Chapitre 5 : Analyse Factorielle des Correspondances Multiples (AFM)

Exercice 27

Nous traiterons des données fictives ou 27 races de chiens sont décrites avec 7 variables qualitatives.

Table 1: Extrait des données chiens

	taille	poids	velocite	intellig	affect	agress	fonction
beauceron	T++	P+	V++	I+	Af+	Ag+	Utilite
basset	T-	P-	V-	I-	Af-	Ag+	Chasse
ber_allem	T++	P+	V++	I++	Af+	Ag+	Utilite
boxer	T+	P+	V+	I+	Af+	Ag+	Compagnie
bull-dog	T-	P-	V-	I+	Af+	Ag-	Compagnie
bull-mass	T++	P++	V-	I++	Af-	Ag+	Utilite

Voici un extrait des données, nous avons 6 variables ordinales, la taille, le poids , la velocite, l'intelligence, l'affectation et l'agressivité, et une variable fonction qui determine l'utilité des chiens, qui peut être utilite, chasse ou compagnie. Pour notre analyse nous ne conserverons que les variables ordinales.

Nous allons réaliser un AFM, avec fonction comme variable supplémentaire.

Table 2: Valeurs propres

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.482	28.896	28.896
Dim.2	0.385	23.084	51.981
Dim.3	0.211	12.657	64.638
Dim.4	0.158	9.453	74.091
Dim.5	0.150	9.008	83.099
Dim.6	0.123	7.398	90.497
Dim.7	0.081	4.888	95.385
Dim.8	0.046	2.740	98.125
Dim.9	0.024	1.413	99.537
Dim.10	0.008	0.463	100.000

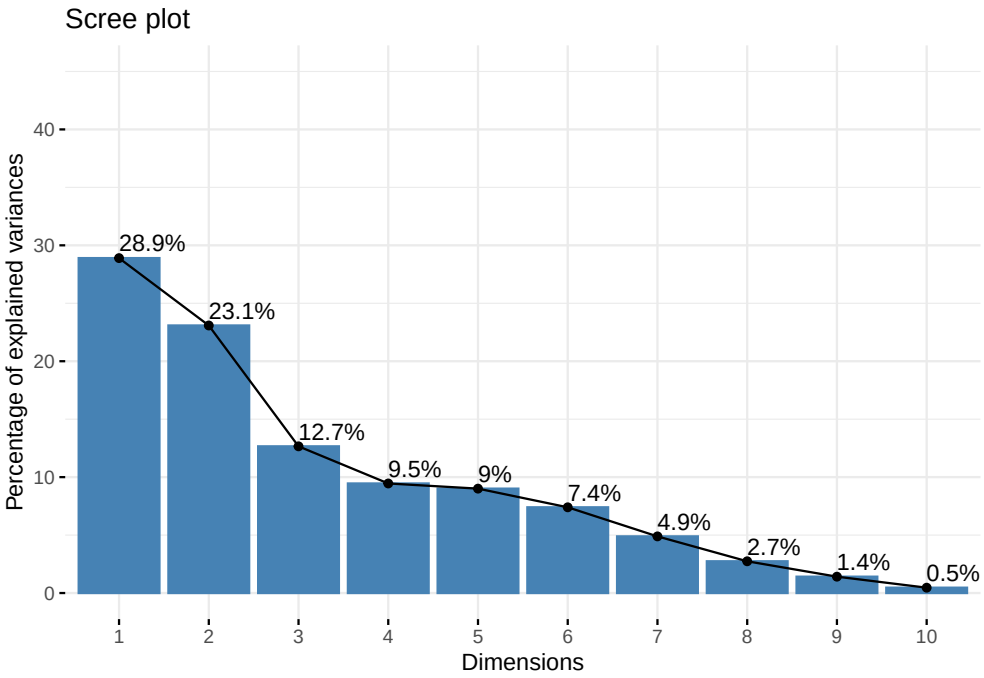

Figure 1: Visualisation des valeurs propres

On commence avec les valeurs propres, on voit avec le tableau qu'à partir de 4 dimensions, plus de 70% de l'inertie total, on conserve donc les 4 premiers axes. Avec le graphique ont voit que l'axe 1 explique 28.9%, l'axe 2 23.1%, l'axe 3 12.7% et l'axe 4 9.5%.

Table 3: Extrait des coordonnée des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	1.185	0.924	-0.616	0.120	-0.020
T+	0.851	-1.232	1.016	0.342	-0.310
T++	-0.837	-0.021	-0.051	-0.170	0.113
P-	1.169	0.824	-0.359	0.165	-0.051
P+	-0.305	-0.819	-0.231	-0.118	-0.190
P++	-1.015	0.974	1.222	0.068	0.615

On se focalise d'abord sur les variables. On obtient dans le tableau ci-dessus les coordonnées afin de tracer le grapahique des variable.

Table 4: Extrait des Cos2 des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	0.491	0.299	0.133	0.005	0.000
T+	0.165	0.345	0.235	0.027	0.022
T++	0.875	0.001	0.003	0.036	0.016
P-	0.575	0.286	0.054	0.011	0.001
P+	0.100	0.722	0.058	0.015	0.039
P++	0.234	0.216	0.339	0.001	0.086

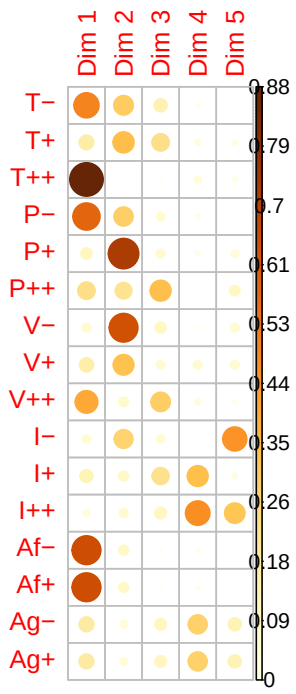

Figure 2: Visualisation des cos2 des variables

On se penche ensuite sur les qualités de représentations (cos2) des variables. On voit à l'aide du graphique et du tableau que une grande taille aura une bonne qualité de représentation sur l'axe 1.

Table 5: Extrait des contributions des variables

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
T-	12.598	9.587	7.772	0.396	0.011
T+	4.642	12.171	15.104	2.297	1.976
T++	13.459	0.010	0.115	1.703	0.783
P-	14.010	8.722	3.013	0.852	0.086
P+	1.674	15.062	2.191	0.768	2.082
P++	6.604	7.609	21.833	0.090	7.763

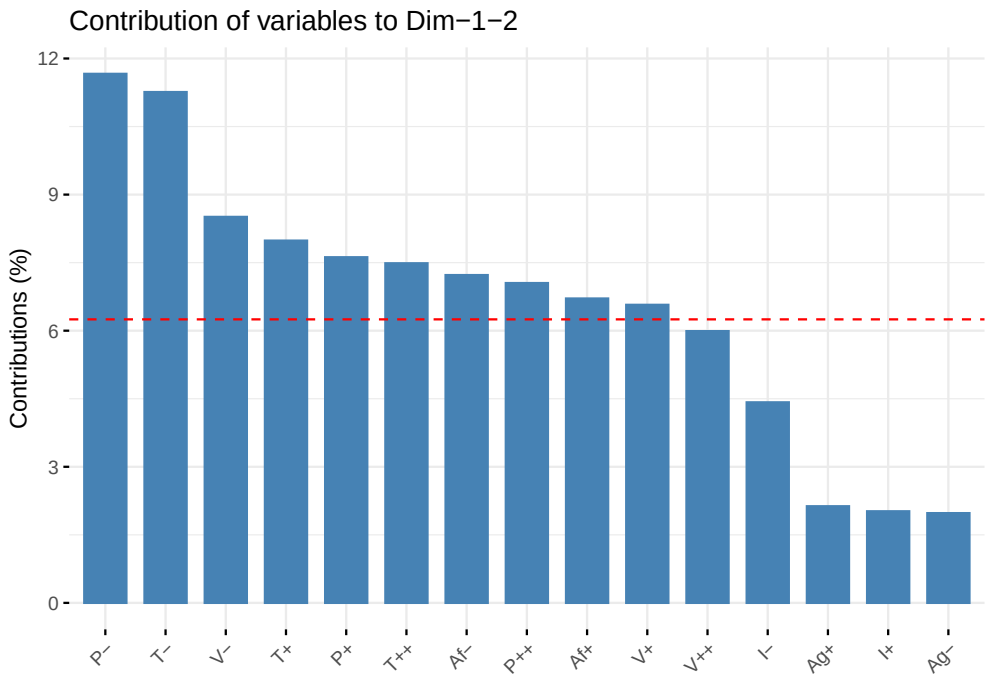

Figure 3: Visualisation des contributions pour les variables

Pour les contributions des variables sur le premier plan. On remarque avec le graphique et le tableau que un faible poids à la meilleur contribution au premier plan. Sur le graphique toutes les variables au dessus de la ligne pointillée rouge peuvent être considéré comme suffisament contribuant au premier plan.

On passe maintenant au individus.

Table 6: Extrait des coordonnées des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	-0.317	-0.418	-0.101	-0.211	-0.119
basset	0.254	1.101	-0.191	0.293	-0.524
ber_allem	-0.486	-0.464	-0.498	0.577	0.276
boxer	0.447	-0.882	0.692	0.260	-0.456
bull-dog	1.013	0.550	-0.163	-0.350	0.331
bull-mass	-0.753	0.547	0.498	0.655	0.722

D'abord avec le tableau des coordonnées.

Table 7: Extrait des Cos2 des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	0.089	0.154	0.009	0.039	0.012
basset	0.034	0.635	0.019	0.045	0.144
ber_allem	0.154	0.140	0.161	0.217	0.049
boxer	0.111	0.433	0.266	0.038	0.115
bull-dog	0.624	0.184	0.016	0.074	0.067
bull-mass	0.271	0.143	0.118	0.205	0.249

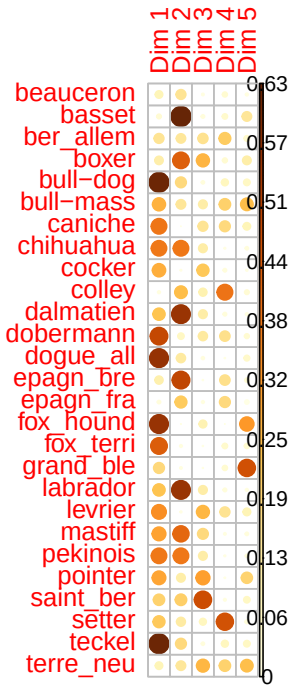

Figure 4: Visualisation des cos2 des individus

Ensuite avec les qualités de représentations des individus. On ne voit par exemple avec le graphique que les teckel et les bull-dog ont la meilleur qualité de représentation sur la dimension 1.

Table 8: Extrait des contributions des individus

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
beauceron	0.774	1.680	0.181	1.051	0.346
basset	0.497	11.674	0.638	2.013	6.774
ber_allem	1.819	2.077	4.357	7.838	1.878
boxer	1.539	7.485	8.408	1.589	5.120
bull-dog	7.897	2.911	0.469	2.878	2.699
bull-mass	4.356	2.879	4.347	10.090	12.858

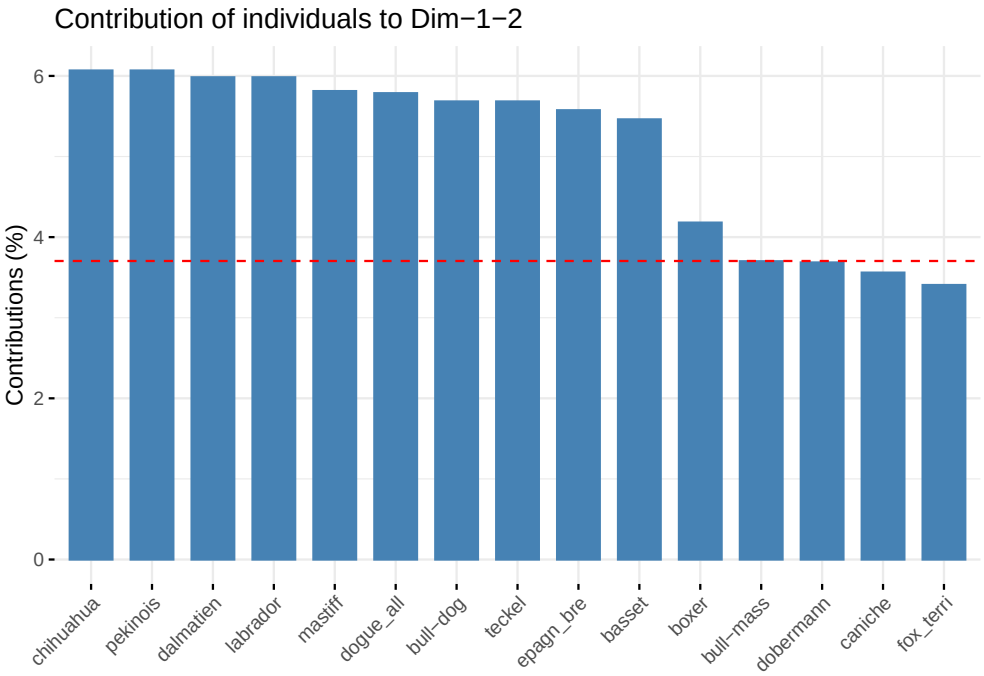

Figure 5: Visualisation des contributions des individus

Pour les contributions des individu sur le premier plan. On remarque avec le graphique les chihuahua ont la meilleur contribution au premier plan. Toutes les individus au dessus de la ligne pointillée rouge peuvent être considéré comme suffisament contribuant au premier plan.

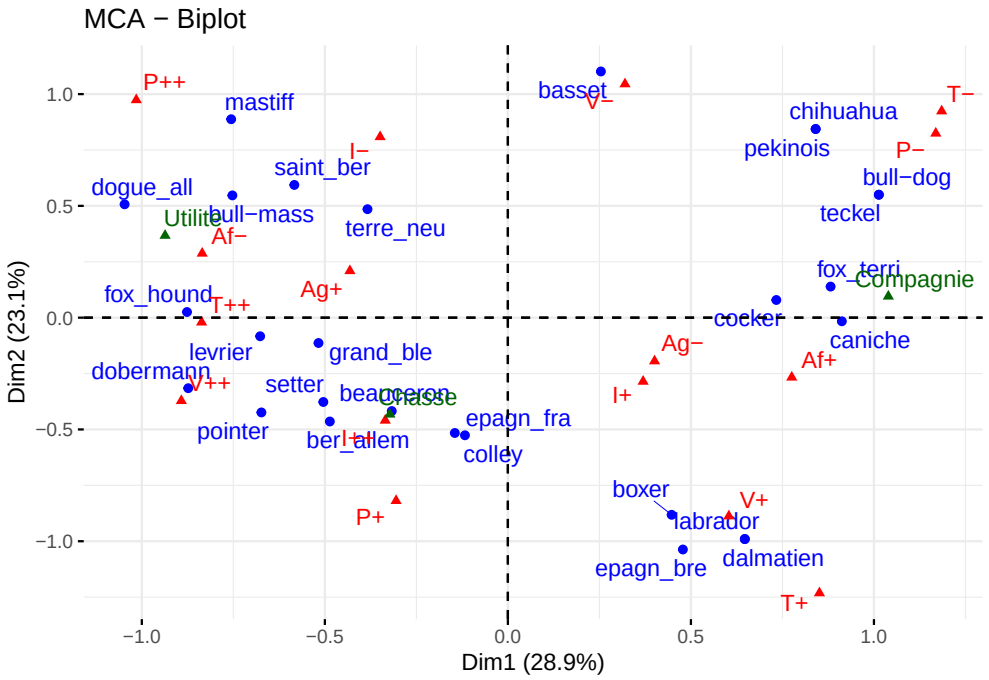

Figure 6: Bitplot

On peut enfin tracer le bitplot. Les individus sont en bleu, les variables sont en rouge et les variables supplémentaires sont en vert foncé.

On peut faire des liens entre les individus et les variables, tous les individus proche les un des autres peuvent être considéré comme des profils similaire. par exemple on voit que les boxers, les labradors, les dalmatiens et les espagn_bre sont silimaires avec une grande taille et une vitesse élevé.

Quand on regarde les variables supplémentaires, on voit qu'elles sont éloigner les une des autres surtout pour les chiens de compagnie qu'on arrive bien à distinguer des deux autres.

On vas s'intéresser au rapports de corrélations entre les variables qualitatives et les deux premièrescomposantes principales

Table 9: Rapports de corrélations entre les variables qualitatives et les deux premières composantes principales

	Dim 1	Dim 2
taille	0.887	0.502
poids	0.644	0.725
velocite	0.411	0.684
intellig	0.127	0.280
affect	0.648	0.077
agress	0.173	0.041

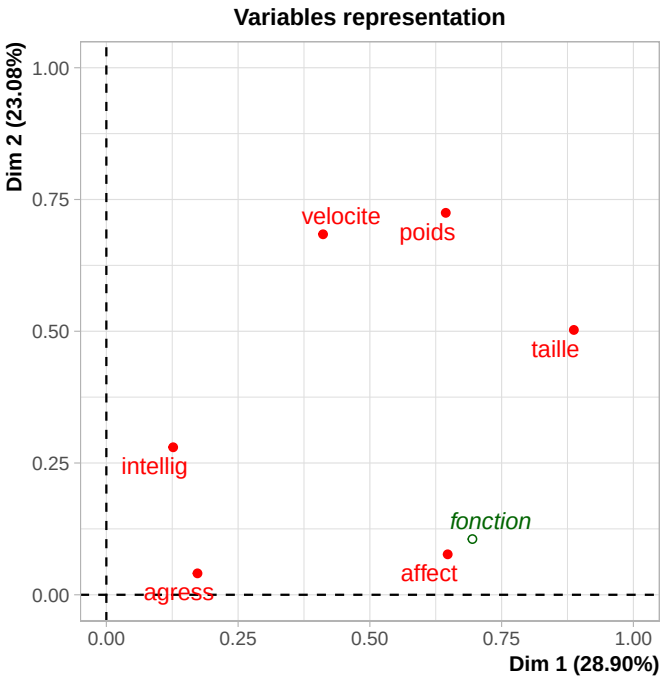

Figure 7: Visualisation des rapports de corrélation

On voit avec le tableau et le graphique que le poids est la variable la plus corrélée à l'axe 1 tandis que le poids et la plus corrélée à l'axe 2.

On décide ensuite de rajouter des données manquantes a nos données, et nous refaisons une AFM, pour voir si elles sont prises en compte.

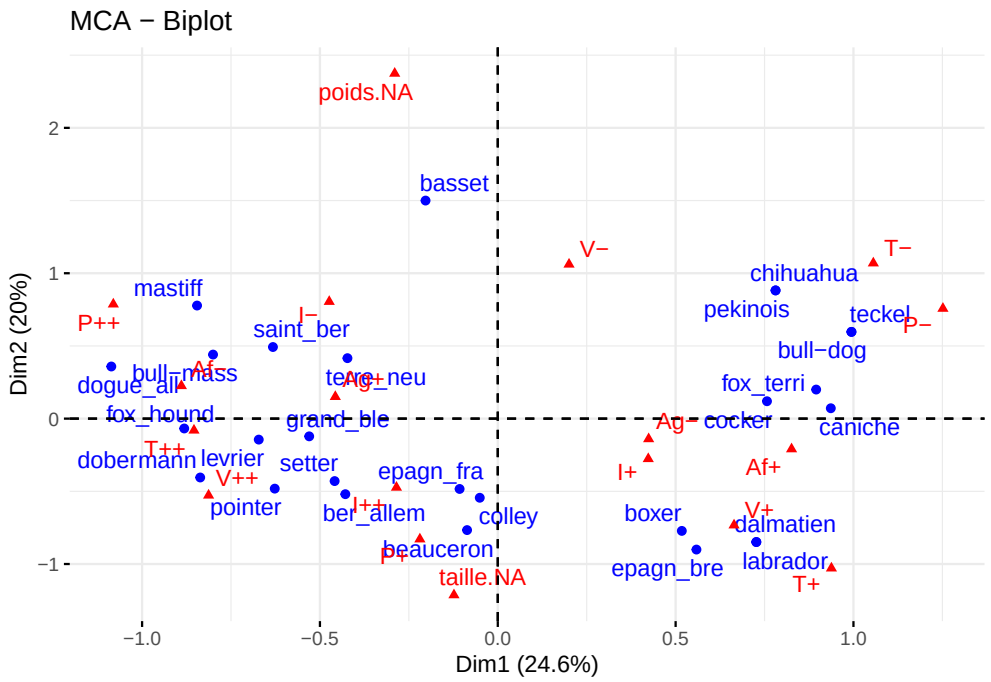

Figure 8: Bitplot avec les données manquantes

Quand on refait le bitplot on voit bien des points supplémentaire avec .NA en sufixe, donc les données manquantes sont prises en compte par la fonction MCA comme des individus classique, ce qui n'est pas correcte.

On veut maintenant comparer l'ACM et l'AFC dans le cas particulier de deux variables qualitative. Noius allons réaliser l'AFC du tableau de contingence croisant les variables taille et poids, et comparer les valeurs propres.

Table 10: Valeurs propre de l'AFC

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.861	91.743	91.743
Dim.2	0.077	8.257	100.000

Table 11: Valeurs propre de l'AFM

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.964	48.193	48.193
Dim.2	0.639	31.958	80.151
Dim.3	0.361	18.042	98.193
Dim.4	0.036	1.807	100.000

Table 12: Valeurs propres AFM avec l'AFC

	Valeurs propres
dim 1	0.964
dim 2	0.639
dim 4	0.036
dim 3	0.361

On retrouve un lien entre les valeurs propres de l'AFC et l'ACM. Quand on joue avec la racine carrée des les valeurs propres de l'AFC, on arrive à retouvé les valeurs propres de l'ACM.

$\frac{1+\sqrt{vpDim1AFC}}{2} = vp\ dim1\ ACM$

$\frac{1+\sqrt{vpDim2AFC}}{2} = vp\ dim2\ ACM$

$\frac{1-\sqrt{vpDim1AFC}}{2} = vp\ dim3\ ACM$

$\frac{1-\sqrt{vpDim2AFC}}{2} = vp\ dim4\ ACM$

Ou vp sont les valeurs propres selon la méthode et la dimension.

Exercice 28

Dans cette partie, nous allons présenter le package R missMDA. Il gére les données manquantes en ACP et en ACM, et de choisir le nombre de composantes par validation croisée. Nous décrirons les principales fonctionnalités de ce package, avec à chaque fois une explication de la méthode.

Overimpute : Évaluez l'ajustement de la distribution prédictive après avoir effectué une imputation multiple

estim_ncpPCA : Estime le nombre de dimensions pour l'Analyse en Composantes Principales par validation croisée

MIFAMD : effectue des imputations multiples pour des données mixtes (continues et catégorielles) en utilisant l'analyse factorielle de données mixtes.

estim_ncpMultilevel : Estimez le nombre de dimensions pour la composante principale multiniveau (ACP multiniveau, AMC multiniveau ou analyse factorielle multiniveau de données mixtes) par validation croisée.

estim_ncpMCA : Estimer le nombre de dimensions pour l'Analyse des Correspondances Multiples par validation croisée

MIPCA : Réalise une imputation multiple avec un modèle ACP. Peut être utilisé comme étape préliminaire pour effectuer une imputation multiple dans l'ACP.

MIMCA : Effectue des imputations multiples pour des données catégorielles en utilisant l'analyse des correspondances multiples.

estim_ncpFAMD : Estime le nombre de dimensions pour l'Analyse Factorielle de Données Mixtes par validation croisée

prelim : Cette fonction effectue des opérations de regroupement et de tri sur un ensemble de données imputées à plusieurs reprises. Elle crée un objet mids qui est nécessaire à l'entrée de with.mids, qui permet d'analyser l'ensemble de données imputées à plusieurs reprises. L'ensemble de données incomplètes d'origine doit être disponible pour que nous sachions où se trouvent les données manquantes.

imputeFAMD : Imputez les valeurs manquantes d'un ensemble de données mixtes (avec des variables continues et catégorielles) en utilisant la méthode des composantes principales "analyse factorielle pour données mixtes" (FAMD). Peut être utilisé comme une étape préliminaire avant d'exécuter FAMD sur un ensemble de données incomplet.

imputeMFA : Impute un jeu de données avec des variables structurées en groupes de variables (groupes de variables continues ou catégorielles).

imputeMCA : Imputez les valeurs manquantes d'un ensemble de données catégoriques en utilisant l'analyse des correspondances multiples (ACM). Peut être utilisé comme une étape préliminaire avant d'effectuer l'ACM sur un ensemble de données incomplet.

imputeCA : Imputez les entrées manquantes d'un tableau de contingence en utilisant l'analyse des correspondances (AC). Peut être utilisé comme une étape préliminaire avant d'effectuer l'AC sur un ensemble de données incomplet.

imputePCA : Impute les valeurs manquantes d'un jeu de données avec le modèle d'analyse en composantes principales. Peut être utilisé comme une étape préliminaire avant d'effectuer une ACP sur un jeu de données complet.

imputeMultilevel : Imputez les valeurs manquantes d'un ensemble de données mixtes multi-niveaux (avec une variable qui regroupe les individus, et avec des variables continues et catégorielles) en utilisant la méthode des composantes principales "analyse factorielle multi-niveaux pour données mixtes".

plot.MIMCA : À partir des ensembles de données imputées multiples, la fonction trace des graphiques pour les individus, les catégories et les dimensions pour l'analyse des correspondances multiples (ACM).

plot.MIPCA : À partir des ensembles de données imputées multiples, la fonction trace des graphiques pour les individus, les variables et les dimensions pour l'analyse en composantes principales (ACP).