imputera à chaque itération, on sera obligé de mettre à jour les moyennes empiriques.

Mise à jour de $W^{(t)}$. On prend les termes de Q qui dépendent de $W^{(t)}$:

$$Q_W = -\frac{1}{2}\sum_{i=1}^{n}\Big[-2\,\xi_i^\top W^{(t)^\top}\Psi^{-1}(x_i-\mu) + \mathrm{tr}\Big(\Psi^{-1}W^{(t)}\,S_i\,W^{(t)^\top}\Big)\Big].$$

On peut écrire

$$Q_W = \mathrm{tr}\Big((\sum_i (x_i-\mu)\xi_i^\top)\Psi^{-1}W^{(t)^\top}\Big) - \tfrac{1}{2}\mathrm{tr}\Big(\Psi^{-1}W^{(t)}\,(\sum_i S_i)\,W^{(t)^\top}\Big) + \text{const.}$$

En utilisant $\mathrm{tr}(UV^\top) = \mathrm{tr}(VU^\top)$ et la dérivée $\partial\mathrm{tr}(CW^{(t)}W^{(t)^\top})/\partial W^{(t)} = 2CW^{(t)}$, on obtient

$$\frac{\partial Q}{\partial W^{(t)}} = \Psi^{-1}A^\top - \Psi^{-1}W^{(t)}\,B \overset{!}{=} 0 \quad\Longrightarrow\quad A^\top = W^{(t)}\,B \quad\Longrightarrow\quad W^{(t+1)} = A^\top B^{-1}.$$

Soit, explicitement en remplacant A et B par leurs expressions,

$$W^{(t+1)} = \Big[\sum_{i=1}^{n}(x_i-\mu^{(t+1)})\,\xi_i^\top\Big]\,\Big[\sum_{i=1}^{n}S_i\Big]^{-1}.$$

Si on écrit en remplacant S_i et ξ_i par leur vrai expression, on a :

$$\begin{aligned}W^{(t+1)} = &\left[\sum_{i=1}^{n}(x_i-\mu^{(t+1)})\left(M^{(t)}W^{(t)\top}\Psi^{(t)-1}(x_i-\mu^{(t)})\right)^\top\right]\\ &\cdot\left[\sum_{i=1}^{n}\left(M^{(t)}+M^{(t)}W^{(t)\top}\Psi^{(t)-1}(x_i-\mu^{(t)})(x_i-\mu^{(t)})^\top\Psi^{(t)-1}W^{(t)}M^{(t)}\right)\right]^{-1}.\end{aligned}$$

En remplacant $\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu^{(t)})(x_i-\mu^{(t)})^\top$ par $\widetilde{\Sigma}$, on obtient la forme suivante :

$$W^{(t+1)} = \widetilde{\Sigma}\Psi^{(t)-1}W^{(t)}M^{(t)}\left[M^{(t)}+M^{(t)}W^{(t)\top}\Psi^{(t)-1}\widetilde{\Sigma}\Psi^{(t)-1}W^{(t)}M^{(t)}\right]^{-1}.$$

Mise à jour de $\Psi^{(t)}$. On considère ici les termes de Q dépendant de Ψ :

$$Q_\Psi = -\frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\|x_i-\mu-Wz_i\|^2_{\Psi^{-1}}\mid x_i\right].$$

On développe :

$$\begin{aligned}\mathbb{E}\Big[(x_i-\mu-Wz_i)(x_i-\mu-Wz_i)^\top\mid x_i\Big] = &(x_i-\mu)(x_i-\mu)^\top - W\xi_i(x_i-\mu)^\top\\ &-(x_i-\mu)\xi_i^\top W^\top + WS_iW^\top.\end{aligned}$$

En sommant sur tous les i, on obtient :

$$\begin{aligned}Q_\Psi = &-\frac{n}{2}\log|\Psi|\\ &-\frac{1}{2}\,\mathrm{tr}\left\{\Psi^{-1}\left[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top - W\sum_{i=1}^{n}\xi_i(x_i-\mu)^\top\right.\right.\\ &\left.\left.-\sum_{i=1}^{n}(x_i-\mu)\xi_i^\top W^\top + W\left(\sum_{i=1}^{n}S_i\right)W^\top\right]\right\}.\end{aligned}$$

En posant les notations :

$$\begin{aligned}
\widetilde{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^{(t)})(x_i - \mu^{(t)})^\top, \\
A &= \sum_{i=1}^{n}(x_i - \mu)\xi_i^\top, \\
B &= \sum_{i=1}^{n} S_i,
\end{aligned}$$

on peut réécrire :

$$Q_\Psi = -\frac{n}{2}\log|\Psi| - \frac{1}{2}\operatorname{tr}\left(\Psi^{-1}\left[n\widetilde{\Sigma} - WA^\top - AW^\top + WBW^\top\right]\right).$$

On maximise Q_Ψ par rapport à Ψ. Le maximum est atteint pour :

$$\Psi^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(x_i - \mu - Wz_i)(x_i - \mu - Wz_i)^\top \mid x_i\right].$$

En remplaçant par l'expression précédente, on obtient :

$$\Psi^{(t+1)} = \widetilde{\Sigma} - \frac{1}{n}W^{(t+1)}A^\top - \frac{1}{n}AW^{(t+1)\top} + \frac{1}{n}W^{(t+1)}BW^{(t+1)\top}.$$

On peut encore développer complètement, en explicitant ξ_i et S_i comme :

$$\begin{aligned}
\xi_i &= M^{(t)}W^{(t)\top}\Psi^{(t)-1}(x_i - \mu^{(t)}), \\
S_i &= M^{(t)} + \xi_i\xi_i^\top,
\end{aligned}$$

et poser :

$$M^{(t)} = \left(W^{(t)\top}\Psi^{(t)-1}W^{(t)} + I\right)^{-1}.$$

Alors on retrouve la forme fermée donnée par Bach and Jordan [2005] :

$$\Psi^{(t+1)} = \widetilde{\Sigma} - \widetilde{\Sigma}\Psi^{(t)-1}W^{(t)}M^{(t)}W^{(t)\top} - W^{(t)}M^{(t)}W^{(t)\top}\Psi^{(t)-1}\widetilde{\Sigma} + W^{(t)}M^{(t)}W^{(t)\top}\Psi^{(t)-1}\widetilde{\Sigma}\Psi^{(t)-1}W^{(t)}M^{(t)}W^{(t)\top}$$

C L'imputation multiple

Imputation multiple L'imputation simple remplace chaque valeur manquante par une seule prédiction, ce qui ignore l'incertitude inhérente à cette prédiction. Cela peut conduire à une sous-estimation de la variance totale, donc à des intervalles de confiance trop étroits et à des tests statistiques trop optimistes.

L'imputation multiple est une méthode plus robuste qui consiste à effectuer plusieurs imputations indépendantes des valeurs manquantes, en générant m jeux de données complétés, chacun reflétant la variabilité due à l'incertitude de l'imputation. Chaque jeu de données est ensuite analysé séparément, et les résultats sont combinés pour produire des inférences globales valides. L'imputation multiple est une méthode d'estimation qui permet de refléter la variabilité des estimateurs.

Principe :

1. Générer m jeux de données complets, en imputant les valeurs manquantes selon un modèle probabiliste (souvent une variante de regression).
2. Analyser chaque jeu de données imputé comme si les données étaient complètes.

Avantages :

- Prend en compte l'incertitude liée aux valeurs manquantes.
- Produit des estimations moins biaisées et des intervalles de confiance plus réalistes.

Limites :

- Plus coûteuse en calculs, car nécessite plusieurs analyses.
- Suppose toujours que le mécanisme de donnée est ignorable (MAR ou MCAR).

Pour combiner les résultats des m estimations, on utilise les règles de Rubin [1987] qui sont en annexe D.

Les principales méthodes d'imputation multiple

MICE (Multivariate Imputation by Chained Equations) La méthode MICE, proposée par Buuren and Groothuis-Oudshoorn [2011], procède par itérations successives d'imputation univariée conditionnelle :

1. On initialise les valeurs manquantes, par exemple par la moyenne ou une régression simple.
2. Pour chaque variable à imputer Y_j, on ajuste un modèle de régression (linéaire, logistique, etc.) de Y_j sur les autres variables Y_{-j}.
3. On remplace les NA de Y_j par des tirages issus de la distribution prédictive du modèle ajusté, afin de refléter l'incertitude.
4. On répète cette chaîne d'imputations jusqu'à convergence.

Après convergence, on obtient un jeu de données complet ; en répétant le procédé m fois, on construit m imputations multiples. Cette méthode est implémentée dans le package `mice` de R via la fonction `mice()` paru dans Buuren and Groothuis-Oudshoorn [2011].

Imputation Multiple par Bootstrap L'imputation par bootstrap vise à intégrer à la fois l'incertitude du modèle d'imputation et la variabilité d'échantillonnage :

1. On tire m échantillons bootstrap des données observées.
2. Sur chaque échantillon, on ajuste un modèle d'imputation (par exemple une régression ou MICE).
3. On impute les valeurs manquantes dans l'échantillon bootstrap à l'aide du modèle ajusté.
4. On rassemble les m jeux de données imputés pour appliquer ensuite les règles de Rubin [1987].

Cette approche reflète non seulement l'incertitude des valeurs manquantes, mais aussi celle liée à l'échantillonnage, et peut être implémentée en combinant des fonctions de bootstrap (par exemple `boot` dans R) avec un algorithme d'imputation de Efron [1979].

D Les règles de Rubin

Règles de Rubin pour l'imputation multiple Les règles de Rubin [1987] fournissent un cadre pour combiner les m estimations issues de ces jeux complets, en tenant compte de :

- l'**incertitude intra-imputation**, liée à la variabilité de l'estimateur sur chacun des jeux complétés ;
- l'**incertitude inter-imputation**, liée aux différences entre les différentes imputations.

Soient, pour $j = 1, \ldots, m$ imputations indépendantes :

- $\hat{\theta}_j$: estimateur du paramètre d'intérêt sur le jeu complet j,
- U_j : variance (intra-imputation) associée à $\hat{\theta}_j$.

1. Moyenne des estimations $\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$

2. Moyenne des variances intra-imputation $\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$

3. Variance inter-imputation $B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \bar{\theta})^2$

4. Variance totale $T = \bar{U} + \left(1 + \frac{1}{m}\right) B$

5. Degrés de liberté effectifs Rubin [1987] propose d'estimer les degrés de liberté par $\nu = (m-1)\left(1 + \frac{\bar{U}}{(1+1/m)\,B}\right)^2$ et l'intervalle de confiance à $(1-\alpha)$ $\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T}$.

E Les étapes de la CCA

1. **Normalisation**. On centre chaque variable de chaque matrice X1, X2.
2. **Matrice de corrélation canonique**. Construire
$$K \;=\; \tilde{\Sigma}_{11}^{-1/2}\,\tilde{\Sigma}_{12}\,\tilde{\Sigma}_{22}^{-1/2}.$$
3. **Décomposition en valeurs singulières (SVD)**. Écrire
$$K = U\,D\,V^{\top},$$
où $D = \mathrm{diag}(\rho_1, \ldots, \rho_d)$ contient les d plus grandes corrélations canoniques.
4. **Directions canoniques**. Les vecteurs de projection s'obtiennent via
$$A = \tilde{\Sigma}_{11}^{-1/2}\,U, \quad B = \tilde{\Sigma}_{22}^{-1/2}\,V;$$
leurs colonnes (a_k, b_k) sont les directions canoniques optimales.
5. **Projection et interprétation**. Les nouvelles variables $u_k = X_1 a_k,\ v_k = X_2 b_k$ sont non corrélées entre elles ($\mathrm{corr}(u_i, u_j) = 0$ si $i \neq j$), et la corrélation $\mathrm{corr}(u_k, v_k) = \rho_k$ est maximale.

Choix du nombre de composantes d

La sélection de d (le nombre de directions retenues) peut se faire :

- en fixant un seuil minimal sur les valeurs ρ_k (ex. $\rho_k > 0.5$),
- en testant statistiquement (test de Wilks),
- ou via des critères d'information (BIC, AIC) et validation croisée.

F Les étapes de la PCCA

1. **Estimation des matrices de covariance** :
$$\tilde{\Sigma}_{11} = \frac{1}{n-1} X_1^{\top} X_1, \quad \tilde{\Sigma}_{22} = \frac{1}{n-1} X_2^{\top} X_2, \quad \tilde{\Sigma}_{12} = \frac{1}{n-1} X_1^{\top} X_2.$$
2. **Matrice de corrélation croisée normalisée** :
$$M = \tilde{\Sigma}_{11}^{-1/2}\,\tilde{\Sigma}_{12}\,\tilde{\Sigma}_{22}^{-1/2}.$$
3. **Décomposition en valeurs singulières (SVD)** :
$$M = U_1 D V_2^{\top},$$
où $D = \mathrm{diag}(\rho_1, \ldots, \rho_d)$ contient les d plus grandes corrélations canoniques, et $U_1 \in \mathbb{R}^{m_1 \times d}$, $V_2 \in \mathbb{R}^{m_2 \times d}$ sont les vecteurs propres associés.

4. **Choix des matrices arbitraires** :

$$M_1 = M_2 = D^{1/2},$$

vérifiant $M_1 M_2 = D = P_d$. Ce choix de matrice influe sur la non unicité de la solution, on revient sur ça juste à la fin des étapes.

5. **Estimations des directions canoniques** :

$$\hat{W}_1 = \Sigma_{11}^{1/2} U_1 D^{1/2} \in \mathbb{R}^{m_1 \times d}, \quad \hat{W}_2 = \Sigma_{22}^{1/2} V_2 D^{1/2} \in \mathbb{R}^{m_2 \times d}.$$

6. **Estimation du bruit** :

$$\hat{\Psi}_1 = \Sigma_{11} - \hat{W}_1 \hat{W}_1^{\top}, \quad \hat{\Psi}_2 = \Sigma_{22} - \hat{W}_2 \hat{W}_2^{\top}.$$

Non unicité de la solution Les matrices $\hat{W}_1$ et $\hat{W}_2$ s'expriment sous la forme :

$$\hat{W}_1 = \tilde{\Sigma}_{11} U_{1d} M_1, \quad \hat{W}_2 = \tilde{\Sigma}_{22} U_{2d} M_2, *$$

où $M_1, M_2 \in \mathbb{R}^{d \times d}$ satisfont la contrainte

$$M_1 M_2^{\top} = P_d,$$

avec $P_d = \mathrm{diag}(\rho_1, \ldots, \rho_d)$ la matrice diagonale des corrélations canoniques soit $P_d = \mathrm{D}$ dans la décomposition en valeur singulière

Cette paramétrisation implique une **infinité de solutions** car pour tout couple (M_1, M_2) vérifiant la contrainte, on obtient des estimations $\hat{W}_1, \hat{W}_2$ différentes, mais qui génèrent le même modèle probabiliste.

Pour obtenir une solution unique on impose

$$M_1 = M_2 = M = P_d^{1/2} R,$$

où $R \in \mathbb{R}^{d \times d}$ est une matrice de rotation orthogonale, i.e. $R R^{\top} = I_d$. Ainsi, les matrices $\hat{W}_1$ et $\hat{W}_2$ s'écrivent

$$\hat{W}_1 = \tilde{\Sigma}_{11} U_{1d} P_d^{1/2} R, \quad \hat{W}_2 = \tilde{\Sigma}_{22} U_{2d} P_d^{1/2} R.$$

Cette restriction élimine l'ambiguïté liée à la multiplicité des couples (M_1, M_2).

En résumé, la **non-unicité** des solutions vient de la possibilité de choisir différentes paires (M_1, M_2) satisfaisant la contrainte $M_1 M_2^{\top} = P_d$, et l'unicité est restaurée en restreignant la solution à $M_1 = M_2 = P_d^{1/2} R$ où on a une rotation R.

G Le package `missMDA`

Dans le cadre des méthodes low rank dans le cas de donnée quantitatives, l'ACP est très populaire. Le package MissMDA de R propose une implémentation d'un algorithme dit ACP itérative régularisé qui permets d'imputer les données manquantes à l'aide de la décomposition en valeur propre.

L'ACP itérative régularisé est une forme avancé de l'ACP itérative, également connue sous le nom d'algorithme EM-PCA (Expectation-Maximization PCA), car elle correspond à un algorithme EM appliqué à un modèle d'effets fixes en PCA (Caussinus, 1986), dans lequel les données sont supposées générées par une structure fixe de faible rang sur S dimensions perturbée par du bruit : $x_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij}, \quad \text{avec } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Ces algorithmes convergent généralement vers un maximum local.

L'ACP itérative fournit de bonnes estimations des paramètres de l'ACP lorsque :

- les variables sont fortement corrélées,
- et le taux de données manquantes est faible.

Cependant, elle souffre rapidement de **sur-apprentissage** en présence de bruit ou de nombreuses valeurs manquantes : les données observées sont bien ajustées, mais la capacité de prédiction est faible.

Pour pallier ce problème, des variantes régularisées ont été proposées, notamment par Josse et al. [2009], avec l'algorithme suivant :

1. **Initialisation** ($\ell = 0$) :
   - Imputer les valeurs manquantes par des valeurs initiales (comme la moyenne des variables),
   - Noter la matrice imputée $X^{(0)}$,
   - Calculer $M^{(0)}$, matrice contenant les moyennes des variables répétées en lignes.
2. **Étape itérative** ($\ell \geq 1$) :
   (a) Effectuer une ACP (SVD de $X^{(\ell-1)} - M^{(\ell-1)}$) pour estimer $U^{(\ell)}$, $V^{(\ell)}$ et $\sqrt{\Lambda^{(\ell)}}$.
   (b) Garder les S premières dimensions et construire la matrice ajustée : $\hat{x}_{ij}^{(\ell)} = \sum_{s=1}^{S} \left(\sqrt{\lambda_s^{(\ell)}} - \frac{\hat{\sigma}^{2(\ell)}}{\sqrt{\lambda_s^{(\ell)}}} \right) u_{is}^{(\ell)} v_{js}^{(\ell)}$, avec : $\hat{\sigma}^{2(\ell)} = \frac{\left\| X^{(\ell-1)} - U^{(\ell)} \sqrt{\Lambda^{(\ell)}} (V^{(\ell)})^\top \right\|^2}{np - nS - pS + S^2}$.
   (c) Mettre à jour les données imputées : $X^{(\ell)} = W * X + (1 - W) * \hat{X}^{(\ell)}$, où W est une matrice binaire indiquant les valeurs observées (1) ou manquantes (0).
   (d) Mettre à jour $M^{(\ell)}$ à partir de $X^{(\ell)}$.
3. **Critère d'arrêt** : on arrête les itérations lorsque le changement entre deux matrices imputées successives est inférieur à un seuil prédéfini : $\sum_{i,j} \left(\hat{x}_{ij}^{(\ell)} - \hat{x}_{ij}^{(\ell-1)} \right)^2 \leq \varepsilon, \quad \text{avec } \varepsilon\text{petit}$.

   Pour me familiariser avec L'EM PCA codé dans le package MissMDA, je l'ai recodé à la main

H Résultats EM-GPCCA RMSE 15%

Repetition	PropMissing	Schema	d	RMSE	ARI
2	0.15	MAR1	12	0.6055	0.9935
4	0.15	MARR	12	0.6068	0.9987
8	0.15	MARR	12	0.6079	0.9974
9	0.15	MAR1	12	0.6083	0.9960
7	0.15	MAR1	12	0.6100	1.0000
9	0.15	Clust	12	0.6101	1.0000
4	0.15	Clust	12	0.6113	0.9987
6	0.15	MARR	12	0.6115	0.9987
7	0.15	MARR	12	0.6118	0.9974
1	0.15	MAR1	12	0.6121	1.0000
6	0.15	MAR1	12	0.6126	0.9974
2	0.15	MARR	12	0.6131	0.9973
4	0.15	MAR1	12	0.6136	0.9973
1	0.15	MARR	12	0.6154	0.9987
5	0.15	MARR	12	0.6165	0.9974
3	0.15	MARR	12	0.6192	0.9973
9	0.15	MARR	12	0.6224	0.9961
8	0.15	MAR1	12	0.6239	0.9933
1	0.15	Clust	12	0.6242	0.9987
5	0.15	Clust	12	0.6248	0.9961

Table 3: Top 20 des meilleures performances (plus petits RMSE) pour 15% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
8	0.15	Self	1	1.2597	0.3261
7	0.15	Self	1	1.2507	0.3243
10	0.15	Clust	2	1.2464	0.7203
2	0.15	Self	1	1.2384	0.3224
9	0.15	Self	1	1.2115	0.5272
4	0.15	Self	1	1.1626	0.5381
6	0.15	Self	1	1.1545	0.5532
3	0.15	Self	1	1.1461	0.5341
1	0.15	Self	1	1.1444	0.5402
10	0.15	Self	1	1.1376	0.5510
5	0.15	Self	1	1.1196	0.5476
10	0.15	Clust	1	1.1090	0.6342
10	0.15	MARR	1	1.0918	0.3914
7	0.15	Clust	1	1.0871	0.5744
3	0.15	MAR1	4	1.0791	0.6972
9	0.15	Self	4	1.0653	0.8833
9	0.15	MAR1	1	1.0574	0.3250
8	0.15	MAR1	4	1.0387	0.9637
6	0.15	Clust	1	1.0171	0.6823
1	0.15	Clust	1	1.0110	0.6860

Table 4: Top 20 des moins bonnes performances (plus grands RMSE) pour 15% de NA

I Résultats EM-GPCCA RMSE 30%

Table 5: Top 20 des meilleures performances (plus petits RMSE) pour 30% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
3	0.3	MARR	12	0.5766	1.0000
1	0.3	MARR	12	0.5810	1.0000
4	0.3	MARR	12	0.5826	0.9987
9	0.3	MARR	12	0.5864	0.9987
5	0.3	MARR	12	0.5879	1.0000
6	0.3	MARR	12	0.5939	0.9987
8	0.3	MARR	12	0.5952	0.9987
2	0.3	MARR	12	0.5961	1.0000
7	0.3	MARR	12	0.5985	0.9975
10	0.3	MARR	12	0.5987	0.9974
6	0.3	Clust	12	0.5987	0.9987
4	0.3	Clust	12	0.6031	0.9986
9	0.3	Clust	12	0.6071	0.9972
9	0.3	MAR1	12	0.6071	0.9987
5	0.3	MAR1	12	0.6092	0.9986
3	0.3	MAR1	12	0.6103	0.9986
2	0.3	MAR1	12	0.6117	1.0000
8	0.3	MAR1	12	0.6126	0.9986
1	0.3	MAR1	12	0.6126	0.9987
7	0.3	MAR1	12	0.6129	0.9946

Table 6: Top 20 des moins bonnes performances (plus grands RMSE) pour 30% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
1	0.3	Self	1	1.3235	0.2315
7	0.3	Clust	2	1.2247	0.6865
8	0.3	Self	1	1.1846	0.3336
7	0.3	Self	1	1.1587	0.3303
7	0.3	Clust	1	1.1237	0.4742
10	0.3	MAR1	1	1.1197	0.3392
2	0.3	Clust	1	1.1141	0.4039
6	0.3	Self	1	1.1110	0.3747
2	0.3	Clust	2	1.1096	0.8575
10	0.3	Self	1	1.1058	0.3902
3	0.3	Self	1	1.1045	0.4053
4	0.3	Self	1	1.0979	0.3918
2	0.3	Self	1	1.0947	0.4016
9	0.3	Self	1	1.0906	0.4471
5	0.3	Self	1	1.0893	0.3734
7	0.3	MAR1	2	1.0831	0.6315
10	0.3	Clust	1	1.0781	0.5283
4	0.3	Self	2	1.0633	0.7299
8	0.3	Clust	1	1.0587	0.3713
3	0.3	Clust	1	1.0509	0.3809

J Résultats EM-GPCCA RMSE 45%

Table 7: Top 20 des meilleures performances (plus petits RMSE) pour 45% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
10	0.45	MAR1	12	0.6334	0.9850
5	0.45	MAR1	12	0.6355	0.8081
1	0.45	MAR1	12	0.6356	0.7644
10	0.45	MAR1	9	0.6365	0.9891
4	0.45	Clust	12	0.6366	0.9867
4	0.45	Clust	9	0.6367	0.9986
4	0.45	MAR1	12	0.6374	0.9799
1	0.45	MAR1	9	0.6393	0.9919
9	0.45	MAR1	12	0.6401	0.9574
7	0.45	MAR1	12	0.6403	0.7530
3	0.45	Clust	12	0.6406	0.9973
5	0.45	MAR1	9	0.6415	0.9892
4	0.45	MAR1	9	0.6419	0.7120
2	0.45	MAR1	12	0.6440	0.7512
9	0.45	MAR1	9	0.6443	0.9919
3	0.45	MAR1	12	0.6456	0.9986
6	0.45	Clust	12	0.6459	0.9945
7	0.45	MAR1	9	0.6461	0.9933
6	0.45	MAR1	12	0.6462	0.7488
5	0.45	Clust	12	0.6470	0.9960

Table 8: Top 20 des moins bonnes performances (plus grands RMSE) pour 45% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
6	0.45	Self	2	4.6958	0.4094
7	0.45	Clust	2	2.8227	0.8505
5	0.45	Self	2	2.0192	0.5761
2	0.45	MARR	1	1.7206	0.5926
7	0.45	Clust	1	1.6077	0.5302
4	0.45	MARR	2	1.5829	0.8612
10	0.45	Clust	1	1.4025	0.4947
7	0.45	Clust	4	1.3927	0.9907
7	0.45	Self	1	1.3466	0.1962
6	0.45	MARR	1	1.3278	0.4683
1	0.45	MARR	1	1.2975	0.4768
3	0.45	MARR	2	1.2858	0.7515
2	0.45	Self	1	1.2857	0.2240
2	0.45	MARR	2	1.2857	0.8235
10	0.45	Clust	2	1.2724	0.8109
4	0.45	Self	1	1.2658	0.2836
3	0.45	Self	1	1.2640	0.3107
9	0.45	Self	2	1.2633	0.6272
5	0.45	Self	1	1.2627	0.2759
8	0.45	Self	1	1.2619	0.2959

K Résultats EM-GPCCA ARI 15%

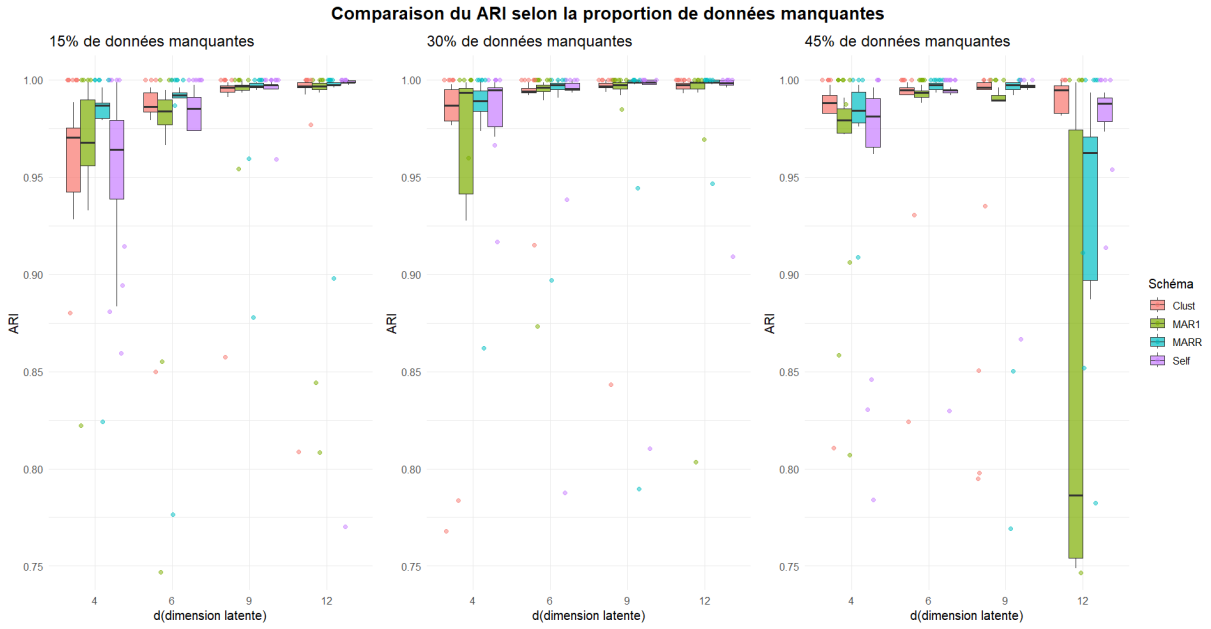


Figure 10: ARI zoomée sur les dimensions 4 à 12.

Table 9: Top 20 des pires performances (plus petites ARI) pour 15% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
2	0.15	Self	1	1.2384	0.3224
7	0.15	Self	1	1.2507	0.3243
9	0.15	MAR1	1	1.0574	0.3250
8	0.15	Self	1	1.2597	0.3261
10	0.15	MARR	1	1.0918	0.3914
9	0.15	MARR	1	0.9841	0.5149
9	0.15	Self	1	1.2115	0.5272
3	0.15	Self	1	1.1461	0.5341
8	0.15	Clust	1	0.9463	0.5372
4	0.15	Self	1	1.1626	0.5381
1	0.15	Self	1	1.1444	0.5402
5	0.15	Self	1	1.1196	0.5476
10	0.15	Self	1	1.1376	0.5510
6	0.15	Self	1	1.1545	0.5532
3	0.15	MARR	1	0.9851	0.5640
8	0.15	MARR	1	0.9769	0.5710
7	0.15	Clust	1	1.0871	0.5744
7	0.15	MARR	1	0.9453	0.5751
5	0.15	MAR1	1	0.9873	0.5761
5	0.15	Clust	1	0.9370	0.5874

Table 10: Top 20 des meilleures performances (plus grandes ARI) pour 15% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
1	0.15	MAR1	12	0.6121	1.0000
3	0.15	Self	12	0.6746	1.0000
5	0.15	Self	12	0.6749	1.0000
7	0.15	MAR1	9	0.7230	1.0000
7	0.15	Self	12	0.6821	1.0000
7	0.15	MAR1	12	0.6100	1.0000
9	0.15	Clust	12	0.6101	1.0000
1	0.15	Self	12	0.6754	0.9987
2	0.15	MARR	9	0.7382	0.9987
6	0.15	Self	12	0.6807	0.9987
9	0.15	Clust	9	0.7476	0.9987
1	0.15	MARR	12	0.6154	0.9987
1	0.15	Clust	12	0.6242	0.9987
3	0.15	Self	9	0.7965	0.9987
3	0.15	MAR1	12	0.6254	0.9987
4	0.15	Self	12	0.6889	0.9987
4	0.15	MARR	12	0.6068	0.9987
4	0.15	Clust	12	0.6113	0.9987
5	0.15	Self	4	0.6963	0.9987
6	0.15	MARR	12	0.6115	0.9987

L Résultats EM-GPCCA ARI 30%

Table 11: Top 20 des pires performances (plus petites ARI) pour 30% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
1	0.3	Self	1	1.3235	0.2315
7	0.3	Self	1	1.1587	0.3303
8	0.3	Self	1	1.1846	0.3336
10	0.3	MAR1	1	1.1197	0.3392
8	0.3	Clust	1	1.0587	0.3713
5	0.3	Self	1	1.0893	0.3734
6	0.3	Self	1	1.1110	0.3747
3	0.3	Clust	1	1.0509	0.3809
10	0.3	Self	1	1.1058	0.3902
4	0.3	Self	1	1.0979	0.3918
2	0.3	Self	1	1.0947	0.4016
2	0.3	Clust	1	1.1141	0.4039
3	0.3	Self	1	1.1045	0.4053
6	0.3	Clust	1	0.9797	0.4261
9	0.3	Self	1	1.0906	0.4471
5	0.3	Clust	1	0.8870	0.4700
7	0.3	Clust	1	1.1237	0.4742
1	0.3	Clust	1	0.9598	0.4784
6	0.3	MARR	1	0.8682	0.4984
10	0.3	Clust	1	1.0781	0.5283

Table 12: Top 20 des meilleures performances (plus grandes ARI) pour 30% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
1	0.3	MARR	9	0.6316	1
1	0.3	Clust	9	0.6534	1
1	0.3	MARR	12	0.5810	1
1	0.3	Clust	12	0.6135	1
2	0.3	MAR1	9	0.6364	1
2	0.3	MARR	9	0.6382	1
2	0.3	MAR1	12	0.6117	1
2	0.3	MARR	12	0.5961	1
3	0.3	MARR	9	0.6286	1
3	0.3	MARR	12	0.5766	1
4	0.3	MARR	4	0.6990	1
4	0.3	Self	9	0.7107	1
5	0.3	Self	6	0.7353	1
5	0.3	MARR	6	0.6765	1
5	0.3	Self	9	0.7028	1
5	0.3	MAR1	9	0.6335	1
5	0.3	MARR	9	0.6255	1
5	0.3	Self	12	0.6742	1
5	0.3	MARR	12	0.5879	1
6	0.3	Self	9	0.7052	1

M Résultats EM-GPCCA ARI 45%

Table 13: Top 20 des pires performances (plus petites ARI) pour 45% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
7	0.45	Self	1	1.3466	0.1962
3	0.45	MARR	1	1.1359	0.2223
2	0.45	Self	1	1.2857	0.2240
10	0.45	Self	1	1.2453	0.2548
5	0.45	Self	1	1.2627	0.2759
8	0.45	Clust	1	1.1613	0.2809
4	0.45	Self	1	1.2658	0.2836
6	0.45	Clust	1	1.0143	0.2839
8	0.45	MARR	1	1.2174	0.2866
2	0.45	Clust	1	1.1914	0.2936
8	0.45	Self	1	1.2619	0.2959
1	0.45	Self	1	1.2608	0.3087
3	0.45	Self	1	1.2640	0.3107
5	0.45	MAR1	1	1.0907	0.3146
4	0.45	MARR	1	1.1130	0.3224
9	0.45	Clust	1	1.0817	0.3224
9	0.45	Self	1	1.2506	0.3317
3	0.45	Clust	1	1.1727	0.3407
5	0.45	Clust	1	1.0456	0.3547
9	0.45	MAR1	1	1.1137	0.3563

Table 14: Top 20 des meilleures performances (plus grandes ARI) pour 45% de NA

Repetition	PropMissing	Schema	d	RMSE	ARI
1	0.45	Clust	9	0.6490	1.0000
1	0.45	Clust	12	0.6486	1.0000
9	0.45	Clust	9	0.6517	1.0000
4	0.45	MARR	9	0.6772	0.9987
1	0.45	MARR	9	0.6679	0.9987
1	0.45	MARR	6	0.7135	0.9987
6	0.45	MARR	6	0.7370	0.9987
6	0.45	MARR	9	0.6772	0.9987
7	0.45	Clust	9	1.2333	0.9987
7	0.45	Clust	12	1.2233	0.9987
1	0.45	Clust	6	0.7330	0.9986
1	0.45	Self	9	0.7657	0.9986
4	0.45	Clust	9	0.6367	0.9986
3	0.45	Self	9	0.7380	0.9986
3	0.45	MAR1	12	0.6456	0.9986
10	0.45	MARR	6	0.7040	0.9986
10	0.45	MARR	9	0.6600	0.9986
7	0.45	MARR	9	0.6534	0.9974
3	0.45	MARR	6	0.7147	0.9974
3	0.45	MARR	9	0.6785	0.9974