A principal components method to impute missing values for mixed data

Ambre Adjevi-Neglokpe, Laura Fuentes Vicente

27 mars 2024

Guidelines in Machine Learning, M2 Maths & IA

Sommaire

Introduction

Motivations

variable 1	variable 2	...	variable n
1.4	NAN	...	Yes
...	...	...	...
0.02	III	...	NAN

- Jeu de données complet $\longrightarrow$ Situation rare
- $\times$ Méthodes d'apprentissage statistique traditionnelle non adaptés
- ✓ **Solutions:** Imputation, adaptation des modèles statistiques, ...

Motivations

Imputation: compléter jeu de données avec valeurs manquantes

Méthodes d'imputation sur données non-mixtes :

- Variables continues :
  - K-plus proches voisins (KNN)
  - Modèle normal multivarié
  - Équation en chaîne
  - Imputation par ACP
- Variables catégorielles :
  - KNN
  - Modèle log-linéaire
  - "Latent Class Model"

Méthodes d'imputation de données mixtes

Méthodes précurseures

- Combinaison Modèle log-linéaire et modèle multivarié [1]
  - × Reproduit désavantages des deux méthodes

- Imputation par équations en chaine [2][3]
  - × Un modèle par variable: coûteux en ressources

Méthodes précurseures

- Combinaison Modèle log-linéaire et modèle multivarié [1]
  - × Reproduit désavantages des deux méthodes
- Imputation par équations en chaine [2][3]
  - × Un modèle par variable: coûteux en ressources
- Imputation par forêts aléatoires [4]
  - Étape 1: Remplacer données manquantes par valeurs initiales
  - Étape 2: Imputer itérativement données manquantes de variable avec moins de valeurs manquantes
  - ✓ Bons résultats: indépendamment du nombre d'individus et type de relations entre variables
  - × Sensible aux hyperparamètres

FAMD dans le cas complet

Basée sur l'ACP et analyse factorielle [5] [6]

Objectifs:

- Réduire dimension des données: maximiser variabilité des points projetés
- Équilibre influence variables continues et catégorielles

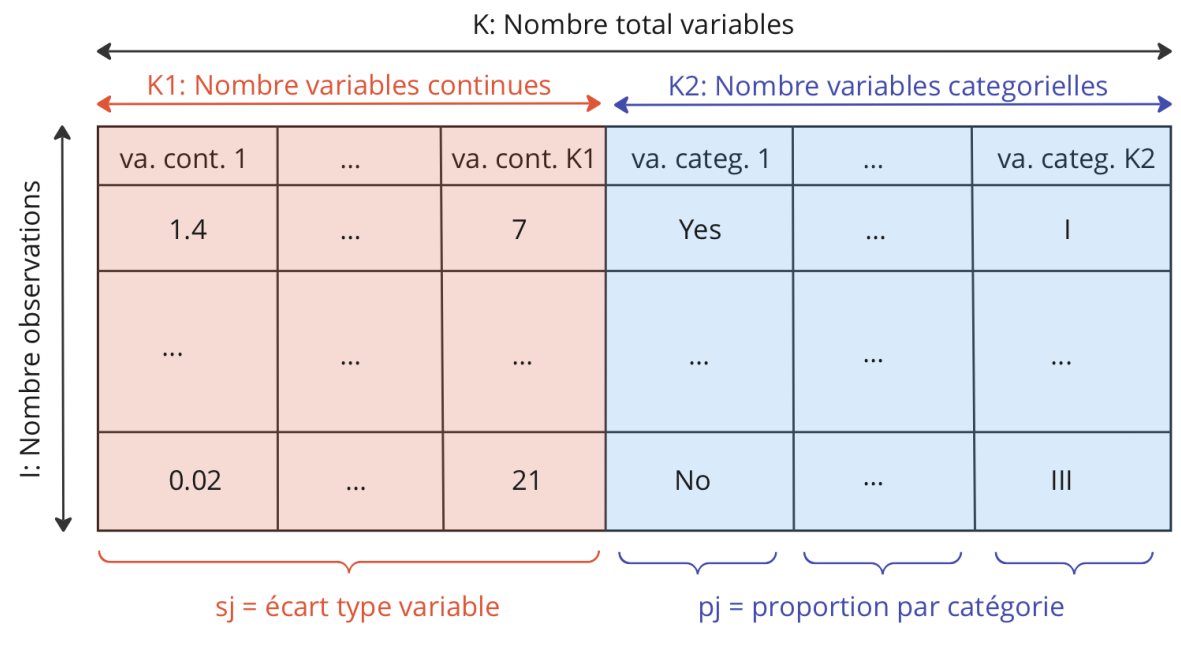
FAMD dans le cas complet

Basée sur l'ACP et analyse factorielle [5] [6]

Objectifs:

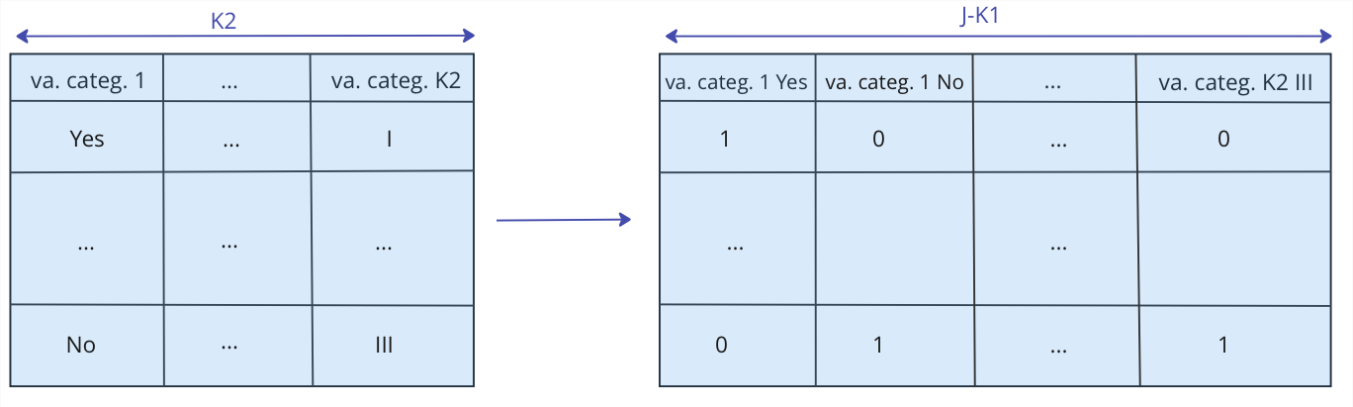
- Réduire dimension des données: maximiser variabilité des points projetés
- Équilibre influence variables continues et catégorielles

Ingrédients:



FAMD dans le cas complet

Étape 1: Encoder variables catégorielles



Étape 2: Pondération

- Variables continues: normalisation s_j $\forall j \in \{1, ..., K_1\}$
- Variables catégorielles: normalisation $\sqrt{p_j}$ $\forall j \in \{K_1 + 1, ..., J\}$

Étape 3: Décomposition SVD sur $(XD_\Sigma^{-1/2} - M)$

$D_\Sigma = diag(s_{x_1}^2, ..., s_{x_{K_1}}^2, p_{K_1+1}, ..., p_J)$

$M_{I \times J}$: matrice moyennes de $XD_\Sigma^{-1/2}$

FAMD itératif

Objectif: Imputer des données mixtes incomplètes

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

$\sigma^2 = \sum_{s=S+1}^{J-K_2} \frac{\lambda_s}{J-K_2-S}$

Propriétés

- Capacité à prendre en compte les **relations entre variables continues et catégorielles**.
- Des **relations linéaires** fortes garantissent des imputations précises.
- Prédictions précises pour les **catégories rares**.
- Choix du **nombre de dimensions**:
  - Faible nombre de dimension : perte d'information.
  - Nombre excessif de dimension : considère du bruit comme du signal.

Implémentations

Implémentations: Jeu de données synthétique

- Choisir dimension S et créer S variables indépendantes $\sim \mathcal{N}(0, I)$
- Répliquer K^s fois chacune des variables s ($s \in 1, ..., S$)

  $\longrightarrow$ Créer S groupes orthogonaux de variables corrélées
- Ajouter de bruit Gaussien
- Créer des variables catégorielles
- Insérer des données manquantes de manière aléatoire

Implémentations: Jeu de données synthétique

Relations linéaires et non-linéaires

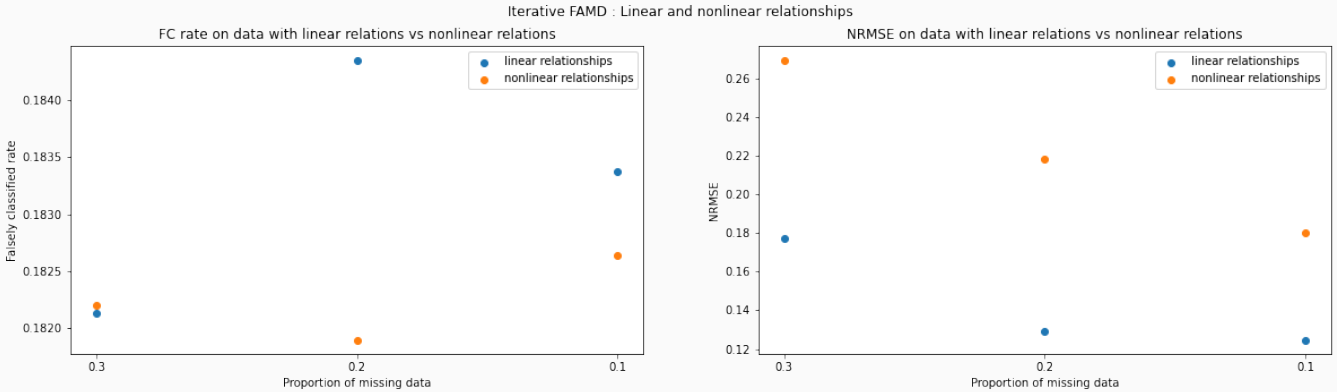


Figure 1: Performances de iFAMD sur de variables liées linéairement vs non-linéairement

Implémentations: Jeu de données synthétique

Catégories Rares

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 2: Performance de iFAMD sur 1000 simulations de données avec catégories rares (Résultats du papier [7])

Implémentations: Jeu de données synthétique

Choix du nombre de dimensions

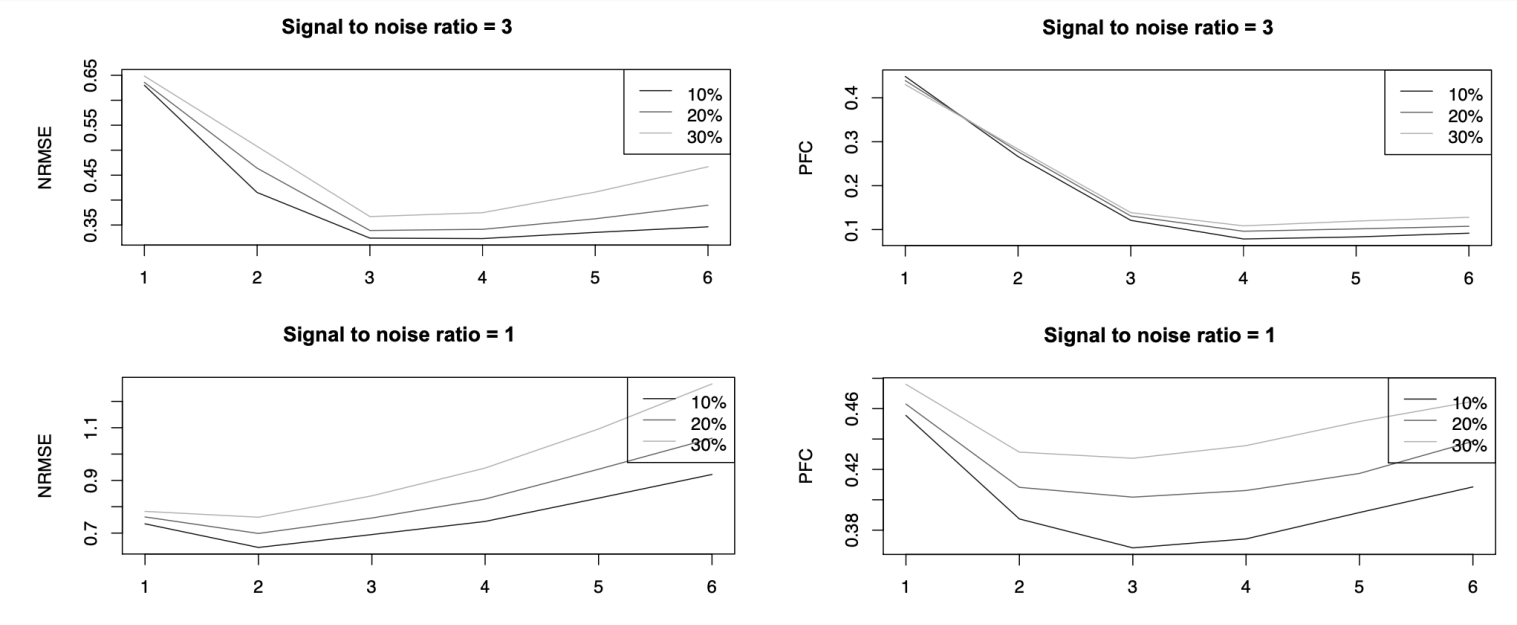


Figure 3: Erreur moyenne d'imputation sur 200 simulations en fonction du nombre de dimensions utilisé (Résultats du papier)

Implémentations: Jeu de données réel

Jeu de données GBSG

	pid	age	meno	size	grade	nodes	pgr	er	hormon	rfstime	status
1	132	49	0	18	2	2	0	0	0	1838	0
2	1575	55	1	20	3	16	0	0	0	403	1
3	1140	56	1	40	3	3	0	0	0	1603	0

- Données examinant l'impact d'un traitement hormonal sur le délai de réapparition du cancer du sein
- German Breast Cancer Study Group [8]
- $I = 686$ femmes; $K_1 = 7$; $K_2 = 4$

Implémentations: Jeu de données réel

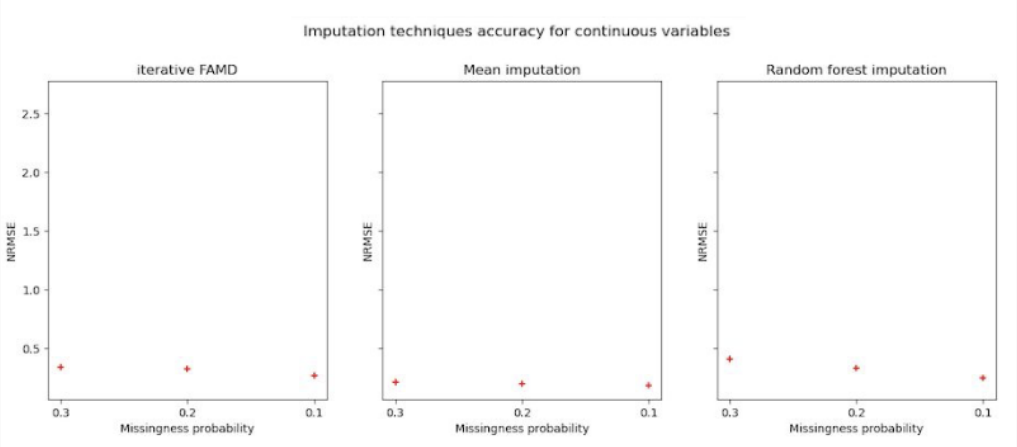


Figure 4: NRMSE en fonction de la probabilité de missingness

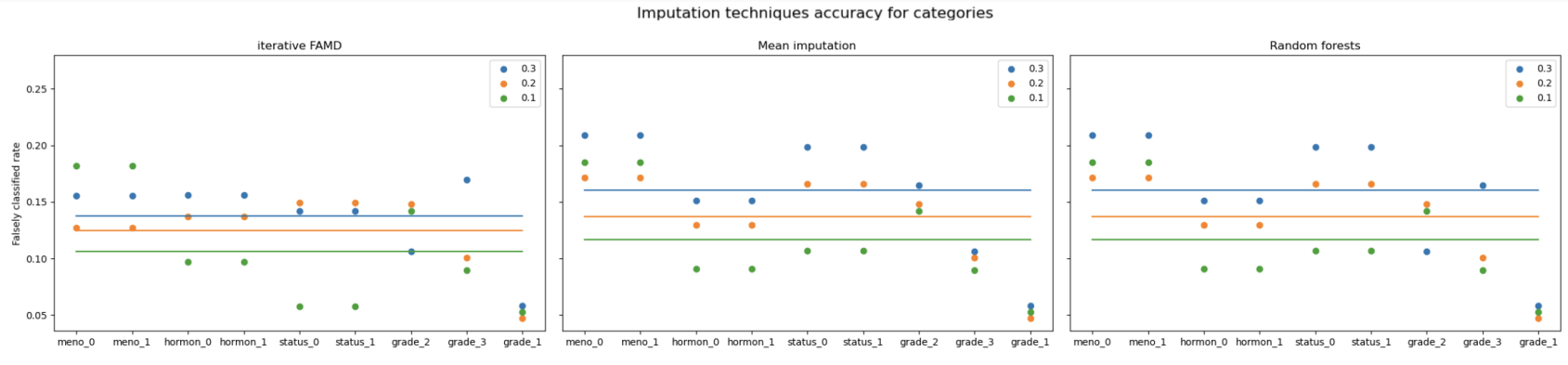


Figure 5: PFC en fonction de la probabilité de missingness

Implémentations: Jeu de données réel

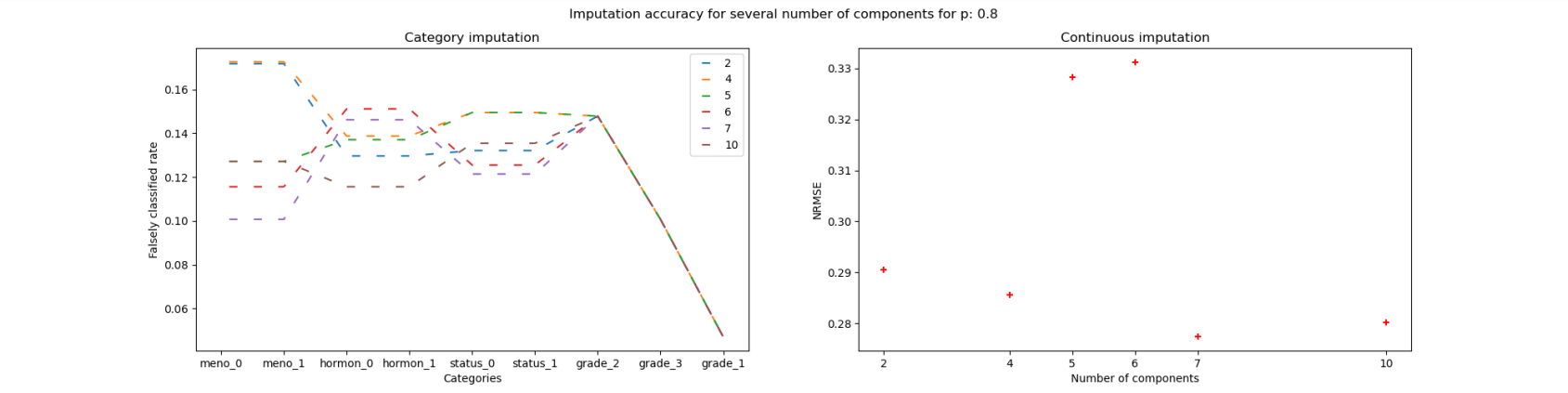


Figure 6: Étude des performances en fonction du nombre de composantes principales

Conclusion

Conclusion

- IFAMD prend en compte les similarités entre individus et les relations entre les variables.
- Efficace lorsque les relations sont **linéaires** et pour prédire les **catégories rares**.
- Performances se dégradent avec la proportion de données manquantes et le manque de relations.
- Hyperparamètre : Dimension (à choisir par **validation croisée**).

Références i

[1] Joseph L Schafer.
Analysis of incomplete multivariate data.
CRC press, 1997.

[2] Stef Van Buuren, Hendriek C Boshuizen, and Dick L Knook.
Multiple imputation of missing blood pressure covariates in survival analysis.
Statistics in medicine, 18(6):681–694, 1999.

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

Références ii

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

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

[6] Michael Greenacre and Jorg Blasius.
Multiple correspondence analysis and related methods.
Chapman and Hall/CRC, 2006.

[7] Vincent Audigier, François Husson, and Julie Josse.
A principal component method to impute missing values for mixed data.
Advances in Data Analysis and Classification, 10:5–26, 2016.

Références iii

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