Rapport sur l'article «A principal components method to impute missing values for mixed data»

ADJEVI-NEGLOKPE Ambre, FUENTES-VICENTE Laura

March 24, 2024

1 Introduction

Lorsqu'un jeu de données présente des données manquantes, l'application directe de méthodes d'apprentissage statistique traditionnelles devient impossible. L'imputation, qui consiste à ajouter des valeurs artificielles à la place des valeurs manquantes, permet de palier à ce problème en rendant les jeux de données complets et donc exploitables. Il existe diverses méthodes d'imputation, chacune offrant des garanties et des propriétés spécifiques.

Dans le cas où toutes les covariables sont continues, des méthodes d'imputations existent, telles que la méthode des K-plus proches voisins (KNN) [Troyanskaya et al., 2001], l'imputation basée sur le modèle normal multivarié [Schafer, 1997], l'imputation par équation en chaîne [Van Buuren et al., 1999] [Van Buuren, 2007] ou bien l'imputation par ACP [Josse et al., 2009]. Cette dernière méthode, basée sur l'ACP, est particulièrement intéressante car elle permet de prendre en compte à la fois les similarités entre les individus et les relations entre les covariables, ce qui permet de garder une structure des données cohérente lors de l'imputation des données manquantes [Ilin and Raiko, 2010].

Dans le cas où toutes les données sont de nature catégorielle, des méthodes d'imputations couramment utilisées incluent, la méthode KNN (méthode non-paramétrique), le modèle log-linéaire [Schafer, 1997] et le "latent class model" [Vermunt et al., 2008].

En présence de données mixtes, les méthodes d'imputation conventionnelles ne sont pas applicables. Ainsi, il devient impératif de concevoir des techniques spécifiquement adaptées à ce contexte, ce qui contribuera à enrichir la littérature sur ce sujet.

Nous commencerons par aborder les méthodes, en explorant les précurseurs, FAMD et FAMD itératif, suivi par l'implémentation et les résultats incluant l'analyse des jeux de données synthétiques et réels. Enfin, nous conclurons notre étude.

2 Méthodes pour données mixtes

2.1 Méthodes précurseures

Pour les données mixtes, comprenant à la fois des variables continues et catégorielles, une approche courante consiste à encoder les données catégorielles sous forme de variables muettes (dummy variables). Cela permet ensuite d'appliquer des méthodes d'imputation conçues pour les données continues. Mais le problème est que les hypothèses que les méthodes d'imputation imposent sur les variables continues ne peuvent pas être imposées sur ces nouvelles variables nominales. Il faut donc créer de nouvelles méthodes d'imputation afin de prendre en compte les différentes natures des variables.

Une des méthodes proposées consiste à effectuer une combinaison du modèle log-linéaire avec le modèle multivarié normal [Schafer, 1997]. Cependant, cette solution reproduit les désavantages inhérents à chacune des méthodes individuelles. L'imputation par équations en chaines [Van Buuren et al., 1999][Van Buuren, 2007] peut être appliquée au cas mixte car elle repose sur le fait de créer un modèle spécifique pour chacune des variables. Cette approche peut toutefois s'avérer inefficace du fait qu'elle nécessite la création d'un modèle distinct pour chaque variable, ce qui peut être peu optimal en termes de ressources et de temps de calcul.

La méthode qui lors de la rédaction du papier offre les résultats de référence est la méthode de [Stekhoven and Bühlmann, 2012], basée sur les forêts aléatoires. La méthode consiste à remplacer les données manquantes par des valeurs initiales pour chaque variable. Puis, de manière itérative, les données manquantes de la variable présentant le moins de données manquantes sont prédites par des forêts aléatoires et ce, jusqu'à stabilisation des prédictions. C'est la méthode avec les meilleures performances, et les imputations offertes par cette méthode sont de bonne qualité, indépendamment du nombre d'individus, de variables, et du type de relations entre les variables. Néanmoins, un des désavantages de cette méthode est sa sensibilité aux hyper-paramètres.

Les auteurs de l'article au coeur de notre projet présentent une nouvelle méthode d'imputation pour des données mixtes incomplètes. Leur technique repose sur une extension du modèle d'Analyse Factorielle pour les Données Mixtes (FAMD) [Audigier et al., 2016].

2.2 Méthode FAMD dans le cas complet

Dans le cas complet, la méthode FAMD permet la description et la visualisation de données mixtes multidimensionnelles [Lebart et al.,] [Greenacre and Blasius, 2006]. L'objectif est de réduire la dimension des données en trouvant le sous-espace qui maximise la variabilité des points projetés, tout en équilibrant l'influence des variables continues et catégorielles.

Cette méthode repose sur deux fondements essentiels. D'une part, l'étude des composantes principales permet d'établir des prédictions basées sur les similarités entre individus, relations entre variables, liens entre individus et variables, facilitant ainsi la réduction de dimension. D'autre part, l'analyse factorielle assure un équilibre dans

l'influence des variables continues et catégorielles lors de la reconstruction des données par composantes principales.

2.2.1 Algorithme

Soit $X_{I\times K}$ un jeu de données, avec I le nombre d'observations et K le nombre de variables. On note K_1 le nombre de variables continues et K_2 le nombre de variables catégorielles, tel que $K = K_1 + K_2$ [Appendix A].

Step 1: Encoder les variables catégorielles La première étape consiste à encoder les variables catégorielles à l'aide d'une matrice indicatrice sur des variables muettes. Ceci permet de transformer le jeu de données et le rendre apte aux méthodes numériques. On a ainsi un nouveau nombre de variables $J = K_1 + \sum_{k=K_1+1}^{K} q_k$, avec q_k le nombre de catégories de la variable k. Le nouveau jeu de données s'écrit: $X_{I\times J}$.

Step 2: Pondération Cette étape consiste à pondérer les variables. Les variables continues, sont ainsi standarisées en les divisant par leur écart type (s_j), ce qui permet de rendre les distances individus indépendante de l'unité. Pour les variables catégorielles, on divise chaque variable muette par $\sqrt{p_j}$, où p_j désigne la proportion d'observations dans la catégorie $j \in \{K_1 + 1, ..., J\}$. Ceci permet de rendre les individus de catégories rares plus distants que ceux de catégories plus fréquentes entre eux. On obtient ainsi $1 - p_j$ l'inertie de x_j, rendant les catégories rares plus importantes dans la construction.

Step 3: Mise en place de l'ACP Enfin, on met en place une ACP sur XD_Σ^{-1}, avec $D_\Sigma = diag(s^2_{x_1}, ..., s^2_{x_{K_1}}, p_{K_1+1}, ..., p_J)$ une matrice diagonale. On remarque que cette technique revient à faire une décomposition SVD sur $XD_\Sigma^{-1/2} - M$ avec $M_{I\times J}$ la matrice des moyennes de $XD_\Sigma^{-1/2}$ pour chaque I. On préserve enfin les S premières composantes principales. Cette étape nous permet ainsi de réduire la dimension des données tout en mettant en lumière les liens entre individus et variables.

2.3 FAMD itératif pour les données manquantes

La méthode itérative FAMD (iFAMD) vise à imputer des données mixtes incomplètes. Le principe est de remplir les données manquantes initialement, puis d'appliquer des FAMD pour: améliorer les valeurs des données manquantes et mettre à jour D_Σ, M [Appendix A], tant qu'une solution stable n'est pas atteinte.

2.3.1 Algorithme

Algorithm 1 FAMD Itératif

Initialisation $l = 0$
- Substituer les données manquantes par une valeur initiale
Moyenne: variables continues
p_j: variables catégorielles (la somme par variable et par individu doit faire 1)

- Calculer D_Σ^0, M^0, $(D_\Sigma^0)^{-1/2}$

while $\sum_{i,j}(\widehat{x}_{ij}^{l-1} - \widehat{x}_{ij}^{l})^2 > \epsilon$ **do**

- Mettre en place la FAMD:
Calculer SVD sur $(X^{l-1}(D_\Sigma^{l-1})^{-1/2} - M^{l-1})$

Modification du terme diagonal de la SVD pour effectuer la régularisation:
$\widehat{\Lambda'}_s^l = (\frac{\widehat{\Lambda}_s^l - \sigma^2}{\sqrt{\widehat{\lambda}_s^l}})$

- Garder les S premières dimensions et reconstruire la matrice:
$\widehat{X}_{I\times J}^l = (\widehat{U}_{I\times S}^l(\widehat{\Lambda'}_{S\times S}^{l})^{1/2}(\widehat{V}_{J\times S}^l)^T + M_{I\times J}^{l-1})((D_\Sigma^{l-1})^{1/2})$

- Mettre-à-jour D_Σ^l, M^l et $X^l = W * X + (1 - W) * \widehat{X}^l$
end while

2.3.2 Propriétés

Les auteurs soulignent plusieurs aspects essentiels de l'algorithme à prendre en compte. En premier lieu, son efficacité dépend souvent de sa capacité à gérer les relations entre les variables continues et catégoriques, améliorant la qualité de l'imputation des données manquantes. De plus, l'impact des interactions entre les variables est crucial, surtout dans des méthodes telles que la FAMD basée sur l'ACP, où des relations linéaires fortes garantissent des imputations précises. Cependant, des relations complexes entre plusieurs variables peuvent devenir un inconvénient, rendant l'imputation plus complexe. Une solution envisageable serait d'introduire une variable modélisant ces interactions.

Comme exposé précédemment, l'algorithme accorde une importance significative aux catégories rares, se traduisant par des prédictions précises pour ces catégories. Enfin, en ce qui concerne le choix du nombre de dimensions, il est essentiel de trouver un équilibre adéquat. En effet, un faible nombre de dimensions peut entraîner une perte d'informations importante, tandis qu'un nombre excessif peut considérer du bruit comme un signal, provoquant des instabilités dans les prédictions. C'est dans ce contexte que les auteurs recommandent l'utilisation de la validation croisée qui se relève adaptée aux données incomplètes. Cette approche permet ainsi de choisir judicieusement le nombre de dimensions tout en évitant des potentiels biais.

3 Implémentations

Afin d'évaluer cette nouvelle méthode, elle a été appliquée à différents datasets où des données manquantes ont été artificiellement insérées. Chacun des datasets présente des particularités propres, afin d'illustrer les différentes propriétés de la méthode iFAMD. Vous pourrez retrouver nos implémentations et résultats dans le repositoire Github ci-joint.

3.1 Jeux de données synthétiques

Dans le papier, les propriétés de Iterative FAMD (iFAMD) sont illustrées grâce à l'implémentation de l'algorithme sur des données générées. La méthode de génération de ces datasets est la suivante :

- Choisir une dimension S et créer S variables indépendantes qui suivent une loi normale centrée réduite.
- Répliquer K^s fois chacune des variables s ($s \in 1, ..., S$) afin de créer S groupes orthogonaux de variables corrélées.
- Ajouter de bruit Gaussien.
- Créer des variables catégorielles.

3.1.1 Influence du bruit

La première propriété sur laquelle nous avons travaillé, est l'influence du bruit SNR sur les performances de l'algorithme. Dans le papier, cet aspect est mentionné mais n'est pas illustré. Nous avons donc réalisé 200 simulations de jeux de données avec des paramètres SNR de 1 et 3, afin de représenter différentes conditions de bruit. Les résultats sont présentés dans [Figure 1].

Premièrement, nous observons que l'erreur d'imputation augmente proportionnellement à la proportion de données manquantes. Deuxièmement, une faible valeur de SNR, associée à un bruit plus prononcé, conduit à une dégradation des performances de l'algorithme, comme illustré dans nos résultats. Ces résultats sont consistants avec ceux mentionnés dans l'article.

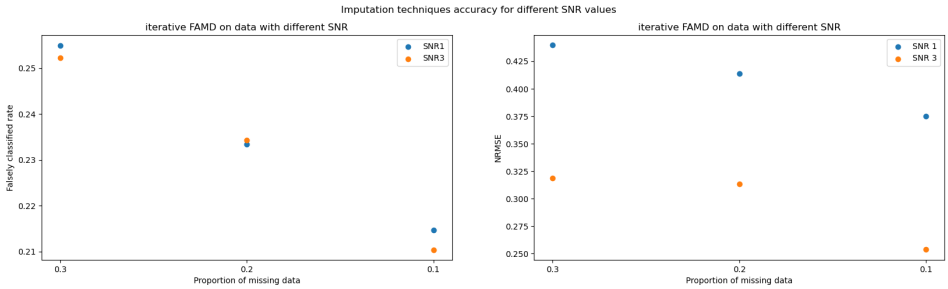

Figure 1: Performance de iFAMD sur des données avec des niveaux de bruit différents

3.1.2 Relations linéaires et non linéaires

La deuxième propriété illustrée est l'influence de la nature linéaire des relations entre les variables sur les performances de l'algorithme. Nous avons créé 200 simulations en suivant la méthode et les paramètres fournis dans le papier. Pour chaque jeu de données linéairement liées, on dérive un jeu de données qui a des variables non-linéairement liées en modifiant deux variables: une variable est passée au carrée et on prend le cosinus de l'autre variable.

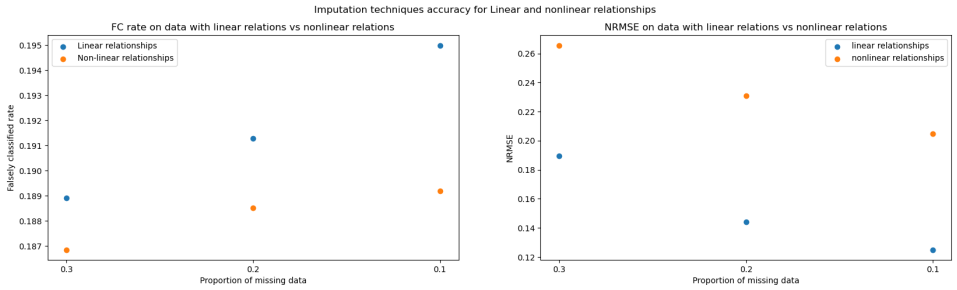

Figure 2: Performance de iFAMD sur des variables liées linéairement vs non-linéairement

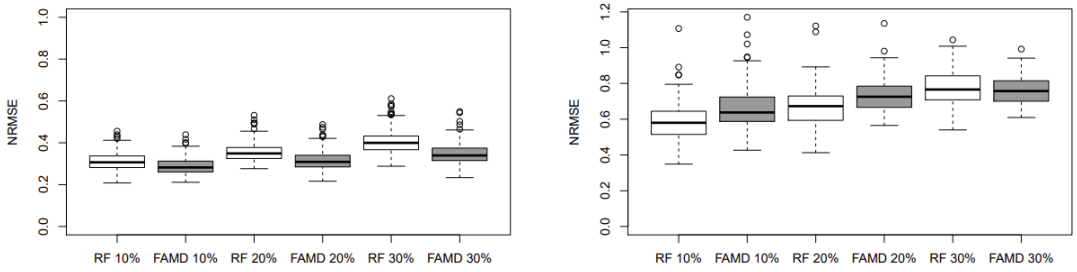

Figure 3: Distribution de la NRMSE quand les relation inter-variables sont linéaires (gauche) et non-linéaire (droite) pour différentes proportions de valeurs manquantes. Les boites blanches correspondent aux performances des Forêts aléatoires, les boîtes grises à celles de iFAMD. (Résultats du papier [Audigier et al., 2016])

Dans [Figure 2, droite], il est observé que les données comportant des variables linéairement liées sont mieux imputées par iFAMD. De même, selon [Figure 3], iFAMD affiche de meilleures performances que les forêts aléatoires sur des données ayant des relations linéaires. Cependant, iFAMD présente des performances inférieures sur des données non linéaires comme revèlent les résultats de [Figure 2, droite]. En revanche, sur [Figure 2, gauche] il est noté que l'imputation des variables catégorielles est de meilleure qualité lorsque les données présentent des relations non linéaires.

Lorsque les variables continues sont liées de manière non linéaire, l'imputation des valeurs manquantes semble initialement plus complexe. Cependant, la FAMD introduit des combinaisons linéaires des indicatrices des variables catégorielles, offrant ainsi une approximation des relations non linéaires. Cela explique les performances observées dans [Figure 2 gauche].

3.1.3 Catégories rares

Des données sont créées avec deux variables catégorielles qui admettent une catégorie rare (associée à la fréquence f). Nous présentons ici [Figure 4] les résultats du papier pour différentes valeurs de f et différents nombres d'individus. L'algorithme iFAMD est plus efficace que la méthode des forêts aléatoires pour imputer les valeurs des individus rares, et l'imputation est meilleure lorsqu'il y a un plus grand échantillon d'individus.

Number of individuals	f	FAMD	Random forests
100	10%	0.060	0.096
100	4%	0.082	0.173
1000	10%	0.042	0.041
1000	4%	0.060	0.071
1000	1%	0.074	0.167
1000	0.4%	0.107	0.241

Figure 4: Performance de iFAMD sur 1000 simulations de données avec catégories rares (Résultats du papier [Audigier et al., 2016])

3.1.4 Choix du nombre de dimensions

La dernière propriété illustrée concerne le choix du nombre de composantes principales pour la reconstruction des données. Nous avons réalisé 200 simulations de jeux de données avec des paramètres SNR de 1 et 3. Le nombre de dimensions utilisé pour la création des données est de S = 2.

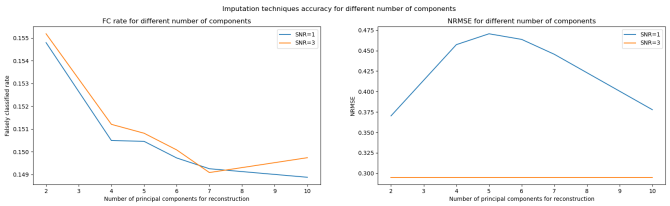

Figure 5: Performance de iFAMD en fonction de la dimension choisie pour différents niveaux de bruit (pour 10% de données manquantes)

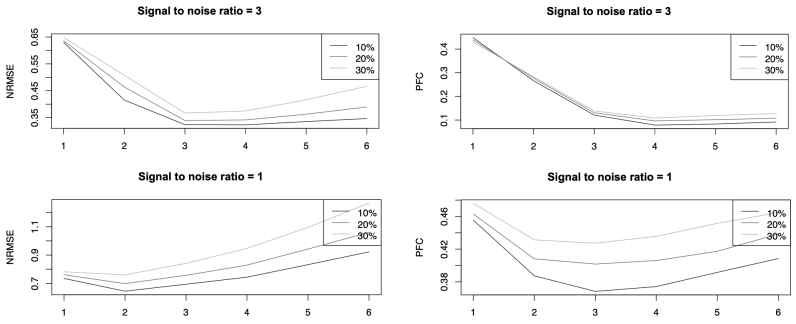

Figure 6: Erreur moyenne d'imputation sur 200 simulations en fonction du nombre de dimensions utilisées dans l'algorithme et pour 3 niveaux de valeurs manquantes (10%,20%,30%): erreur pour les variables continues à gauche et pour les variables catégorielles à droite. Le rapport signal sur bruit est égal à 3 pour les simulations représentées en haut, et à 1 pour les simulations représentées en bas (Résultats du papier [Audigier et al., 2016]).

Lorsque le rapport signal sur bruit (SNR) est élevé, l'erreur d'imputation décroît jusqu'à atteindre une valeur optimale, comme illustré dans [Figure 5 et 6].

En revanche, lorsque le SNR est bas (indiquant un niveau élevé de bruit), il est préférable de choisir un nombre plus restreint de dimensions [Figure 6]. Notre étude a élargi le choix du nombre de dimensions, démontrant qu'au-delà d'un certain seuil, les performances s'améliorent [Figure 5 droite]. Au niveau des variables catégorielles, nos résultats diffèrent de ceux présentés dans l'article, car nous observons des résultats similaires pour des valeurs de SNR différentes [Figure 5 gauche].

3.2 Jeu de données réel

Afin d'évaluer la souplesse et les performances de l'algorithme, nous l'avons mis en œuvre sur l'un des ensembles de données suggérés dans l'article : l'ensemble de données du «Groupe d'Étude sur le Cancer du Sein Allemand» (GBSG2)[Sauerbrei and Royston, 1999] [Appendix A]. Ce jeu de données complet, examine l'effet d'un traitement hormonal sur la période de survie sans récurrence du cancer du sein. Il comprend un total de 686 patientes et 11 variables, dont K_1=7 sont continues et K_2=4 sont catégorielles.

Pour se ramener au cœur de l'étude, nous avons caché artificiellement une partie des données. Notre démarche consistait à tirer sur une loi uniforme de support $[0, 1]$ et cacher les données supérieures à un seuil $p \in \{0.7, 0.8, 0.9\}$.

Dans un premier temps, nous avons testé les capacités de l'algorithme pour différentes proportions de données manquantes avec un nombre de composantes principales fixé égal à 5.

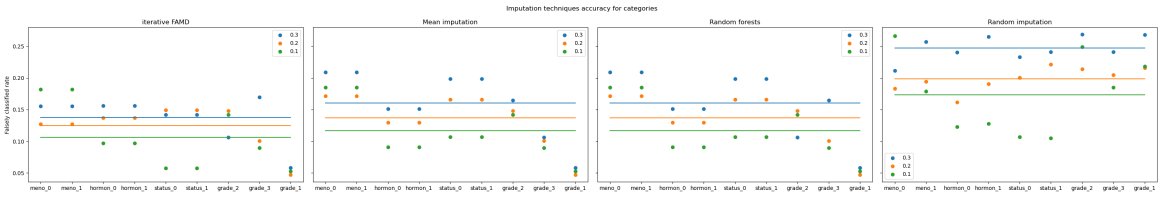

Figure 7: PFC en fonction de la probabilité de missigness

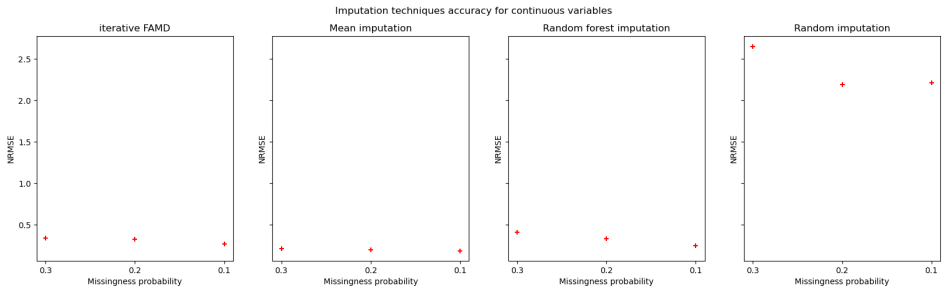

Figure 8: NRMSE en fonction de la probabilité de missigness

Il est observé dans [Figures 7 et 8] que lorsque la proportion de données manquantes augmente, on assiste a une augmentation de l'erreur d'imputation. En ce qui concerne les variables catégorielles, l'algorithme iFAMD affiche de meilleures performances que

les autres méthodes d'imputation (imputation par moyenne, imputation aléatoire, imputation par forêt aléatoire), quelle que soit la proportion de données manquantes. Pour les variables continues, les performances de l'algorithme iFAMD sont significativement meilleures que celles de l'imputation aléatoire, mais moins bonnes que celles de l'imputation par moyenne. Elles sont également très similaires à celles obtenues avec les forêts aléatoires.

Nous avons testé dans un second temps, les performances de l'algorithme en fonction du nombre de composantes principales choisi pour une proportion de données manquantes fixée à 20%.

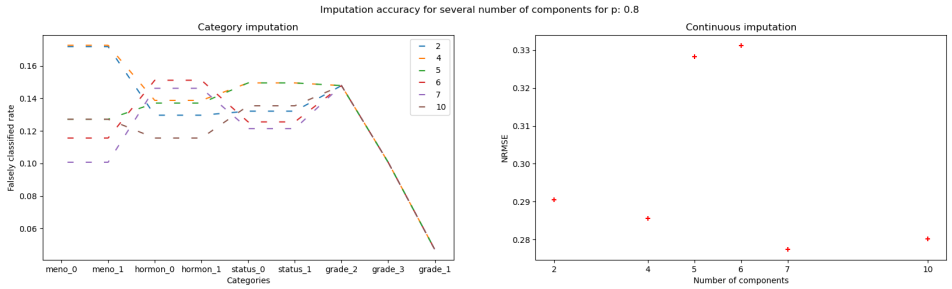

Figure 9: Étude des performances en fonction du nombre de composantes principales

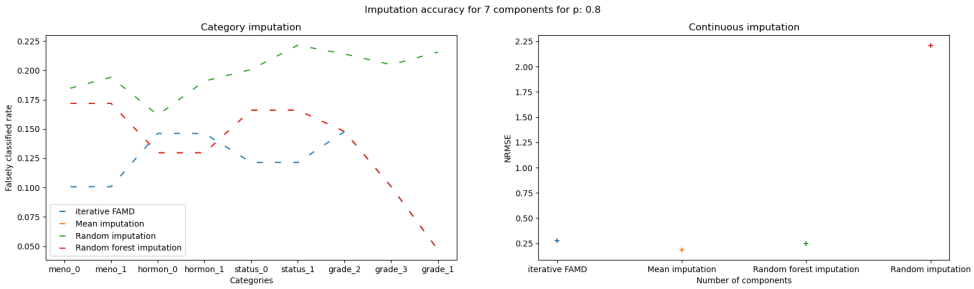

Figure 10: Comparaison avec d'autres méthodes d'imputation

Au niveau des variables catégorielles [Figure 9, gauche], on observe que le nombre de composantes principales entraîne des performances fluctuantes en fonction de la catégorie. Par exemple, pour 10 composantes principales, les meilleurs résultats sont obtenus pour la variable «hormon», tandis que les résultats pour «meno» et «status» sont loin d'être les meilleurs. Dans ce contexte, il apparaît que le nombre optimal de composantes principales pour obtenir les meilleurs résultats généralisés se situe autour de 7.

Concernant les variables continues [Figure 9, droite], on peut remarquer que des performances surprénamment bonnes sont obtenues pour des nombres extrêmes de composantes principales (très petits ou grands). Cependant, cet effet est nuancé par l'observation faite sur les variables catégorielles. Ainsi, la meilleure performance est obtenue avec 7 composantes principales.

En comparant les performances de l'algorithme iFAMD pour 7 composantes principales avec celles des autres algorithmes [Figure 10], les mêmes conclusions sont tirées qu'auparavant. En effet, au niveau des catégories, cette technique demeure généralement bien meilleure que les autres méthodes, tandis qu'au niveau des variables continues, elle se situe assez proche des autres techniques.

4 Conclusion

Dans le cas complet, la méthode FAMD permet de décrire des données mixtes en détectant des similarités entres les individus et en prenant en compte les relations entre les variables continues et catégorielles. La méthode d'imputation pour données mixtes basée sur FAMD permet d'imputer les données manquantes en prenant compte à la fois des similarités inter-individus et des relations entre les variables continues et catégorielles. Les prédictions offertes par cette méthode sont particulièrement efficaces lorsque les données sont liées linéairement. Les données catégorielles manquantes sont aussi très bien prédites, en particulier lorsqu'elles font partie d'une catégorie rare. Cette méthode concurrence celle basée sur les forêts aléatoires, tant sur le plan de la qualité des prédictions que sur celui du temps computationnel. Néanmoins, les performances de Iterative FAMD se dégradent avec la proportion de données manquantes, le manque de relations entre les variables, et sont très dépendantes de la nature des relations entre les variables.

Le nombre de dimensions utilisées pour la reconstruction des données est un hyperparamètre qui peut être choisit par validation croisée. Sa détermination influe énormément au niveau des performances. Comme nous l'avons vu, le choix d'un nombre trop faible de dimension entraîne une grande perte d'information tandis qu'un nombre trop élevé introduit du bruit.

Une manière d'améliorer cette méthode serait de créer plusieurs imputations des données manquantes et de combiner les résultats pour obtenir une estimation dont la variabilité est moindre. C'est l'une des pistes disponibles afin d'élargir la littérature au sujet de l'imputation de données mixtes.

A Appendix

Présentation du jeu de données réel (GBSG2)
Ensemble de données examinant l'impact d'un traitement hormonal sur le délai de réapparition du cancer du sein, collecté par le «German Breast Cancer Study Group» [Sauerbrei and Royston, 1999]. L'échantillon comprend 686 femmes et comprend 7 variables continues ainsi que 4 variables catégorielles.

- Variables continues:
  - pid: ID de la patiente
  - age: age du patient (en années)
  - size: taille de la tumeur (en millimètres)
  - pgr: récepteur de progestérone (en fmol)
  - er: recepteur d'estrogènes (en fmol)
  - nodes: nombre positif de nodes
  - rfstime: temps de récurrence (en jours)
- Variables catégorielles
  - meno: état de menopause (pré-menopause, post-menopause)
  - grade: grade de la tumeur (I,II,III)
  - hormon: traitement hormonal (oui, non)
  - status: état (présence, absence)

Nomenclature

D_Σ	matrice diagonale $diag(s^2_{x_1}, ..., s^2_{x_{K1}}, p_{K_1+1}, ..., p_J)$
I	nombre d'individus
J	nombre total de variables après encodage $J = K_1 + \sum_{k=K_1+1}^{K} q_k$
K	nombre total de variables avant encoder les catégories
K_1	nombre de variables continues
K_2	nombre de variables catégorielles
$M_{I,J}$	matrice contenant dans chaque ligne, le vecteur des moyennes de $XD_\Sigma^{-1/2}$
p_j	proportion d'individus dans la catégorie $j \in \{K_1{+}1, ..., J\}$ avec $1{-}p_j$ =Inertie(x_j)
S	Dimensions de variabilité préservés
s_j	écart type des variables continues
$X_{I\times H}$	jeu de données avec $H = K$ ou J

References

[Audigier et al., 2016] Audigier, V., Husson, F., and Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10:5–26.

[Greenacre and Blasius, 2006] Greenacre, M. and Blasius, J. (2006). *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC.

[Ilin and Raiko, 2010] Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000.

[Josse et al., 2009] Josse, J., Husson, F., et al. (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la société française de statistique*, 150(2):28–51.

[Lebart et al.,] Lebart, L., Morineau, A., and KM. Warwick (1984), multivariate descriptive statistical analysis.

[Sauerbrei and Royston, 1999] Sauerbrei, W. and Royston, P. (1999). Database gbsg2.

[Schafer, 1997] Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

[Stekhoven and Bühlmann, 2012] Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

[Van Buuren, 2007] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

[Van Buuren et al., 1999] Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694.

[Vermunt et al., 2008] Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1):369–397.