La catégorisation

François Husson

Laboratoire de mathématiques appliquées - Agrocampus Rennes

husson@agrocampus-ouest.fr

Description du recueil par catégorisation

Historique :

- proposée en 1970 par des psychologues
- mis en œuvre pour la première fois en sensoriel en 1989 par Lawless

Deux étapes :

1. regroupement des produits en fonction de leur ressemblance globale
2. description de chaque groupe de produits par des mots

Mise en place de la dégustation

- Chaque juge est dans un box individuel
- Tous les produits sont apportés simultanément
- Les produits sont codés comme pour un recueil classique
- Il est possible de revenir sur un produit
- Le juge énumère les groupes de produits et les mots associés au groupe sur une feuille blanche (avec le numéro du juge)

```
Numéro du juge : 18
Groupe 1 : 617, 172, 621 : fruité
Groupe 2 : 891, 268      : fort, entêtant
Groupe 3 : 145, 387, 433 : fleuri, fraîcheur
Groupe 4 : 925, 719, 546 : marine
```

- Bien vérifier que tous les produits apparaissent 1 fois et 1 seule

Intérêts du recueil par catégorisation

- Tâche de description facile
- Tâche de description rapide
- Ne nécessite pas d'entraînement
- Peut être effectuée par des consommateurs
- Etape préliminaire/complémentaire du profil sensoriel classique
- Permet l'obtention de descripteurs

Exemple : description de parfums

Les produits

Angel

Aromatics Elixir

Chanel n°5

Cinéma

Coco Mademoiselle

L'instant

Lolita Lempicka

Pleasures

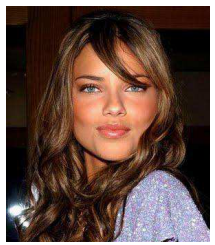
Pure Poison

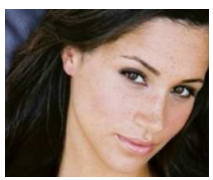
Shalimar

J'adore (ET)

J'adore (EP)

Les juges





Exemple : description de parfums

Etape 1 : constituer les groupes

Etape 2 : verbaliser chaque groupe

« oriental, Patchouli oil »

« gourmand, vanille»

« épicé, aldehyde »

« floral, vert »

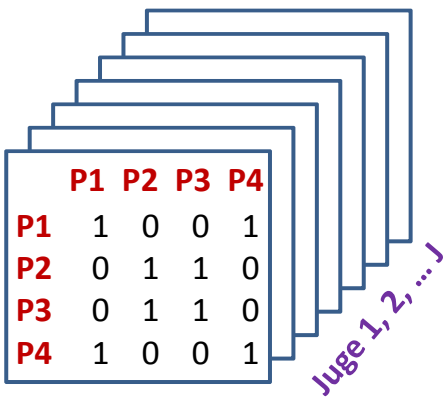
« orange »

«boisé »

Quel tableau de données analyser ?

	P1	P2	P3	P4
P1	80	50	22	15
P2	50	80	40	60
P3	58	40	80	10
P4	65	20	70	80

- le tableau de cooccurences $\Longrightarrow$ MDS
  rq : ni information individuelle ni information sur les mots

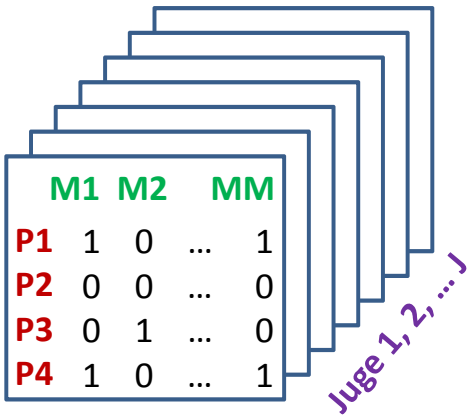


- les tableaux individuels de cooccurences (tableaux de 0 et de 1) $\Longrightarrow$ distatis
  rq : pas d'information sur les mots

	M1	M2	...	MM
P1	20	15	...	3
P2	17	21	...	5
P3	5	2	...	19
P4	3	2	...	24

- le tableau produit x mot $\Longrightarrow$ AFC
  rq : ni information sur les associations de produits ni information individuelle

Quel tableau de données analyser ?



- les tableaux individuels produit x mot
  $\Longrightarrow$ AFMTC
  rq : pas d'information sur les associations

	J1	J2	...	JJ
P1	G1	G3	...	G6
P2	G1	G4	...	G6
P3	G2	G4	...	G7
P4	G2	G5	...	G7

- le tableau produit x juge avec un numéro de groupe dans chaque cellule
  $\Longrightarrow$ ACM
  rq : pas d'information sur les mots

	J1	J2	...	JJ
P1	M1	M3	...	M6
P2	M1	M4	...	M6
P3	M2	M4	...	M7
P4	M2	M5	...	M7

- le tableau produit x juge avec les mots dans chaque cellule $\Longrightarrow$ ACM

Ces méthodes sont comparées dans la thèse de Marine Cadoret

Exemple : description de parfums

Codage des données :

		
	vanille	orange
	boisé	vieux
	épicé	vieux
	orange	fort
	orange	fort
	orange	fort
	vanille	fort
	floral	fort
	boisé	vieux
	boisé	vieux
	floral	orange
	floral	orange

vanille	boisé	épicé	oriental	floral	orange	vieux	fort	orange
1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	0
0	0	0	0	0	1	0	1	0
0	0	0	0	0	1	0	1	0
0	0	0	1	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	0	0	1	0	0	1	0
0	1	0	0	0	0	1	0	0
0	1	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	1
0	0	0	0	1	0	0	0	1

L'ACM : distance entre produits

$$d^2_{i,i'} = \frac{I}{J} \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k}$$

	vanille	boisé	épicé	oriental	floral	orange	vieux	fort	orange
	1	0	0	0	0	0	0	0	1
	0	1	0	0	0	0	1	0	0
	0	0	1	0	0	0	1	0	0
	0	0	0	0	0	1	0	1	0
	0	0	0	0	0	1	0	1	0
	0	0	0	1	0	0	0	1	0
	1	0	0	0	0	0	0	1	0
	0	0	0	0	1	0	0	1	0
	0	1	0	0	0	0	1	0	0
	0	1	0	0	0	0	1	0	0
	0	0	0	0	1	0	0	0	1
	0	0	0	0	1	0	0	0	1
	2	3	1	1	3	2	4	5	3

L'ACM : distance entre produits

$$d^2_{i,i'} = \frac{I}{J} \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k} = 0$$

	vanille	boisé	épicé	oriental	floral	orange	vieux	fort	orange
	1	0	0	0	0	0	0	0	1
	0	1	0	0	0	0	1	0	0
	0	0	1	0	0	0	1	0	0
	0	0	0	0	0	1	0	1	0
	0	0	0	0	0	1	0	1	0
	0	0	0	1	0	0	0	1	0
	1	0	0	0	0	0	0	1	0
	0	0	0	0	1	0	0	1	0
	0	1	0	0	0	0	1	0	0
	0	1	0	0	0	0	1	0	0
	0	0	0	0	1	0	0	0	1
	0	0	0	0	1	0	0	0	1
	2	3	1	1	3	2	4	5	3

- $d_{i,i'} = 0$ si les produits i et i' sont systématiquement ensemble
- i et i' sont d'autant plus proches qu'ils ont été mis ensemble par beaucoup de juges

L'ACM : distance entre produits

$$d^2_{i,i'} = \frac{I}{J}\sum_{k=1}^{K}\frac{(x_{ik}-x_{i'k})^2}{I_k} = \frac{I}{J}\left(\frac{1}{3}+\frac{1}{2}\right) = 0.83 \times \frac{I}{J}$$

	vanille	boisé	épicé	oriental	floral	orange	vieux	fort	orange
	1	0	0	0	0	0	0	0	1
	0	1	0	0	0	0	1	0	0
	0	0	1	0	0	0	1	0	0
	0	0	0	0	0	1	0	1	0
	0	0	0	0	0	1	0	1	0
	0	0	0	1	0	0	0	1	0
	1	0	0	0	0	0	0	1	0
	0	0	0	0	1	0	0	1	0
	0	1	0	0	0	0	1	0	0
	0	1	0	0	0	0	1	0	0
	0	0	0	0	1	0	0	0	1
	0	0	0	0	1	0	0	0	1
	2	3	1	1	3	2	4	5	3

- $d_{i,i'} = 0$ si les produits i et i' sont systématiquement ensemble
- i et i' sont d'autant plus proches qu'ils ont été mis ensemble par beaucoup de juges

L'ACM : distance entre produits

$$d^2_{i,i'} = \frac{I}{J}\sum_{k=1}^{K}\frac{(x_{ik}-x_{i'k})^2}{I_k} = \frac{I}{J}\left(\frac{1}{1}+\frac{1}{2}\right) = 1.5\times\frac{I}{J}$$

	vanille	boisé	épicé	oriental	floral	orange	vieux	fort	orange
	1	0	0	0	0	0	0	0	1
	0	1	0	0	0	0	1	0	0
	0	0	1	0	0	0	1	0	0
	0	0	0	0	0	1	0	1	0
	0	0	0	0	0	1	0	1	0
	0	0	0	1	0	0	0	1	0
	1	0	0	0	0	0	0	1	0
	0	0	0	0	1	0	0	1	0
	0	1	0	0	0	0	1	0	0
	0	1	0	0	0	0	1	0	0
	0	0	0	0	1	0	0	0	1
	0	0	0	0	1	0	0	0	1
	2	3	1	1	3	2	4	5	3

- $d_{i,i'} = 0$ si les produits i et i' sont systématiquement ensemble
- i et i' sont d'autant plus proches qu'ils ont été mis ensemble par beaucoup de juges
- la modalité k contribue de façon inversement proportionnelle à sa taille (un produit particulier est éloigné)

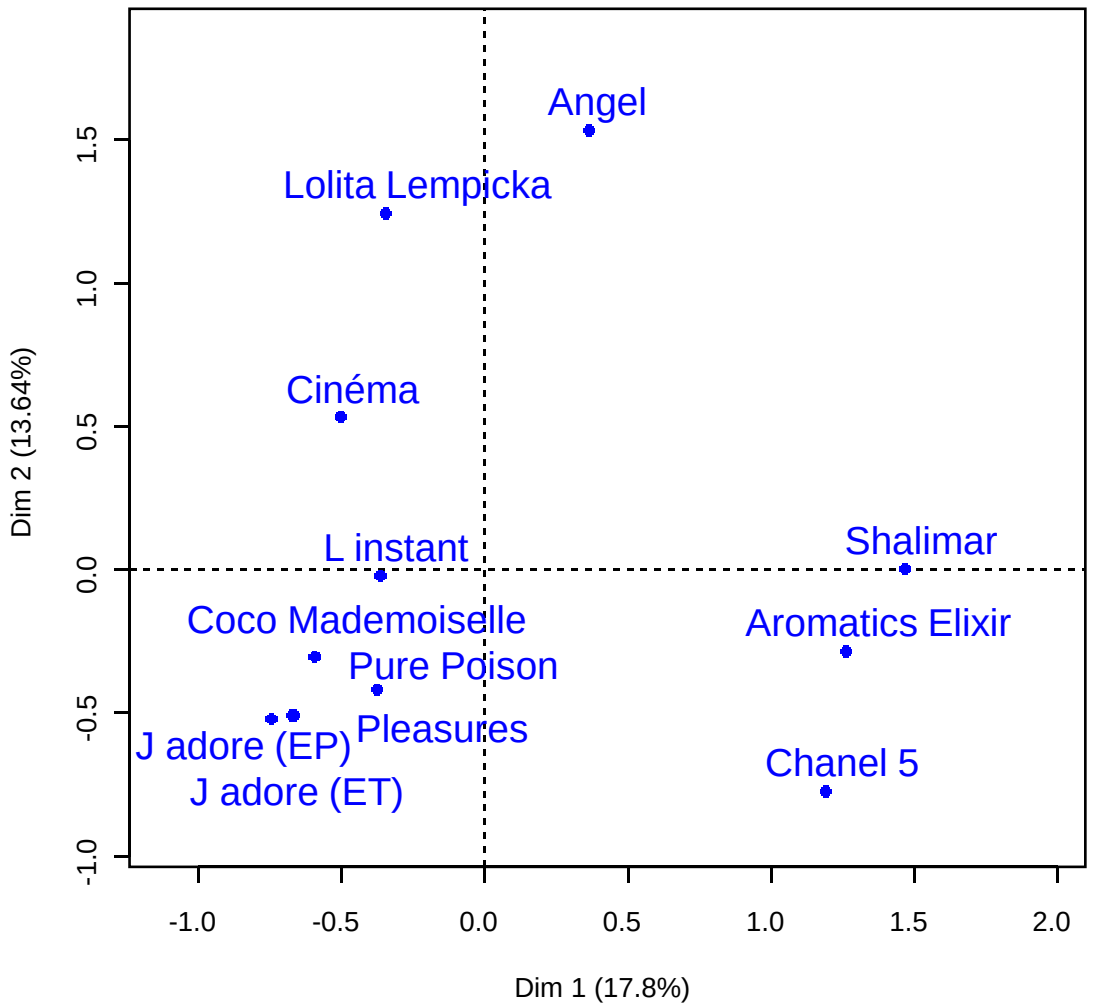
L'ACM : distance entre mots

$$d^2_{k,k'} = I \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2$$

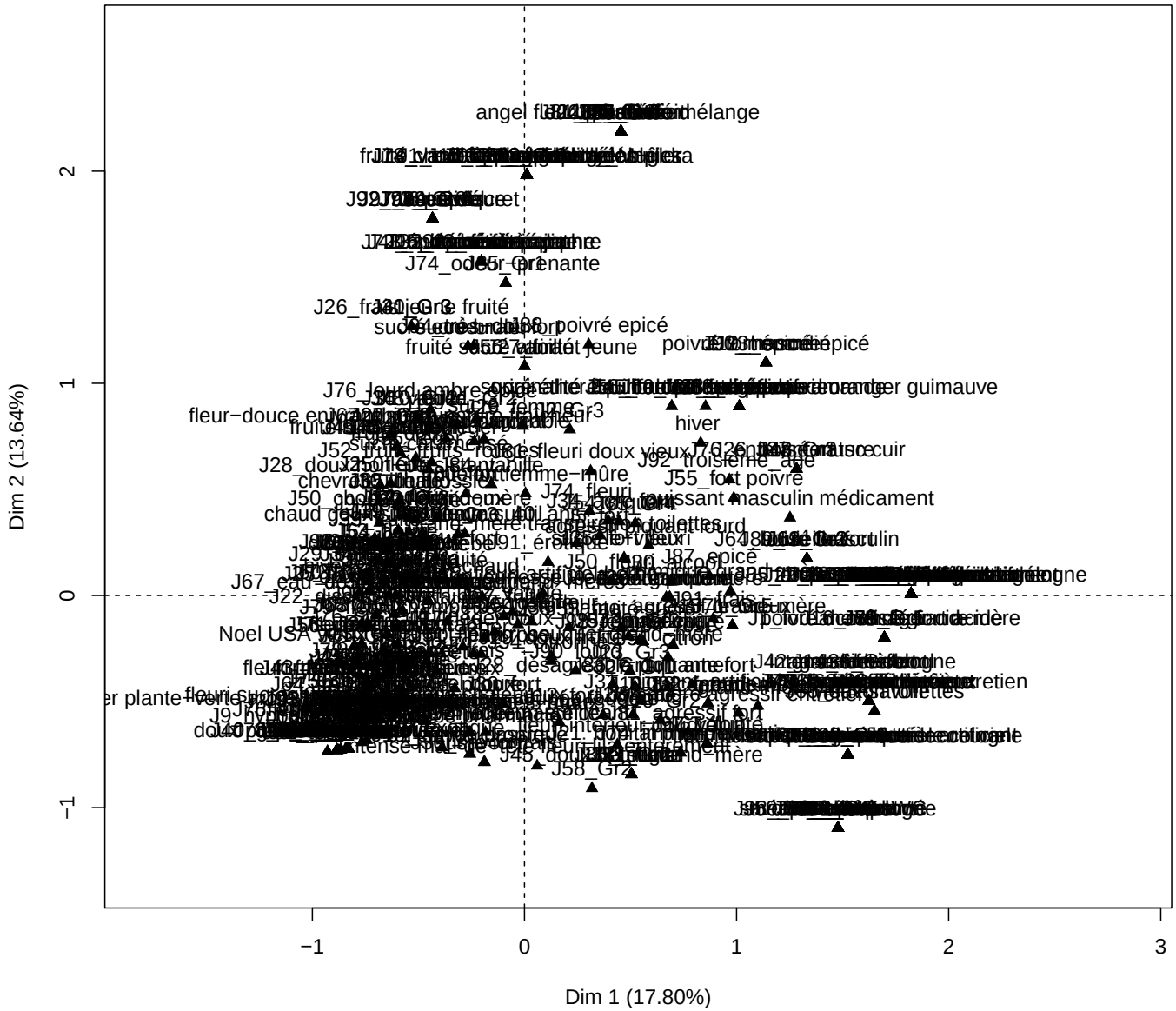
- Deux mots (deux modalités) sont d'autant plus éloignés qu'ils ont peu de parfums (d'individus) en commun : autrement dit, que le nombre de parfums décrits par le mot k et le mot k' est petit
- Deux mots sont superposés s'ils caractérisent exactement les mêmes parfums

Représentation des parfums

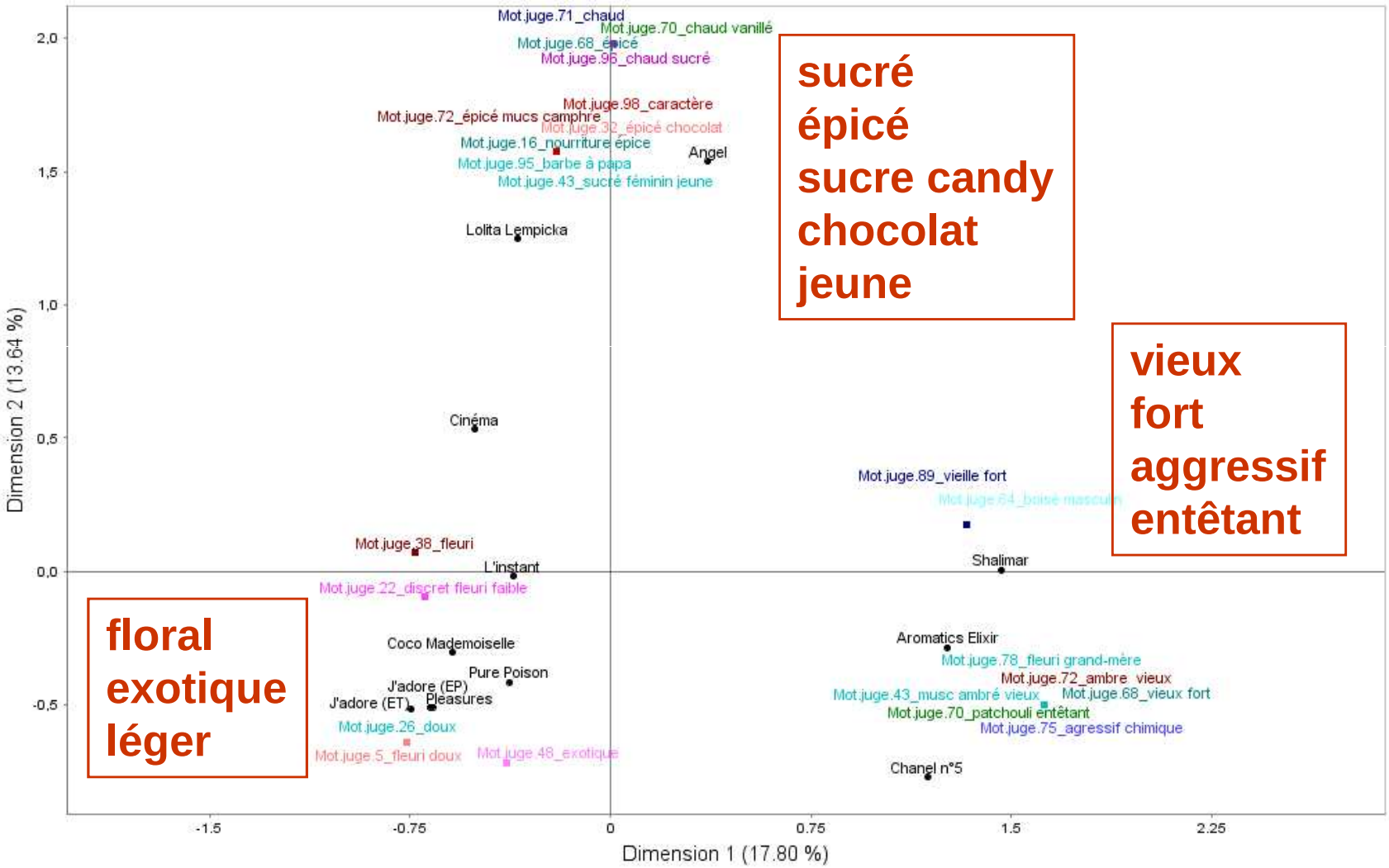
```
res.fast <- fast(parfums, sep.words=" ")
```



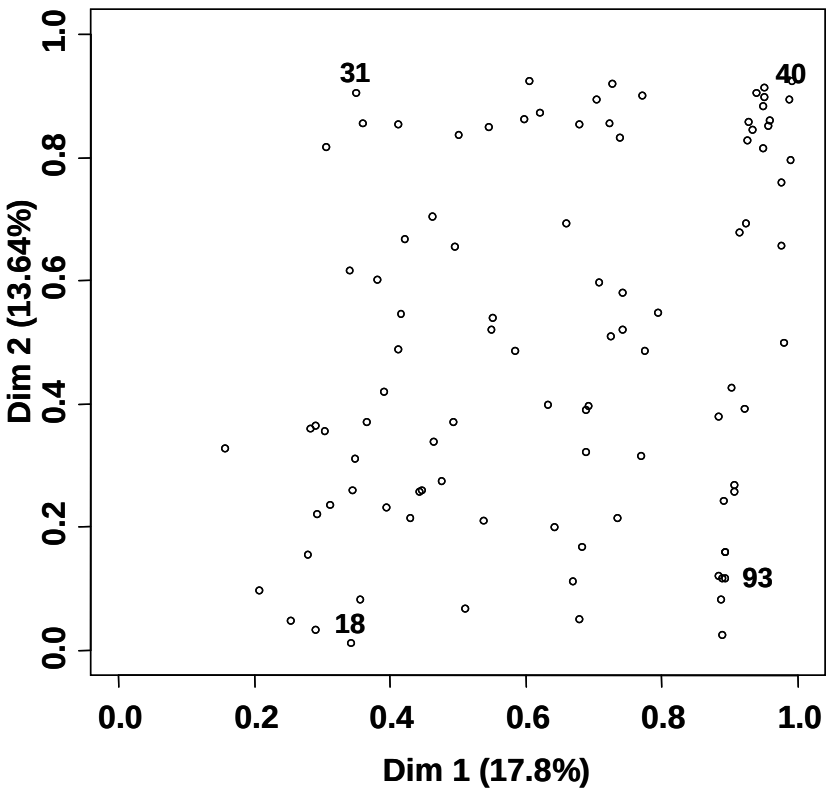
Représentation des parfums et des mots

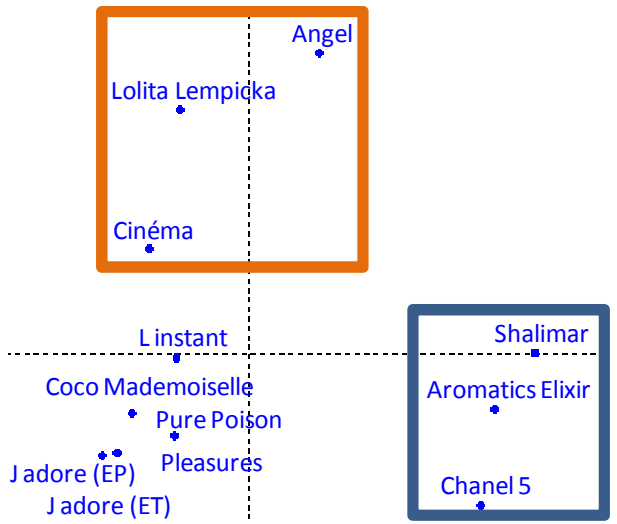


Représentation des parfums et des mots



Représentation des juges





	juge 18	juge 31	juge 40	juge 93
Shalimar	2	1	4	4
Aromatics Elixir	2	2	5	3
Chanel n°5	3	4	5	3
Coco Mademoiselle	3	1	2	2
J'adore (EP)	1	1	1	1
J'adore (ET)	3	1	1	2
L'instant	2	1	2	1
Pleasures	3	1	1	1
Pure Poison	1	2	2	2
Angel	3	5	6	1
Cinéma	3	3	3	2
Lolita Lempicka	1	3	3	2

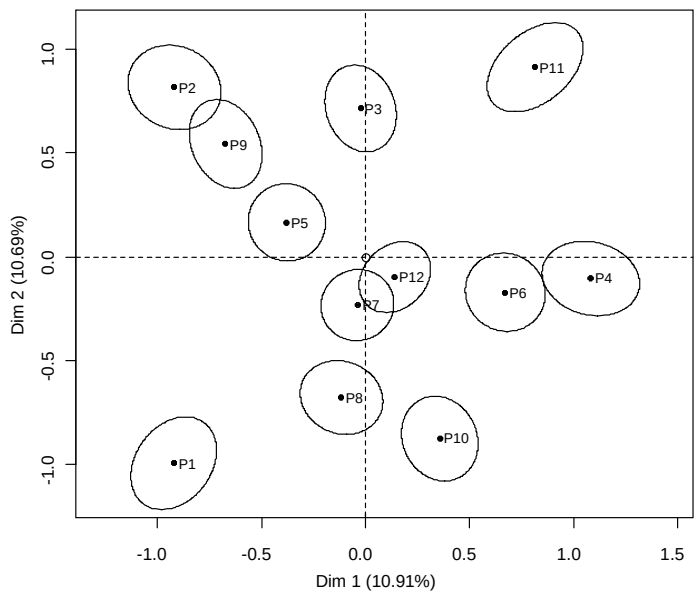
Une (mauvaise) idée pour construire des ellipses de confiance

Principe de construction :

1. Faire l'ACM
2. Utiliser la position des mots pour obtenir la position d'un produit vu par un juge
3. Construire l'ellipse de confiance à partir des J positions d'un produit

Evaluation de la méthode par perturbation du jeu de données :

- Pour chaque juge, intervertir au hasard les produits
- La structure globale du jeu de données est cassée, les produits ne sont plus différenciés par tous les juges (mais on conserve le nb de groupes par juge)

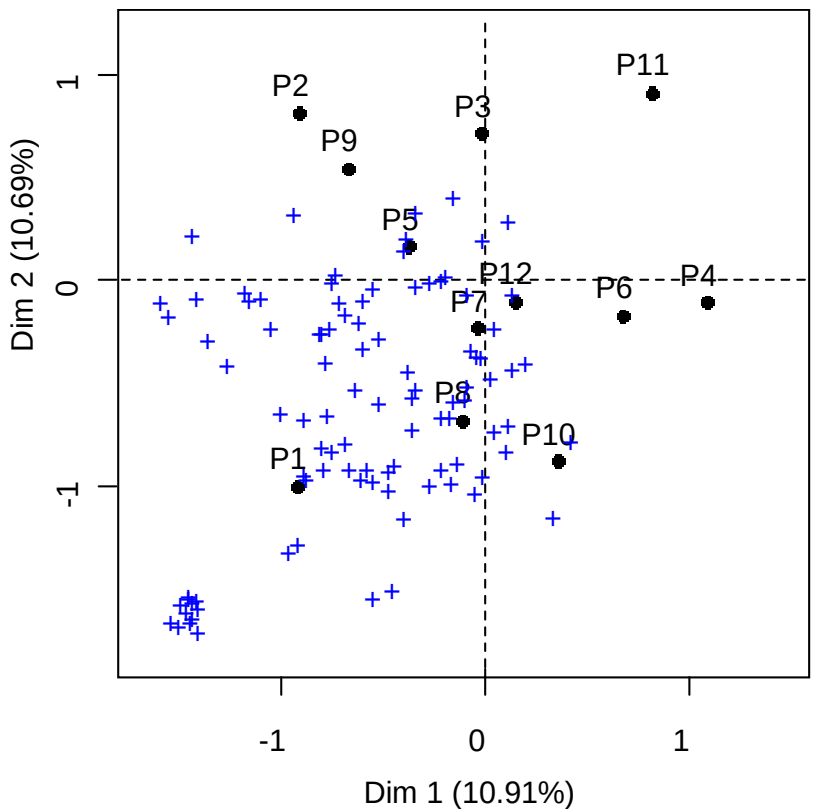


Problème : le graphe met en évidence des différences entre produits

Une (mauvaise) idée pour construire des ellipses de confiance

Pourquoi les ellipses sont-elles autant séparées sur un jeu de données non-structuré ?

- Projections du produit 1 vu par chaque juge sont dans une même région du graphe
- L'ellipse est petite car construite autour d'un centre de gravité de beaucoup de points



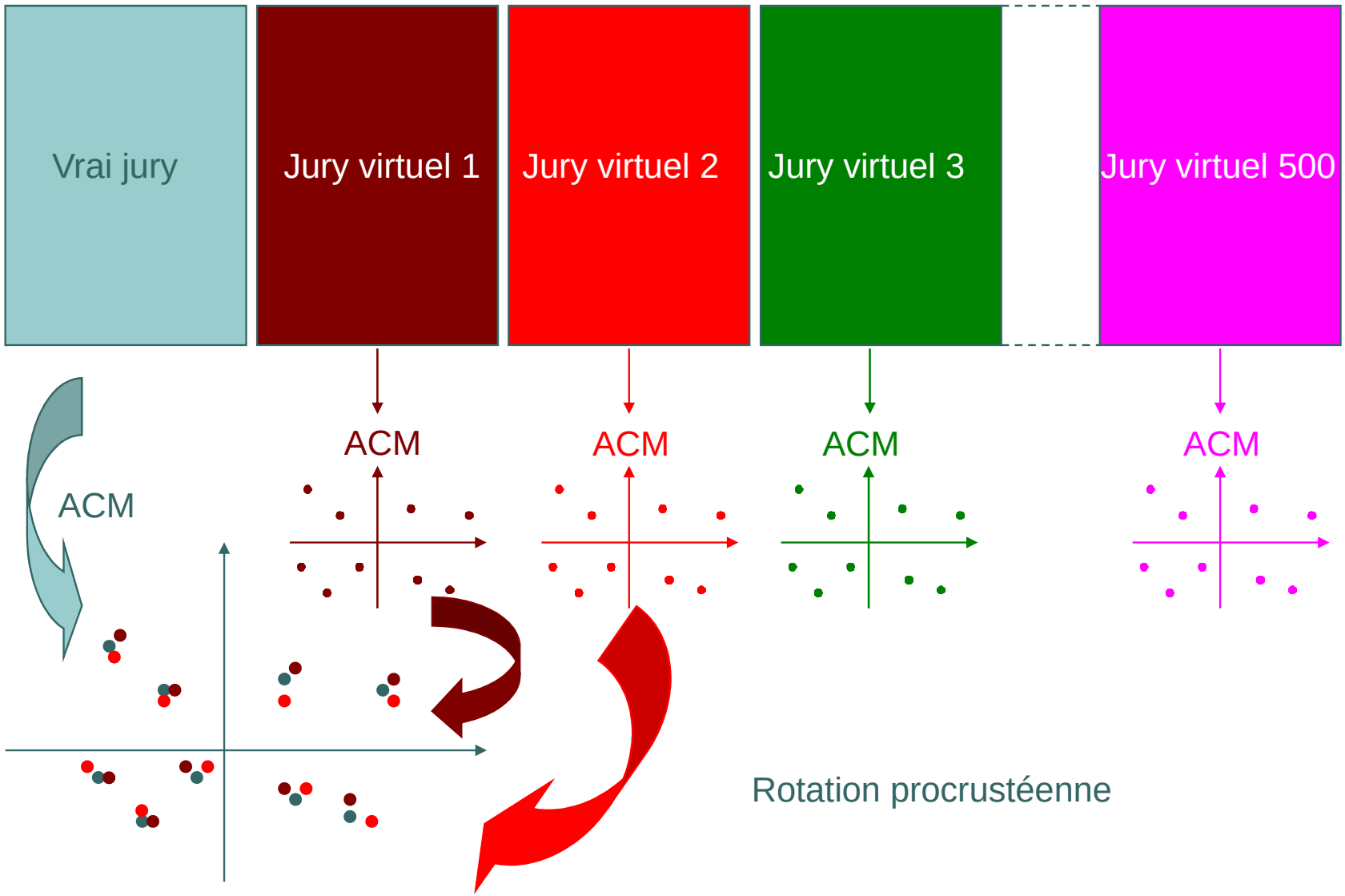
Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total

Le bootstrap total consiste à bootstraper les individus statistiques, refaire une analyse complète pour chaque réplication et enfin concaténer les résultats des échantillons bootstraps

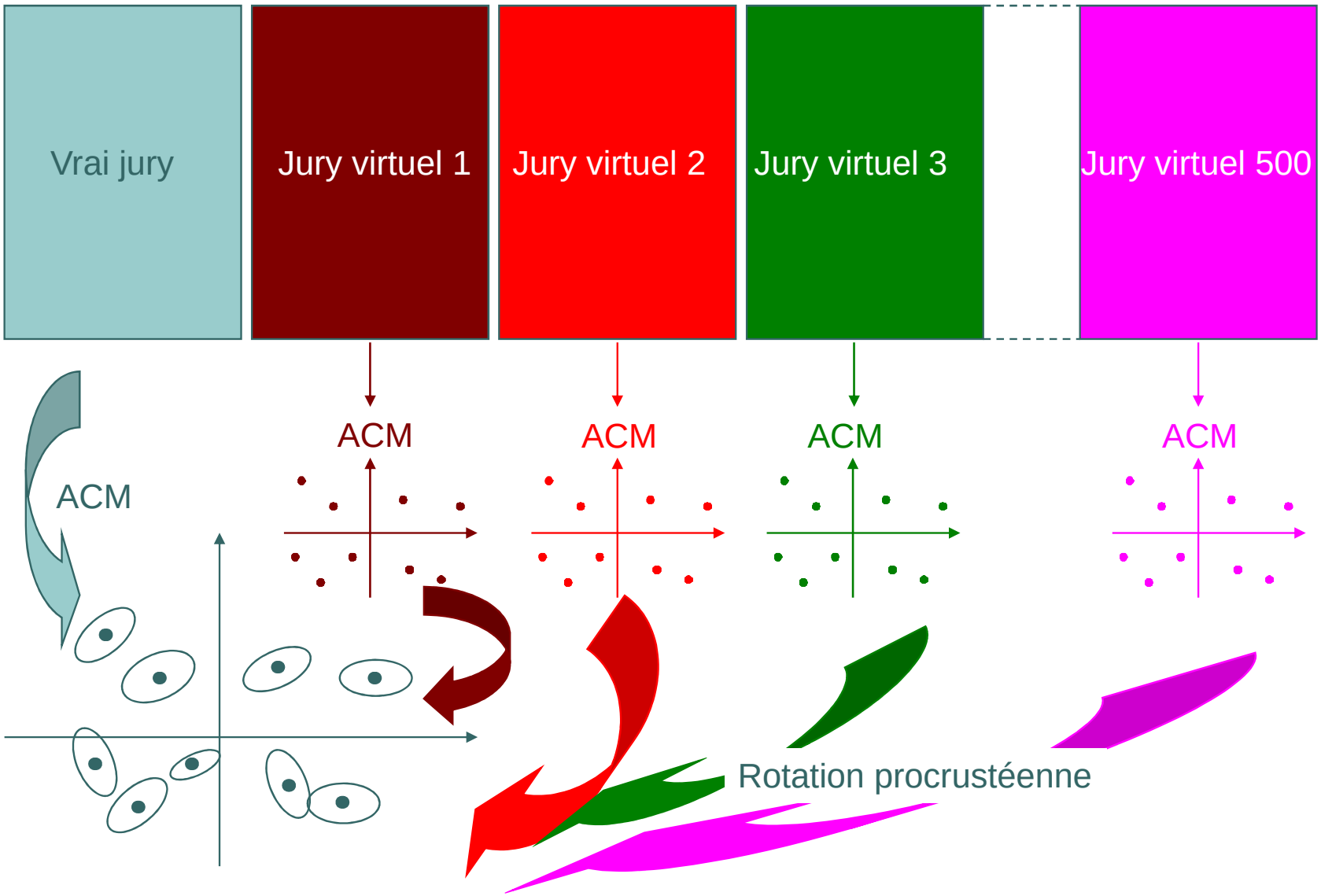
Description de l'algorithme en catégorisation :

1. Faire l'ACM sur les données du vrai jury
2. Répéter
   - Construire un jury virtuel en choisissant au hasard des juges dans le vrai jury
   - Faire l'ACM sur le jury virtuel
   - Faire une rotation procrustéenne du plan d'ACM obtenu par le jury virtuel sur le plan de l'ACM obtenu avec le vrai jury
3. Construire des ellipses de confiance autour de chaque produit à partir des positions de chaque jury virtuel

Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total

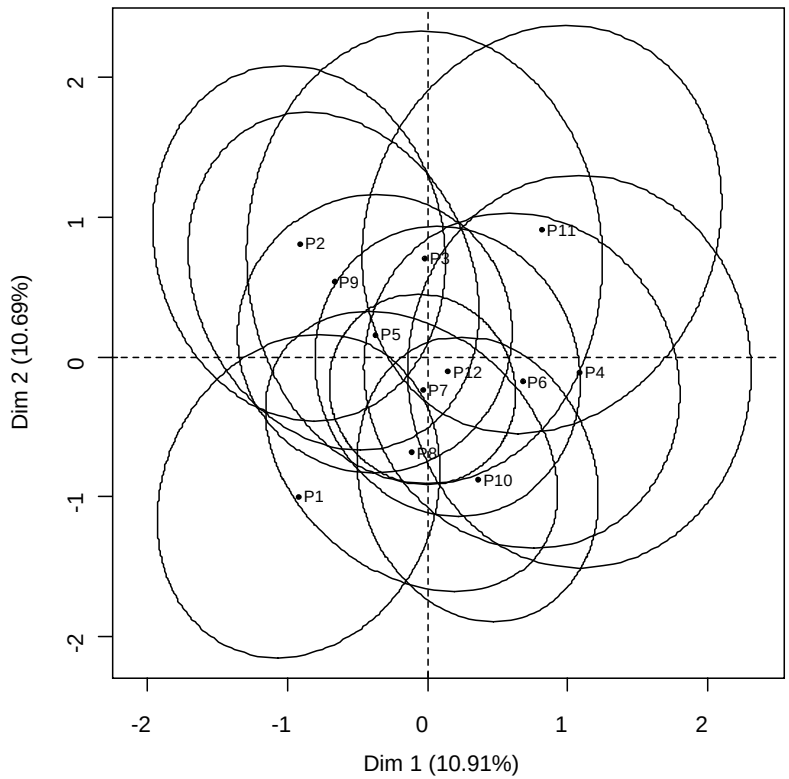


Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total



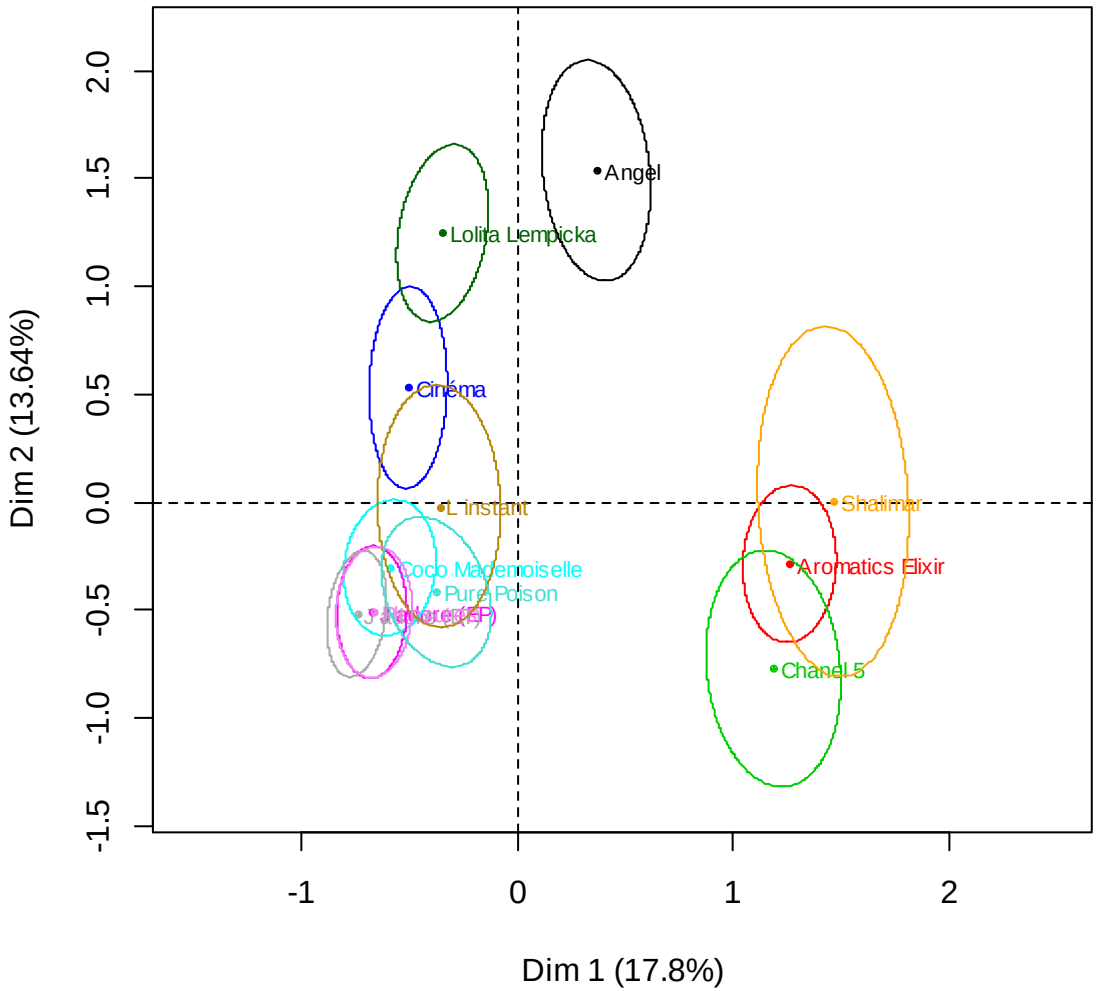
Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total

Evaluation de la méthode sur données non-structurées (perturbation aléatoire du jeu de données)



$\Longrightarrow$ Aucune mise en évidence de produits : résultat attendu pour données non-structurées

Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total



Une (bonne) idée pour construire des ellipses de confiance : le bootstrap total

Besoin de choisir le nombre de dimensions de l'ACM pour faire la rotation procrustéenne
Choix difficile du nombre de dimensions : souvent 2 dimensions

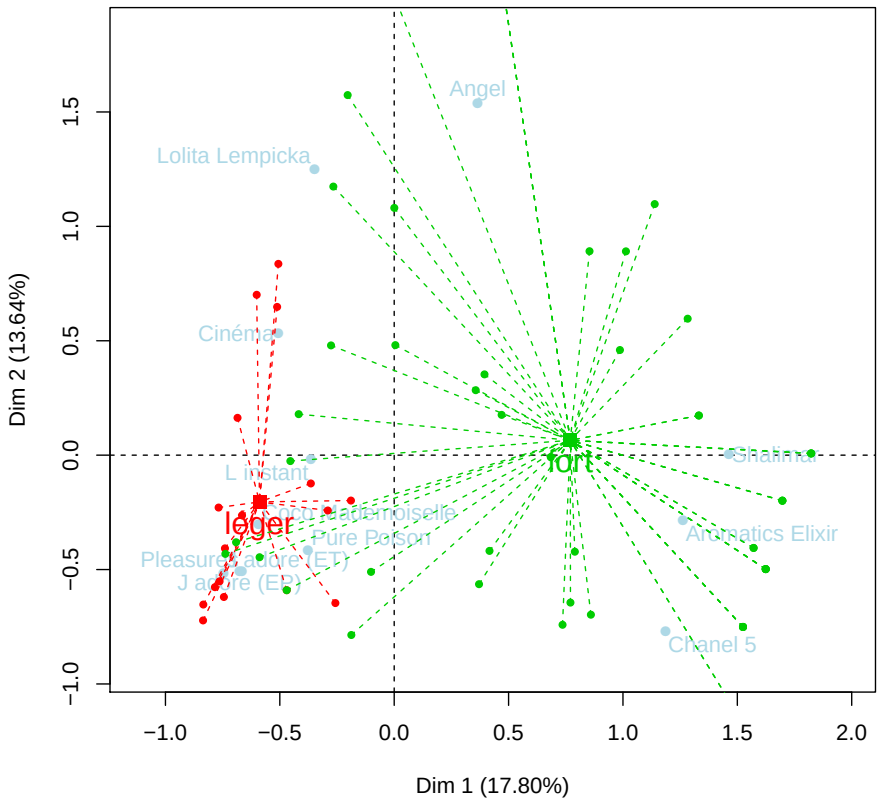
Utilisation possible de cet algorithme pour des données de catégorisation, napping, napping catégorisé, tri hiérarchique, profil flash, et pour les données de QDA.

Algorithme disponible dans la fonction `boot` de SensoMineR

Recherche automatique de mots consensuels

Un mot est consensuel s'il est utilisé de la même façon par les juges

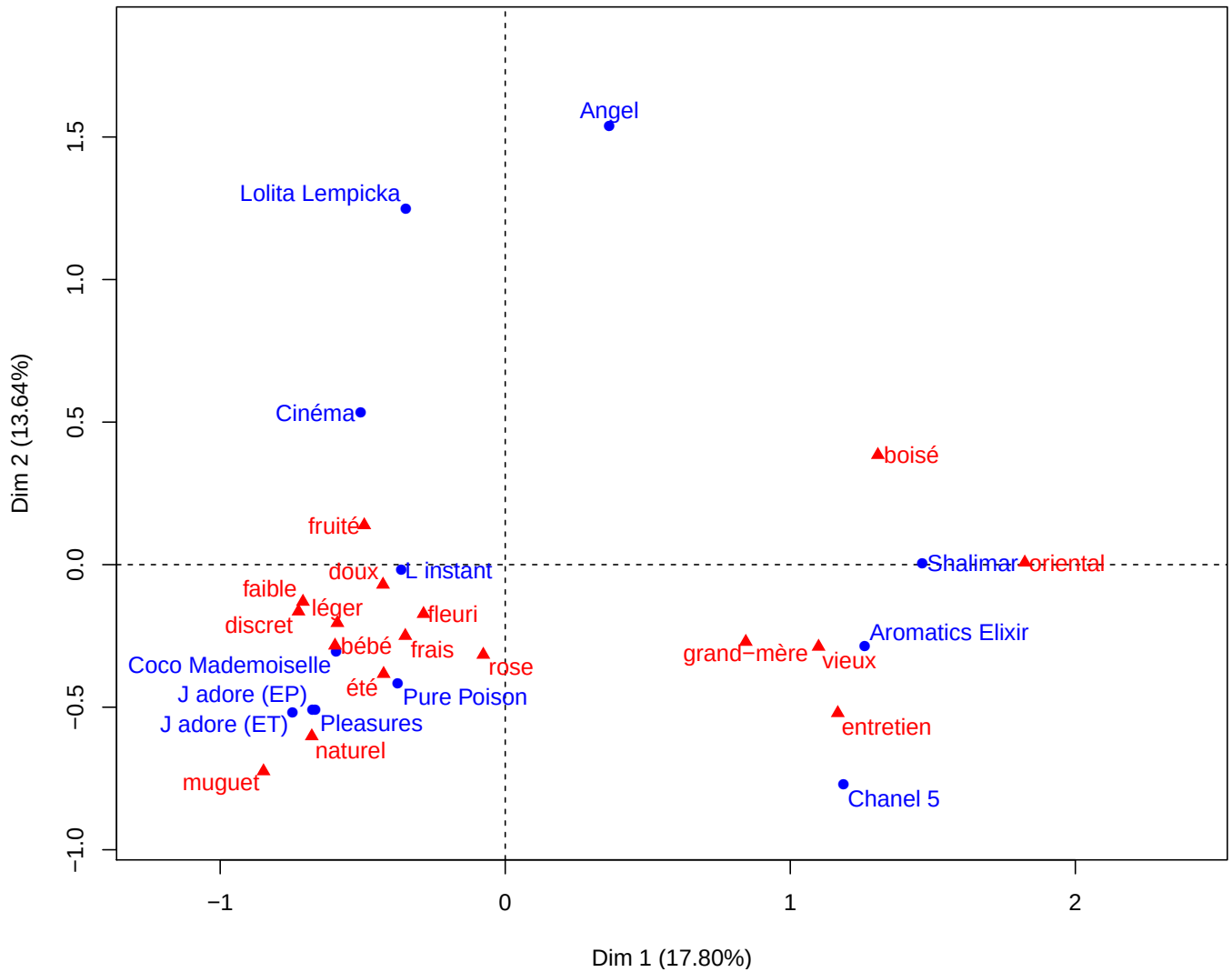
- Un mot cité par P juges a P positions
- Consensus sur un mot si la variabilité des P projections de ce mot est significativement plus petite que la variabilité des positions de P mots choisis au hasard



Quel intérêt de commenter un mot non-consensuel ?

Recherche automatique de mots consensuels

```
res.consensual <- ConsensualWords(res.fast, nbtimes=2, proba=0.05)
```



Analyse textuelle

Le mot *Vanillé* caractérise-t-il le parfum Angel ?

	Angel	Pas Angel	Total
Vanillé	5	8	13
Pas Vanillé	119	1414	1533
Total	124	1422	1546

Principe : une urne contient 1546 boules, sur 13 boules est écrit le mot *vanillé*, on tire 124 boules.

H_0 : la fréquence F du mot *Vanillé* suit une loi $\mathcal{H}(1546, 13, 124)$

Peut-on remettre en cause cette hypothèse ?
$\Longrightarrow$ 5 provient-il d'une loi hypergéométrique $\mathcal{H}(1546, 13, 124)$?

Angel

	Intern	% glob	% Intern freq	Glob freq	p.value	v.test
vanillé	4.032	0.841	5	13	0.005	2.829

$$\frac{5}{124} = 0.04032 \quad ; \quad \frac{13}{1546} = 0.00841 \quad ; \quad P\left[F \geq 5 \mid F \sim \mathcal{H}(1546, 13, 124)\right] = 0.005$$

$\Longrightarrow$ Rejet de H_0, le mot *Vanillé* est sur-employé pour Angel

Analyse textuelle

```
res.fast$textual
```

```
Angel
          Intern %  glob %  Intern freq  Glob freq  p.value  v.test
vanillé      4.032   0.841            5         13    0.005   2.829
épicé        4.839   1.488            6         23    0.015   2.426
sucré       12.097   6.598           15        102    0.026   2.225
fort        13.710   8.215           17        127    0.041   2.042

Chanel n°5
             Intern%  glob%  Intern freq  Glob freq  p.value  v.test
savon          7.752  1.423           10         22    0.000   4.515
toilettes      3.101  0.712            4         11    0.019   2.341
grand-mère     6.202  2.523            8         39    0.025   2.236
chimique       3.876  1.164            5         18    0.026   2.220
fort          13.953  8.215           18        127    0.029   2.183
vieux          3.876  1.229            5         19    0.033   2.126
```

Quelques références

- Cadoret M. (2010). Analyse factorielle multiple de données de catégorisation : application aux données sensorielles. *Thèse de doctorat*. `http://marine.cadl.free.fr/These_Marine_Cadoret.pdf`
- Cadoret M., Lê S. & Pagès J. (2009). A Factorial Approach for Sorting Task data (FAST). *Food Quality and Preference*. 20, 410–417.
- Cadoret M. & Husson F. (2013). Construction and evaluation of confidence ellipses applied at sensory data . *Food Quality and Preference*, 28, 106–115.
- Kostov B., Bécue-Bertaut M. & Husson F. (2014). An original methodology for the analysis and interpretation of word-count based methods : multiple factor analysis for contingency tables complemented by consensual words. *Food Quality and Preference*, 32, 35–40.

Les fonctions de SensoMineR :

```
fast              # fait l'ACM, l'analyse textuelle, les zones de confiance
boot              # construit les ellipses de confiance (appelée par fast)
ConsensualWords   # trouve les mots consensuels
```