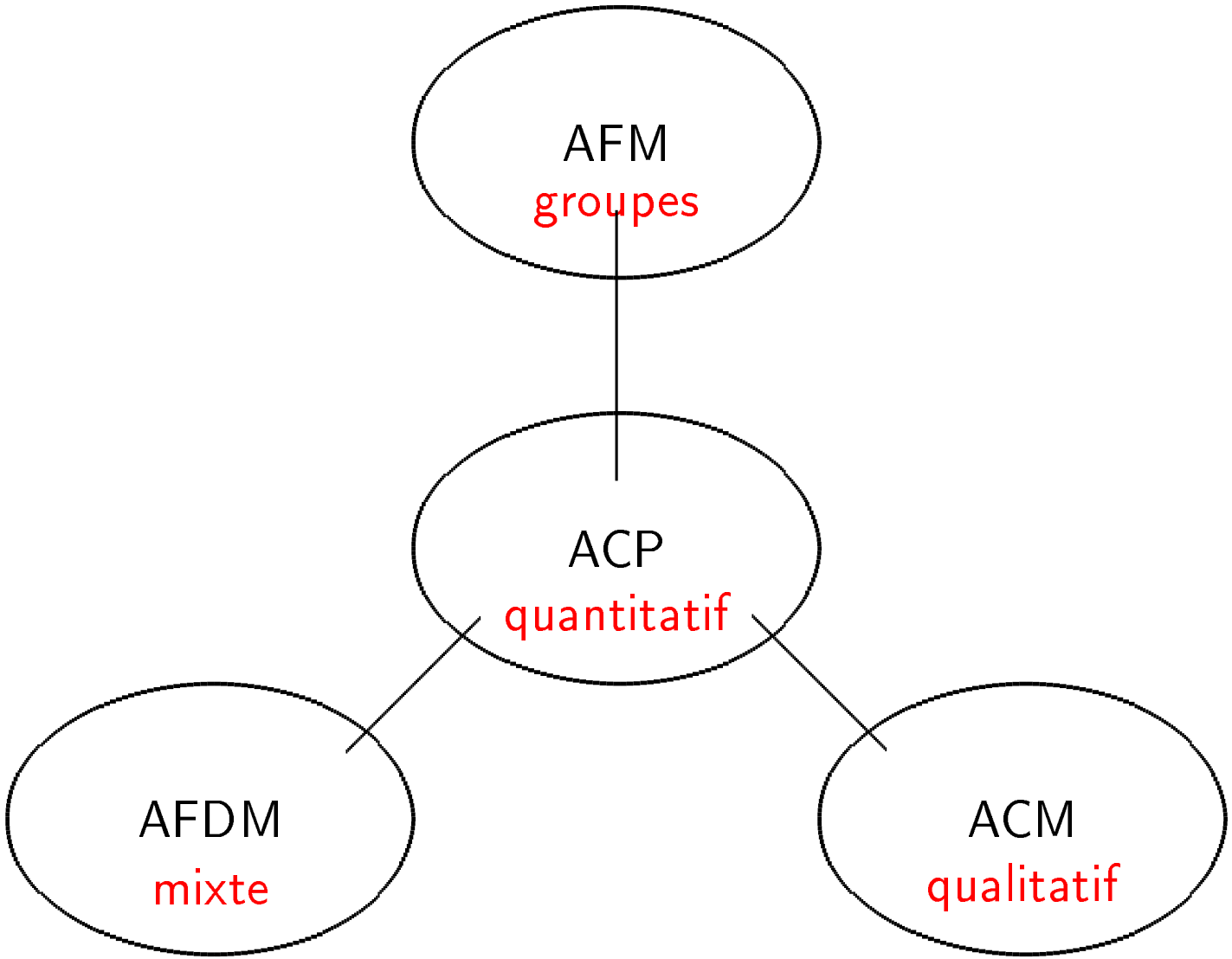
Imputation des données manquantes pour des données mixtes via les méthodes factorielles grâce à `missMDA`

Vincent Audigier & Julie Josse & François Husson

Agrocampus Rennes

useR! 2012, Bordeaux, 2-3 Juillet 2012

L'ACP au cœur des méthodes d'analyse factorielle



L'analyse factorielle pour imputer

- `missMDA` pour des analyses factorielles avec données manquantes
- comment effectuer une ACP sur tableau incomplet ?
- comment cela permet d'imputer ?

L'ACP sur données manquantes

- ACP sur données complètes : minimiser

$$\mathcal{C} = ||\mathbf{X}_{I\times J} - \mathbf{F}_{I\times S}\mathbf{U}'_{S\times J}||^2$$

- ACP sur données manquantes : minimiser

$$\mathcal{C} = ||\mathbf{W}_{I\times J} * \left(\mathbf{X}_{I\times J} - \mathbf{F}_{I\times S}\mathbf{U}'_{S\times J}\right)||^2$$

$$w_{ij} = \begin{cases} 0 \text{ si } x_{ij} \text{ est manquant} \\ 1 \text{ sinon} \end{cases}$$

Algorithme : ACP itérative

1. initialisation $\ell = 0$: imputation par la moyenne $\longrightarrow \mathbf{X}^0$
2. itération ℓ :
   (a) ACP sur $\mathbf{X}^{\ell-1} \longrightarrow$ recherche de $\hat{\mathbf{F}}^{\ell}_{I\times S}$ et $\hat{\mathbf{U}}^{\ell}_{J\times S}$ (estimation)
   (b) $\mathbf{X}^{\ell} \leftarrow \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{F}}^{\ell}\hat{\mathbf{U}}^{\ell'}$ (imputation)
3. l'étape 2 est répétée jusqu'à la convergence

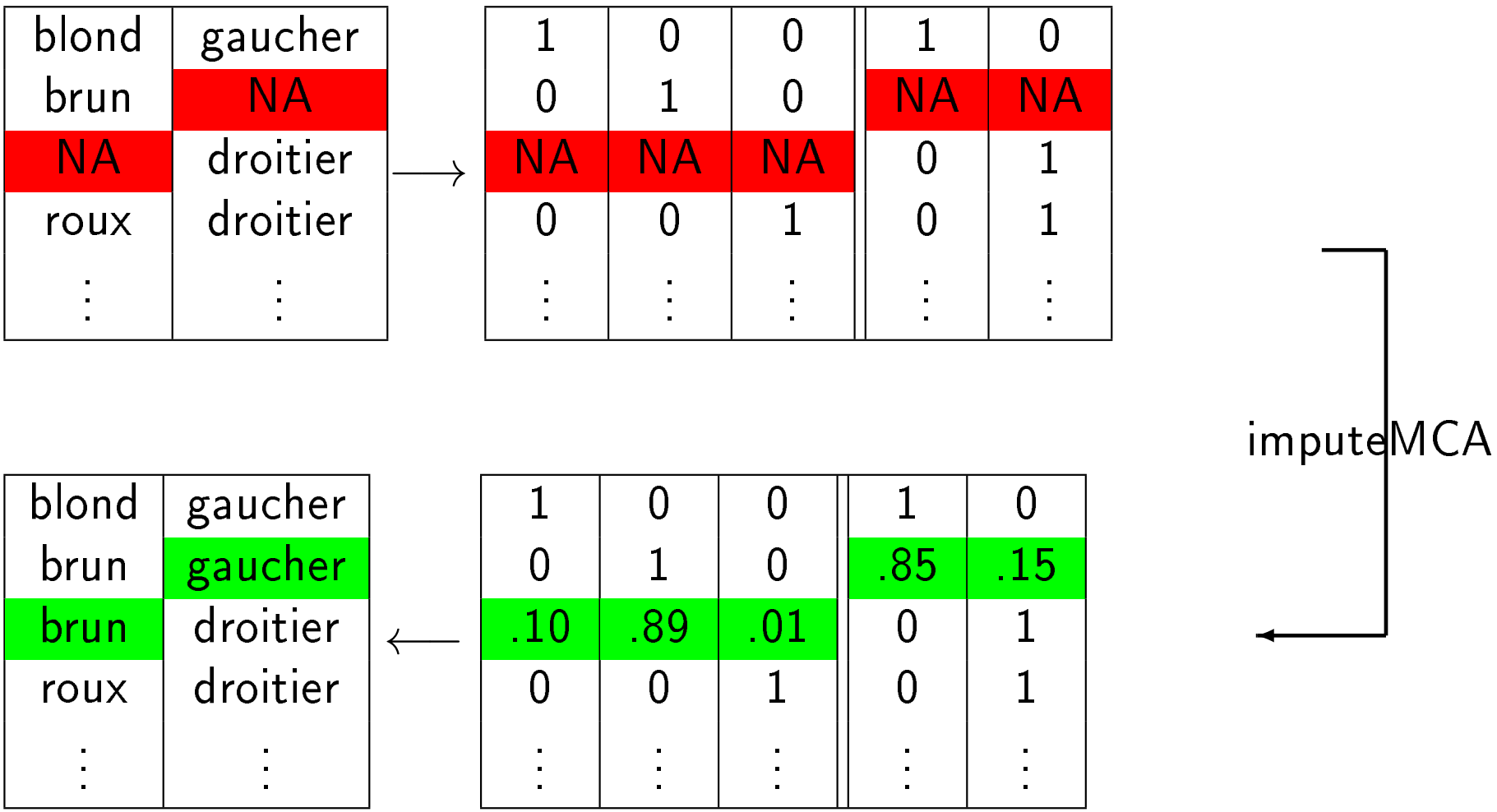
ACP itérative pour imputer des données quantitatives

Sweet	Acid	Bitter	Pulp	Typ
NA	NA	2.83	NA	5.21
5.46	4.13	3.54	4.62	4.46
NA	4.29	3.17	6.25	5.17
4.17	6.75	NA	1.42	3.42
...	...	...	...	...
NA	NA	NA	7.33	5.25
4.88	5.29	4.17	1.50	3.50

$\xrightarrow{imputePCA}$

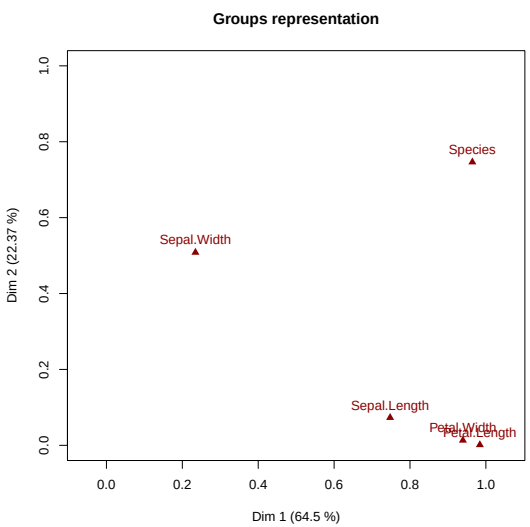
Sweet	Acid	Bitter	Pulp	Typ
5.54	4.13	2.83	5.89	5.21
5.46	4.13	3.54	4.62	4.46
5.45	4.29	3.17	6.25	5.17
4.17	6.75	4.73	1.42	3.42
...	...	...	...	...
5.71	3.87	2.80	7.33	5.25
4.88	5.29	4.17	1.50	3.50

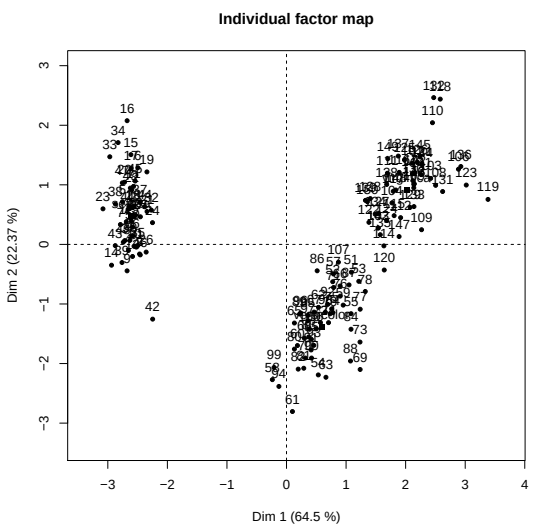
Imputer des données qualitatives

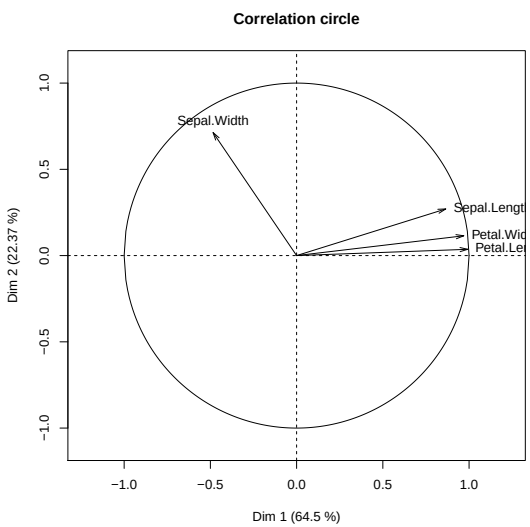


L'Analyse Factorielle des Données Mixtes

Quantitatif		Qualitatif
Tps à l'étranger	Nb pts TOEIC	Nationalité
5.0	301	FR
6.4	567	CZ
7.9	900	NL
⋮	⋮	⋮







Imputer des données mixtes

Tps	Pts	Nationalité
NA	301	FR
6.4	NA	CZ
7.9	900	NA
⋮	⋮	⋮

→

Tps	Pts	FR	CZ	NL
NA	301	1	0	0
6.4	NA	0	1	0
7.9	900	NA	NA	NA
⋮	⋮	⋮	⋮	⋮

impute**AFDM**

Tps	Pts	FR	CZ	NL
4.9	301	1	0	0
6.4	549	0	1	0
7.9	900	.1	.1	.8
⋮	⋮	⋮	⋮	⋮

←

Tps	Pts	Nationalité
4.9	301	FR
6.4	549	CZ
7.9	900	NL
⋮	⋮	⋮

Propriétés

- point fort : prise en compte des relations entre individus et entre variables
- point faible : détermination du nombre d'axes
- remarque : régularisation nécessaire

Littérature

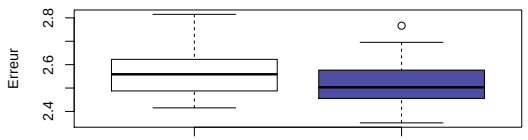
- peu de méthodes existantes
- Stekhoven et Bühlmann (2011) proposent une imputation par forêts aléatoires

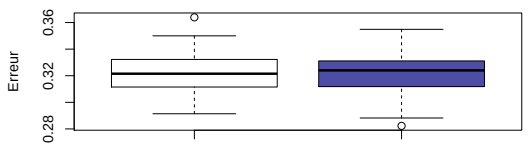
Deux critères d'erreurs pour s'y comparer :

- taux de mauvais classement pour les variables qualitatives
- RMSE normalisée pour les variables quantitatives

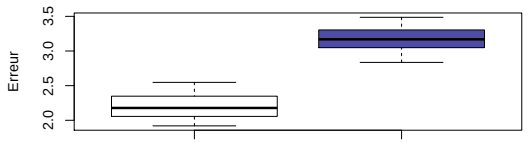
Simulations

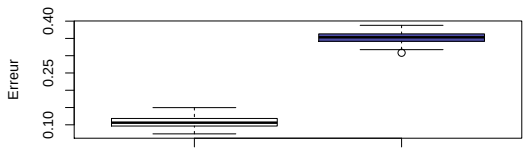
Mixte



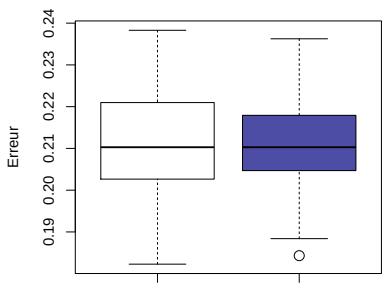


Mixte

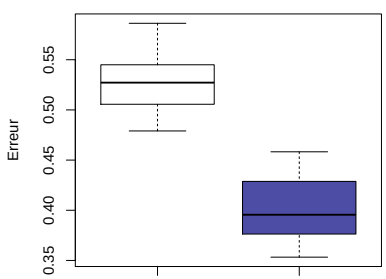




Qualitatif



Quantitatif



Conclusion

`missMDA` permet :

- d'imputer un tableau de données quantitatives et/ou qualitatives
- d'effectuer toute analyse factorielle avec données manquantes (à l'aide de `FactoMineR` par exemple)