Analyse Factorielle Multiple

François Husson

Laboratoire de mathématiques appliquées - AGROCAMPUS OUEST

husson@agrocampus-ouest.fr

Plan

Description sensorielle de vins de Loire

- 10 vins blancs du Val de Loire : 5 Vouvray - 5 Sauvignon
- descripteurs sensoriels : acidité, amertume, odeur agrume, etc.

Description sensorielle de vins de Loire

- 10 vins blancs du Val de Loire : 5 Vouvray - 5 Sauvignon
- descripteurs sensoriels : acidité, amertume, odeur agrume, etc.

	O. fruité	O. passion	O. citron	...	Sucré	Acidité	Amertume	Astringence	Intensité arôme	Persistance arôme	Intensité visuel	Cépage
S Michaud	4,3	2,4	5,7	...	3,5	5,9	4,1	1,4	7,1	6,7	5,0	Sauvignon
S Renaudie	4,4	3,1	5,3	...	3,3	6,8	3,8	2,3	7,2	6,6	3,4	Sauvignon
S Trotignon	5,1	4,0	5,3	...	3,0	6,1	4,1	2,4	6,1	6,1	3,0	Sauvignon
S Buisse Domaine	4,3	2,4	3,6	...	3,9	5,6	2,5	3,0	4,9	5,1	4,1	Sauvignon
S Buisse Cristal	5,6	3,1	3,5	...	3,4	6,6	5,0	3,1	6,1	5,1	3,6	Sauvignon
V Aub Silex	3,9	0,7	3,3	...	7,9	4,4	3,0	2,4	5,9	5,6	4,0	Vouvray
V Aub Marigny	2,1	0,7	1,0	...	3,5	6,4	5,0	4,0	6,3	6,7	6,0	Vouvray
V Font Domaine	5,1	0,5	2,5	...	3,0	5,7	4,0	2,5	6,7	6,3	6,4	Vouvray
V Font Brûlés	5,1	0,8	3,8	...	3,9	5,4	4,0	3,1	7,0	6,1	7,4	Vouvray
V Font Coteaux	4,1	0,9	2,7	...	3,8	5,1	4,3	4,3	7,3	6,6	6,3	Vouvray

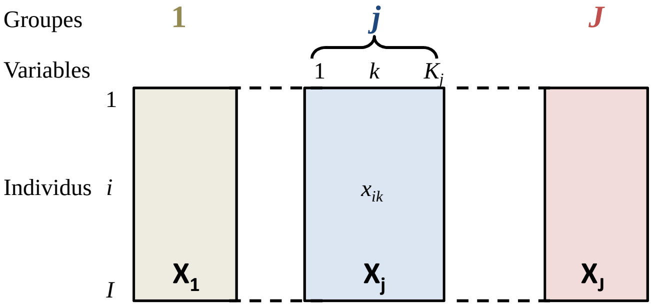
Description sensorielle de vins : comparaison de jurys

- 10 vins blancs du Val de Loire : 5 Vouvray - 5 Sauvignon
- description sensorielle de 3 jurys : œnologue, conso., étudiant
- notes hédoniques de 60 consommateurs : appréciation globale

	Expert (27)	Conso (15)	Etudiant (15)	Appréciation (60)	Cépage (1)
Vin 1					
Vin 2					
...					
Vin 10					

- Comment caractériser les vins ?
- Les vins sont-ils décrits de la même façon par les différents jurys ? Y-a t'il des spécificités par jury ?

Tableaux multiples



Exemples avec des variables **quantitatives et/ou qualitatives** :

- génomique : ADN, expression, protéines
- questionnaires : santé des étudiants (consommations de produits, conditions psychologiques, sommeil, signalétique)
- économie : indicateurs économiques chaque année

Objectifs

- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables ET les relations entre variables

Prendre en compte la structure en groupes

- Etudier globalement les ressemblances et les différences entre groupes (voir les spécificités de chaque groupe)
- Etudier les ressemblances et les différences entre groupes du point de vue individuel
- Comparer les typologies issues des analyses séparées

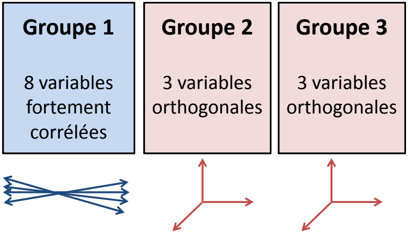
$\Rightarrow$ Equilibrer l'influence de chaque groupe dans l'analyse

Plan

Equilibrer l'influence des groupes de variables

En ACP : normer équilibre l'influence de chaque variable (dans le calcul des distances entre individus i et i')
En AFM, on équilibre les groupes

1ère idée : diviser chaque variable par l'inertie totale du groupe auquel elle appartient



2ème idée : diviser chaque variable par la (racine carrée de la) 1ère valeur propre du groupe auquel elle appartient

Equilibrer l'influence des groupes de variables

"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize"
Benzécri

L'AFM est une ACP pondérée :

- calculer la 1ère valeur propre λ_1^j du groupe de variables j ($j = 1, ..., J$)
- réaliser l'ACP globale sur le tableau pondéré :

$$\left[\frac{X_1}{\sqrt{\lambda_1^1}}; \frac{X_2}{\sqrt{\lambda_1^2}}; ...; \frac{X_J}{\sqrt{\lambda_1^J}}\right]$$

X_j correspond ici au tableau j centré ou centré-réduit

Equilibrer l'influence des groupes de variables

	Avant pondération			Après pondération		
	Expert	Etudiant	Conso	Expert	Etudiant	Conso
λ_1	11.74	7.89	7.17	1.00	1.00	1.00
λ_2	6.78	3.83	2.59	0.58	0.49	0.36
λ_3	2.74	1.70	1.63	0.23	0.22	0.23

- Même poids pour toutes les variables d'un même groupe : la structure du groupe est préservée
- Pour chaque groupe, la variance de la principale dimension de variabilité (première valeur propre) est égale à 1
- Aucun groupe ne peut générer à lui seul la première dimension
- Un groupe multidimensionnel contribue à plus de dimensions qu'un groupe uni-dimensionnel

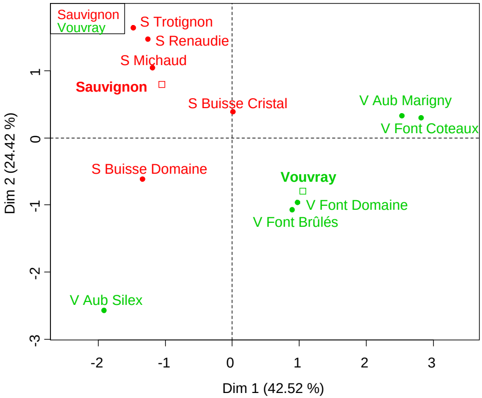
L'AFM une ACP pondérée

⇒ Mêmes représentations qu'en ACP

- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables
- Etudier les relations entre variables
- Décrire les individus à partir des variables

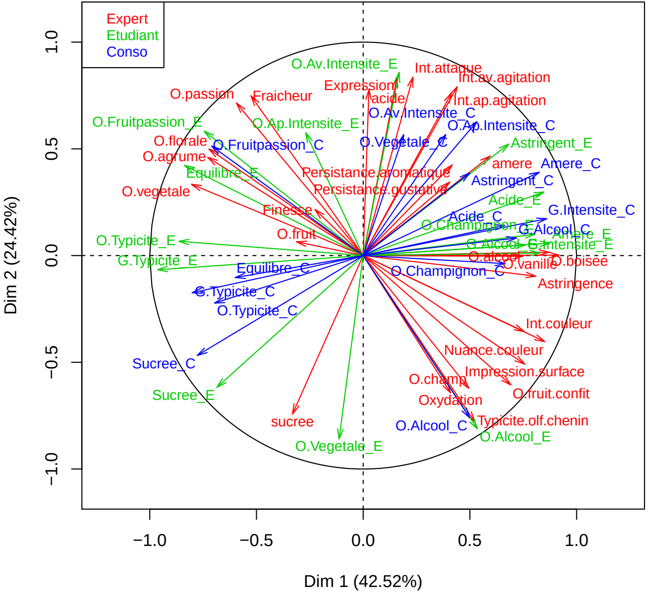
⇒ Mêmes sorties (coordonnées, cosinus, contributions)
⇒ Ajouter des individus et variables (quantitatives et qualitatives) en supplémentaire

Représentation des individus

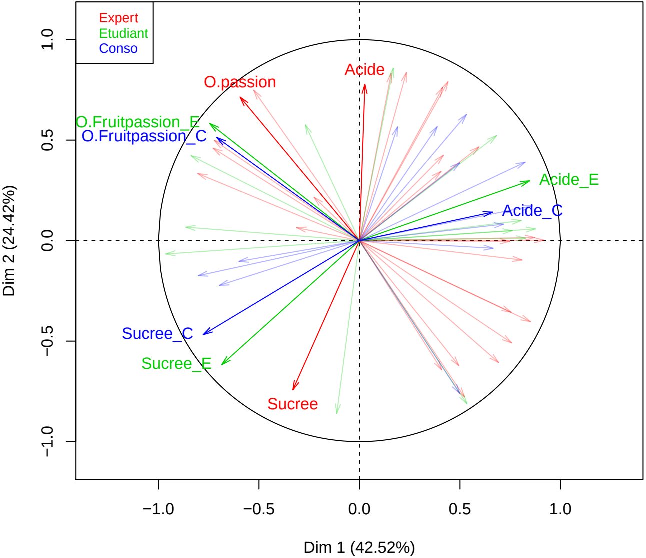


- Les deux cépages sont bien séparés
- Les Vouvray sont plus différents du point de vue sensoriel
- Plusieurs groupes de vins, ...

Représentation des variables



Représentation des variables



Plan

Première composante de l'AFM

En ACP (rappel) : $\arg\max_{v_1 \in \mathbb{R}^I} \sum_{k=1}^{K} cov^2(x_{.k}, v_1)$

En AFM :

$$\arg\max_{v_1 \in \mathbb{R}^I} \sum_{j=1}^{J} \sum_{k \in K_j} cov^2\left(\frac{x_{.k}}{\sqrt{\lambda_1^j}}, v_1\right) = \arg\max_{v_1 \in \mathbb{R}^I} \sum_{j=1}^{J} \underbrace{\frac{1}{\lambda_1^j} \sum_{k \in K_j} cov^2(x_{.k}, v_1)}_{\mathcal{L}_g(K_j, v_1)}$$

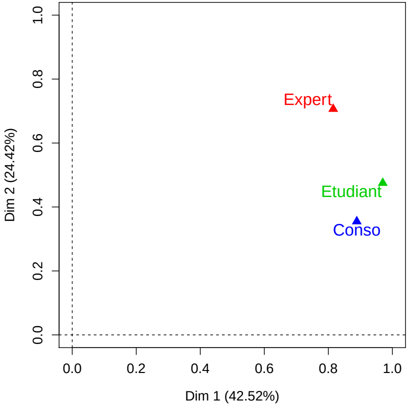
$\mathcal{L}_g(K_j, v_1)$ = inertie projetée de toutes les variables de K_j sur v_1 $\Rightarrow$ La première composante principale de l'AFM est la variable qui maximise la liaison avec tous les groupes au sens du $\mathcal{L}_g$

$$0 \leq \mathcal{L}_g(K_j, v_1) \leq 1$$

$\mathcal{L}_g = 0$: toutes les variables du groupe j sont non-corrélées à v_1
$\mathcal{L}_g = 1$: v_1 confondue avec la 1ère composante principale de K_j

Représentation des groupes

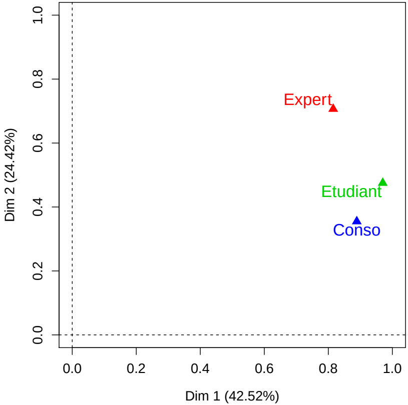
$\Rightarrow$ Utilisation de la mesure $\mathcal{L}_g$ pour représenter les groupes
Le groupe j a pour coordonnées $\mathcal{L}_g(K_j, v_1)$ et $\mathcal{L}_g(K_j, v_2)$



- 1ère dimension commune à tous les groupes
- 2ème dimension due au groupe Expert
- 2 groupes sont proches quand ils induisent la même structure

Représentation des groupes

⇒ Utilisation de la mesure $\mathcal{L}_g$ pour représenter les groupes
Le groupe j a pour coordonnées $\mathcal{L}_g(K_j, v_1)$ et $\mathcal{L}_g(K_j, v_2)$



- 1ère dimension commune à tous les groupes
- 2ème dimension due au groupe Expert
- 2 groupes sont proches quand ils induisent la même structure

⇒ Ce graphe fournit une comparaison synthétique des groupes
⇒ Les positions relatives des individus sont-elles similaires d'un groupe à l'autre ?

Mesures de similarité entre groupes

- Coefficient $\mathcal{L}_g$ mesure de liaison entre 2 groupes de variables :

$$\mathcal{L}_g(K_j, K_m) = \sum_{k \in K_j} \sum_{l \in K_m} cov^2 \left(\frac{x_{.k}}{\sqrt{\lambda_1^j}}, \frac{x_{.l}}{\sqrt{\lambda_1^m}} \right)$$

- Coefficient $\mathcal{L}_g$ comme un indice de dimensionalité d'un groupe

$$\mathcal{L}_g(K_j, K_j) = \frac{\sum_{k=1}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2} = 1 + \frac{\sum_{k=2}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2}$$

- $RV(K_j, K_m) = \dfrac{\mathcal{L}_g(K_j, K_m)}{\sqrt{\mathcal{L}_g(K_j, K_j)}\sqrt{\mathcal{L}_g(K_m, K_m)}}$ $\qquad 0 \leq RV \leq 1$

  $RV = 0$: toutes les variables de K_j et K_m sont non-corrélées
  $RV = 1$: les deux nuages de points sont homothétiques

Mesures de similarité entre groupes

```
> res$group$Lg
          Expert   Etudiant   Conso    MFA
Expert      1.45
Etudiant    1.17       1.29
Conso       0.94       1.04    1.25
MFA         1.33       1.31    1.21   1.44

> res$group$RV
          Expert   Etudiant   Conso    MFA
Expert      1.00
Etudiant    0.85       1.00
Conso       0.70       0.82    1.00
MFA         0.92       0.96    0.90   1.00
```

- Les Experts donnent une description plus riches ($\mathcal{L}_g$ supérieur)
- Les groupes Etudiant et Expert sont liés (RV $=$ 0.85)
- Le groupe Etudiant est le plus proche de la configuration commune (RV $=$ 0.96)

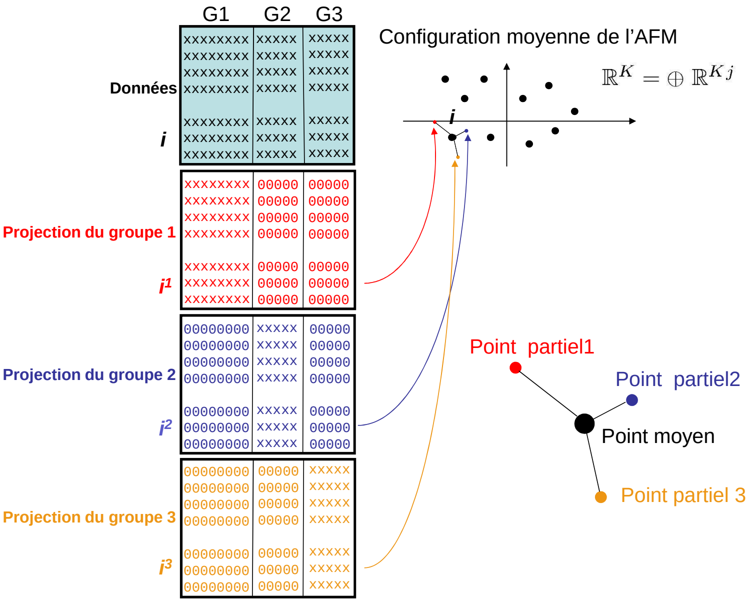
Représentation des points partiels

⇒ Comparaison des groupes à partir des individus

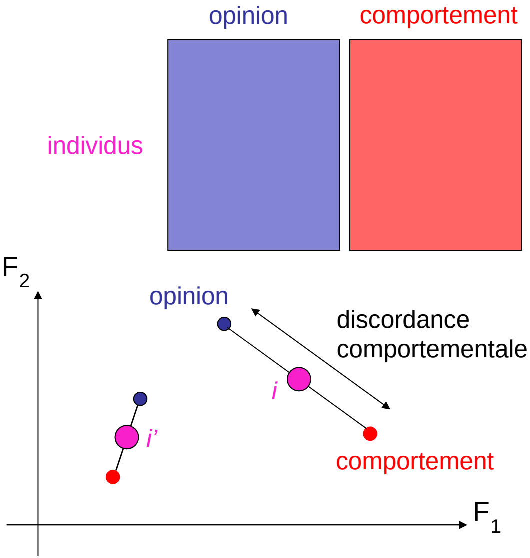
⇒ Comparaison des typologies fournies par chaque groupe dans un espace commun

⇒ Y a-t-il a des individus très particuliers pour certains groupes de variables ?

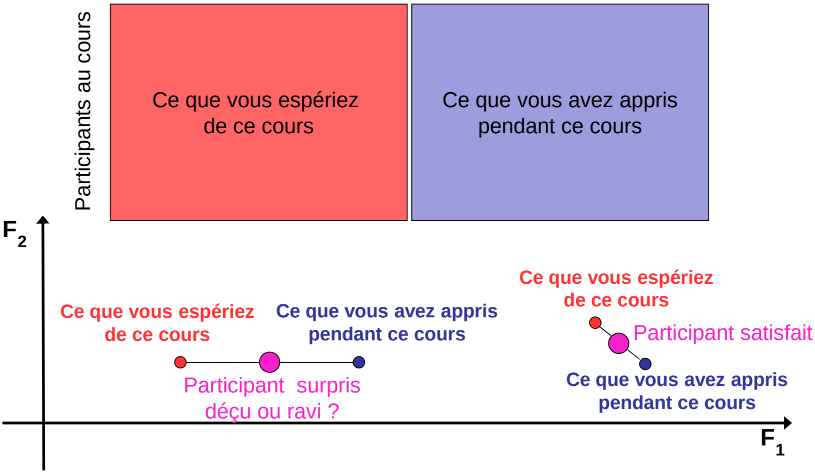
Projection des points partiels



Points partiels



Points partiels



Relations de transition

Les relations de transition s'appliquent pour les points moyens

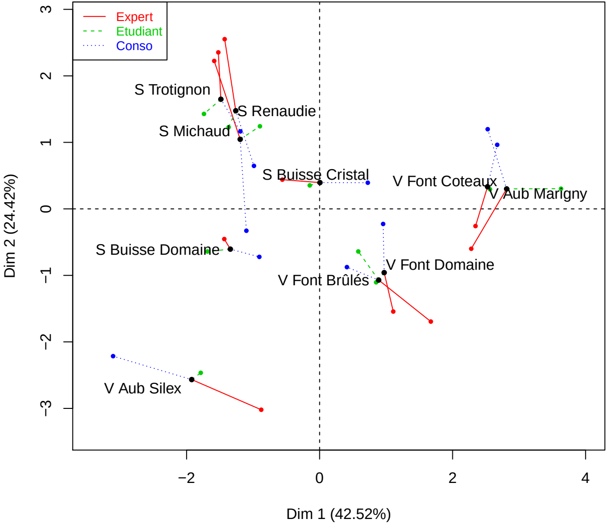
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \left(\frac{1}{\lambda_1^j} \sum_{k=1}^{K_j} x_{ik} G_s(k) \right)$$

et les points partiels

$$F_s(i^j) = J \times \frac{1}{\sqrt{\lambda_s}} \frac{1}{\lambda_1^j} \sum_{k=1}^{K_j} x_{ik} G_s(k)$$

$\Rightarrow$ La représentation superposée avec points moyens et points partiels peut être interprétée dans un cadre unique

Représentation des points partiels



- Point partiel = représentation d'un individu vu par un groupe
- Un individu est au barycentre de ses points partiels

Ratio d'inertie

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(F_{is})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s} - F_{is})^2$$

inertie totale = inertie inter individus + inertia intra individu

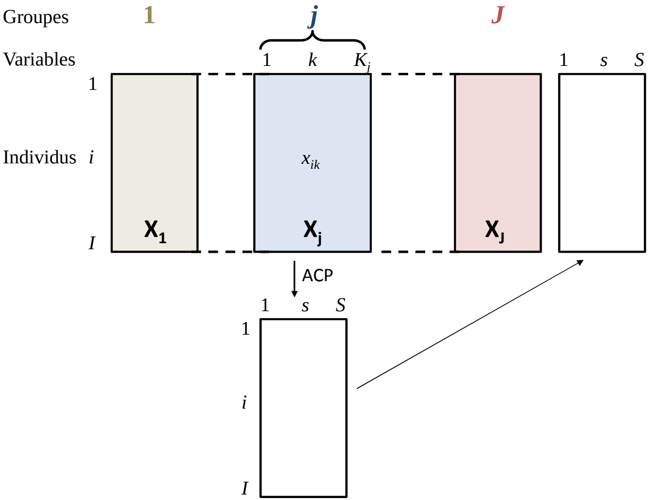
$$\frac{\text{Inertie inter sur l'axe } s}{\text{Inertie totale sur l'axe } s} = \frac{J\ \sum_{i=1}^{I}(F_{is})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s})^2}$$

```
> res$inertia.ratio
Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
 0.93    0.82    0.78    0.54    0.53
```

- Sur la première dimension, les coordonnées des points partiels sont proches (0.93 proche de 1)
- L'inertie intra d'un axe peut être décomposée par individu

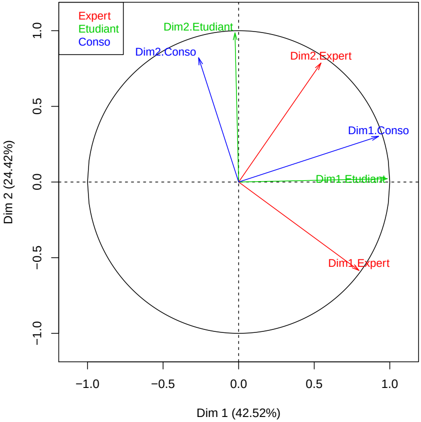
Relation avec les facteurs des analyses séparées

Les analyses séparées donnent-elles des résultats comparables à ceux de l'AFM globale ?



Relation avec les facteurs des analyses séparées

⇒ Les composantes principales des analyses séparées sont projetées en supplémentaires



- Les dimensions de l'ACP sur les données étudiant coïncident avec les dimensions de l'AFM
- Les deux premières dimensions de chaque groupe sont bien projetées

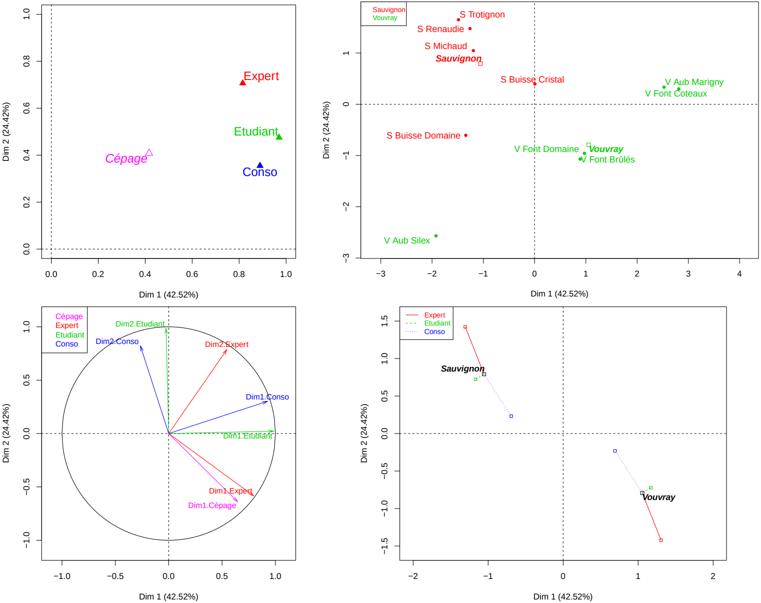
Plan

Données qualitatives

- Equilibrer les groupes de variables dans une analyse globale
- Représentation usuelle pour le traitement de données qualitatives (individus et modalités)
- Représentations spécifiques (graphe des groupes, représentation superposée, représentation des axes partiels des analyses séparées)

$\Rightarrow$ Même démarche en remplaçant ACP par ACM

Données qualitatives



Données mixtes

$\Rightarrow$ Groupes composés de variables quantitatives et groupes composés de variables qualitatives

L'AFM fonctionne "localement" comme :

- une ACP pour les variables quantitatives
- une ACM pour les variables qualitatives

La pondération de l'AFM permet d'analyser les deux types de variables ensemble

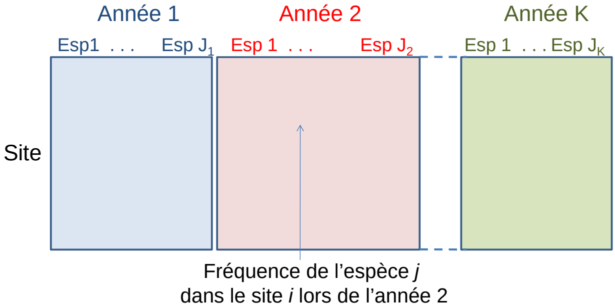
Cas particulier : si chaque groupe est composé d'une seule variable
$\Longrightarrow$ **Analyse Factorielle de Données Mixtes** (AFDM)

L'AFM sur tableaux de contingence

L'AFM a été étendue aux tableaux de contingence : AFMTC
Les tableaux doivent avoir une même dimension en commun

Exemples

- enquête dans plusieurs pays (CSP $\times$ questions par pays)
- écologie : sites $\times$ espèces par année



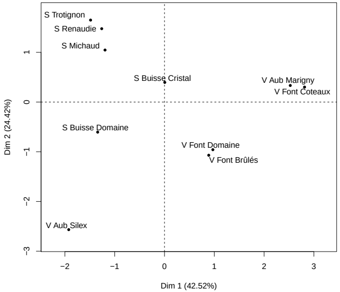
Représentation d'information supplémentaire

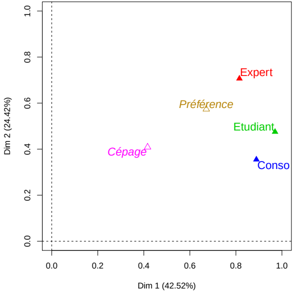
	Expert (27)	Conso (15)	Etudiant (15)	Appréciation (60)	Cépage (1)
Vin 1					
Vin 2					
…					
Vin 10					

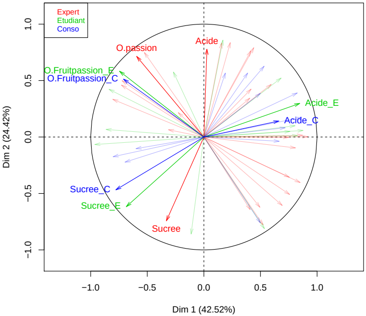
Questions :

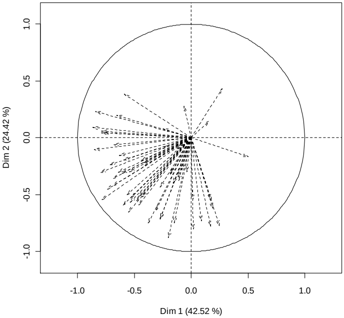
- Les préférences sont-elles liées à la description sensorielle ?
- La variable cépage explique-t-elle des différences sensorielles ?

Représentation d'un groupe supplémentaire quantitatif









Indicateurs : contribution et qualité de représentation

- Individus et variables : mêmes calculs qu'en ACP
- Contribution du groupe k à la construction de l'axe s :

$$Ctr_s(k) = \frac{F_{ks}}{\sum_{k=1}^{K} F_{ks}} \ (\times 100)$$

```
> res$group$contrib
         Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
Expert   30.49  45.99  33.68  44.59  40.60
Etudiant 36.27  30.92  35.07   9.20  14.72
Conso    33.24  23.09  31.25  46.20  44.68
```

- Qualité de représentation du groupe k sur un sous-espace : $\cos^2$ entre le point k et son projeté

```
> res$group$cos2
         Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
Expert    0.46   0.34   0.03   0.03   0.01
Etudiant  0.73   0.17   0.03   0.00   0.00
Conso     0.63   0.10   0.03   0.03   0.02
```

Description des dimensions

Par des variables quantitatives :

- corrélation entre chaque variable et la composante principale de rang s est calculée
- les coefficients de corrélation sont triés et les coefficients significatifs sont conservés

```
> dimdesc(res)
           $Dim.1$quanti                                   $Dim.2$quanti
                corr p.value                                    corr p.value
O.vanille       0.92 1.8e-04    O.Av.Intensite_C            0.86  0.0015
Amere_C         0.88 9.0e-04    Int.attaque                 0.84  0.0026
O.boisee        0.87 1.0e-03    Expression                  0.83  0.0028
G.Intensite_E   0.86 1.4e-03    Int.av.agitation            0.79  0.0064
Nuance.couleur  0.85 1.8e-03    Acide                       0.78  0.0081
Acide_E         0.85 2.0e-03    Int.ap.agitation            0.76  0.0110
   ...           ...     ...         ...                     ...     ...
Equilibre_C    -0.84 2.5e-03    Typicite.olf.chenin        -0.78  0.0081
O.Typicite_C   -0.86 1.3e-03    O.Alcool_C                 -0.81  0.0044
G.Typicite_C   -0.96 7.7e-06    O.Vegetale_C               -0.86  0.0014
```

Description des dimensions

Par des variables qualitatives :

- construire une analyse de variance avec les coordonnées des individus ($F_{.s}$) expliquées par la variable qualitative
  - un test F par variable
  - pour chaque catégorie, un test de Student (t-test)

```
> dimdesc(res)
$Dim.1$quali
              R2     p.value
cepage 0.4162427 0.04396733

$Dim.1$category
           Estimate     p.value
Vouvray    1.055053 0.04396733
Sauvignon -1.055053 0.04396733
```

```
$Dim.2$quali
              R2     p.value
cepage 0.4084123 0.04667455

$Dim.2$category
            Estimate     p.value
Sauvignon  0.7920973 0.04667455
Vouvray   -0.7920973 0.04667455
```

Mise en œuvre d'une AFM

1. Définir la structure du jeu de données (la composition des groupes)
2. Définir les groupes actifs et les éléments supplémentaires
3. Réduire ou non les variables ?
4. Réaliser l'AFM
5. Choisir le nombre de dimensions à interpréter
6. Interprétation simultanée du graphe des individus et des variables
7. Etude des groupes
8. Analyses partielles
9. Utilisation d'indicateurs pour enrichir l'interprétation

Fonction `MFA` du package `FactoMineR`

Conclusion

- AFM : une méthode multi-tableaux pour les variables quantitatives, qualitatives et les tableaux de fréquences
- L'AFM équilibre l'influence de chaque tableau
- Représentation l'information apportée par tous les tableaux dans un référentiel commun

- Sorties classiques (individus, variables)
- Sorties spécifiques (groupes, analyses séparées, points partiels)

Bibliographie

- Escofier, B. & Pagès, J. (2008). *Analyses Factorielles Simples et Multiples : Objectifs, Méthodes et Interprétation*. Dunod, 4e édition.
- Pagès, J. (2013). *Analyse factorielle multiple avec R*. EDP.