

# Le Multidimensional Scaling et la cartographie des préférences

Gilbert Saporta

Conservatoire National des Arts et Métiers

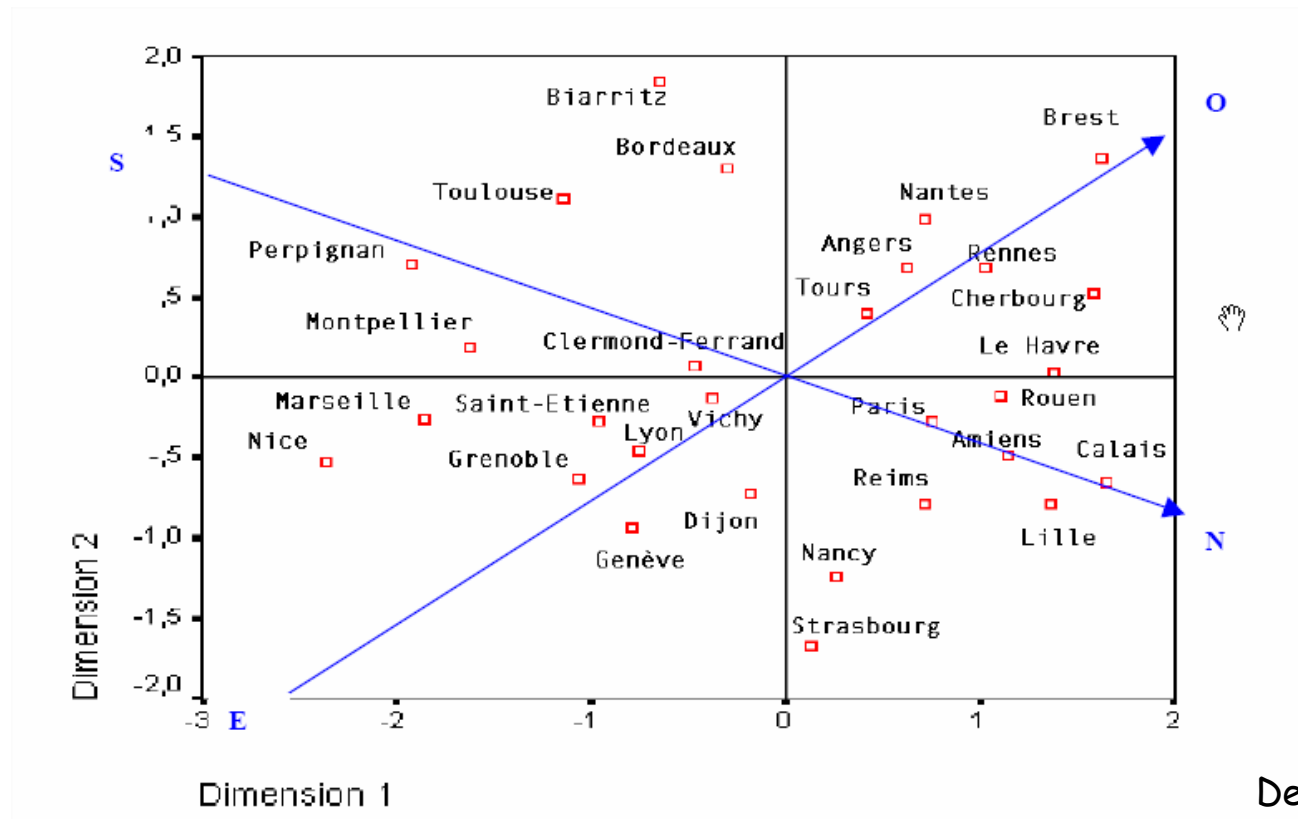
<http://cedric.cnam.fr/~saporta>

Janvier 2012

# Multidimensional scaling

- Egalement appelé « positionnement multidimensionnel », « analyse des proximités »
- Objectif: à partir d'un ou plusieurs tableaux de distances ou de dissimilarités entre  $n$  objets, reconstituer une image dans un espace euclidien

- Exemple: reconstituer une carte connaissant le tableau des distances entre n villes de France





# 1. Le cas « classique »: tableau de distances euclidiennes; *principal coordinate analysis*

- Axiomes:

$$d_{ij} \geq 0$$

$$d_{ij} = d_{ji}$$

$$d_{ij} \leq d_{ik} + d_{kj}$$

$$d_{ij} = \left( (\mathbf{e}_i - \mathbf{e}_j)' \mathbf{M} (\mathbf{e}_i - \mathbf{e}_j) \right)^{\frac{1}{2}}$$

- **Rappels d'ACP**

- Tableau de données  $X$

- Facteurs principaux  $Vu = \lambda u$        $nV = X'X$

- Composantes principales  $c = Xu$

$$1/n Wc = \lambda c$$

$W = XX'$  matrice  $n \times n$  des produits  
scalaires

- Si on trouve  $W$  à partir de la matrice des (carrés des) distances  $D$ , le problème est résolu

# La formule de Torgerson

On pose:

$$d_{i.}^2 = \frac{1}{n} \sum_{j=1}^n d_{ij}^2$$
$$d_{.j}^2 = \frac{1}{n} \sum_{i=1}^n d_{ij}^2$$
$$d_{..}^2 = \frac{1}{n} \sum_{j=1}^n d_{i.}^2 = \frac{1}{n} \sum_{j=1}^n d_{.j}^2$$

$d_{..}^2$  n'est autre que deux fois l'inertie I

$$w_{ij} = -\frac{1}{2} \left( d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2 \right)$$

- Démonstration
  - Formule du triangle

$$d_{ij}^2 = \|e_i\|^2 + \|e_j\|^2 - 2w_{ij}$$

$$w_{ij} = -\frac{1}{2} \left( d_{ij}^2 - \|e_i\|^2 - \|e_j\|^2 \right)$$

$$\frac{1}{n} \sum_j w_{ij} = -\frac{1}{2} \left( d_{i.}^2 - \|e_i\|^2 - \frac{1}{n} \sum_j \|e_j\|^2 \right) = -\frac{1}{2} \left( d_{i.}^2 - \|e_i\|^2 - \frac{d^2}{2} \right) = 0$$

si les axes sont centrés sur g car  $\frac{1}{n} \sum_j w_{ij} = \langle e_i; \frac{1}{n} \sum_j e_j \rangle$

$$\|e_i\|^2 = d_{i.}^2 - \frac{d^2}{2} \quad \text{et} \quad \|e_j\|^2 = d_{.j}^2 - \frac{d^2}{2}$$

- Matriciellement
  - Opérateur de centrage

$$\mathbf{A} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}'$$

- Double centrage en lignes et en colonnes

$$\mathbf{W} = -\frac{1}{2} \mathbf{A}\mathbf{D}\mathbf{A}$$

- Les coordonnées sur les axes principaux sont donnés par les vecteurs propres de  $W$
- Les vecteurs propres doivent être normalisés comme suit

$$\frac{1}{n} \sum_{i=1}^n c_i^2 = \lambda$$

- Le nombre de valeurs propres non nulles donne la dimension de l'espace
- Distance euclidienne si aucun  $\lambda$  négatif

## 2. La méthode de la constante additive

- Si  $d$  n'est pas euclidienne. En ajoutant  $c^2$  à tous les carrés de distance, on peut la rendre euclidienne

$$\delta_{ij}^2 = d_{ij}^2 + c^2 \quad \text{et} \quad \delta_{ii} = 0$$

$$\mathbf{W}_\delta = \mathbf{W}_d + \mathbf{W}_c$$

$$\mathbf{W}_c = -\frac{1}{2} \mathbf{A} \begin{pmatrix} 0 & c^2 & c^2 & c^2 \\ c^2 & 0 & c^2 & c^2 \\ c^2 & c^2 & 0 & c^2 \\ c^2 & c^2 & c^2 & 0 \end{pmatrix} \mathbf{A} = -\frac{c^2}{2} \mathbf{A} (\mathbf{1}\mathbf{1}' - \mathbf{I}) \mathbf{A}$$

*puisque*  $\mathbf{1}\mathbf{1}' = n(\mathbf{I} - \mathbf{A})$

$$\mathbf{W}_c = -\frac{c^2}{2} \mathbf{A} ((n-1)\mathbf{I} - n\mathbf{A}) \mathbf{A} = \frac{c^2}{2} \mathbf{A}$$

- Les vecteurs propres de  $W_\delta$  sont les mêmes que ceux de  $W_d$  car ils sont centrés.
- Leurs valeurs propres sont augmentées de  $c^2/2$
- Il suffit alors de prendre  $c^2 = 2|\lambda_{\min}|$
- Transforme directement une dissimilarité (pas d'inégalité triangulaire) en une distance euclidienne

F.Cailliez a résolu en 1983 le problème consistant à ajouter la plus petite constante à la distance d'origine :

cette constante est la plus grande valeur propre de la matrice carrée suivante de taille  $2n$

$$\begin{pmatrix} 0 & 2W_d \\ -I & -4W_{\sqrt{d}} \end{pmatrix}$$

$W_{\sqrt{d}}$  est la matrice de Torgerson où les carrés sont remplacés par les distances.

**Dans les deux cas, il ne faut pas que la constante soit trop grande**

# 3. Le MDS semi metrique

- Rechercher une configuration de n points dans un espace de dimension p fixée à l'avance, dont les interdistances  $\delta_{ij}$  soient proches des dissimilarités  $d_{ij}$

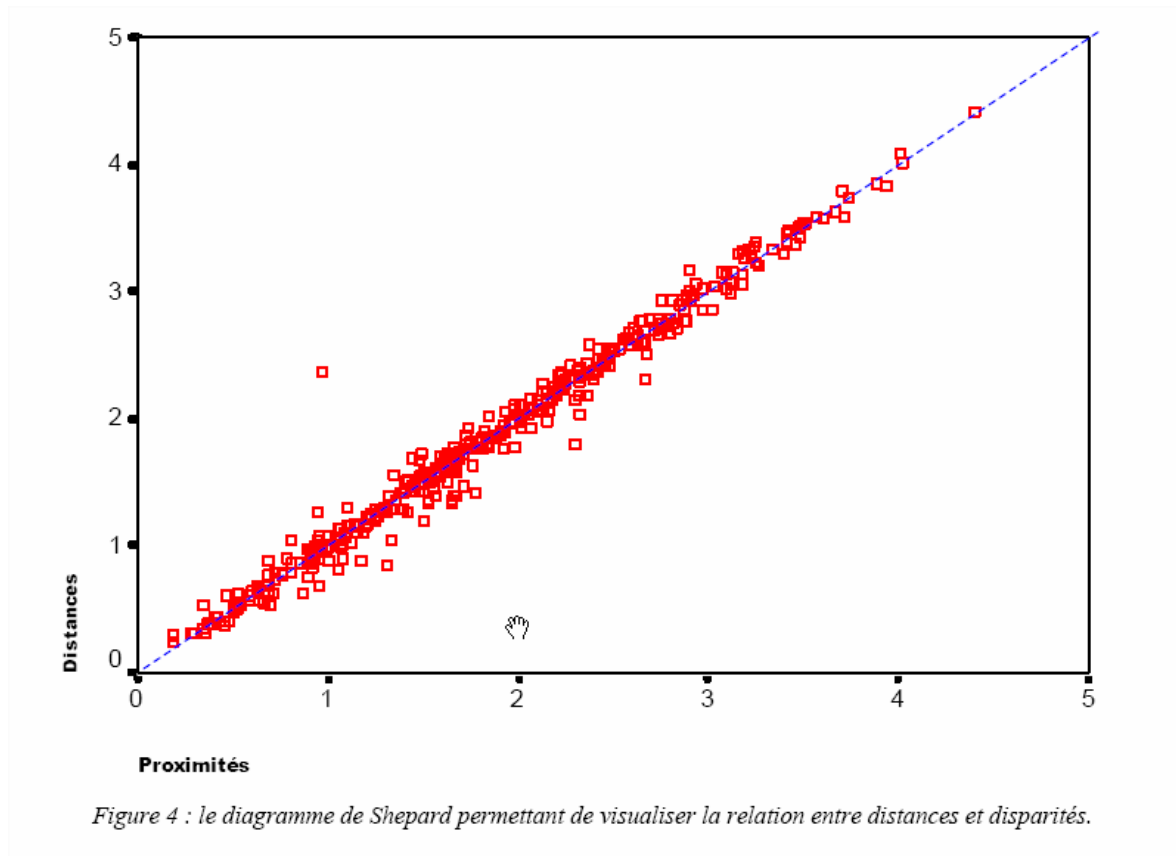
- La méthode de Kruskal de minimisation du STRESS

$$\min \frac{\sum_{i,j} (\delta_{ij} - f(d_{ij}))^2}{\sum_{i,j} (\delta_{ij})^2}$$

- f transformation monotone (on ne garde que l'information portée par l'ordre des dissimilarités)

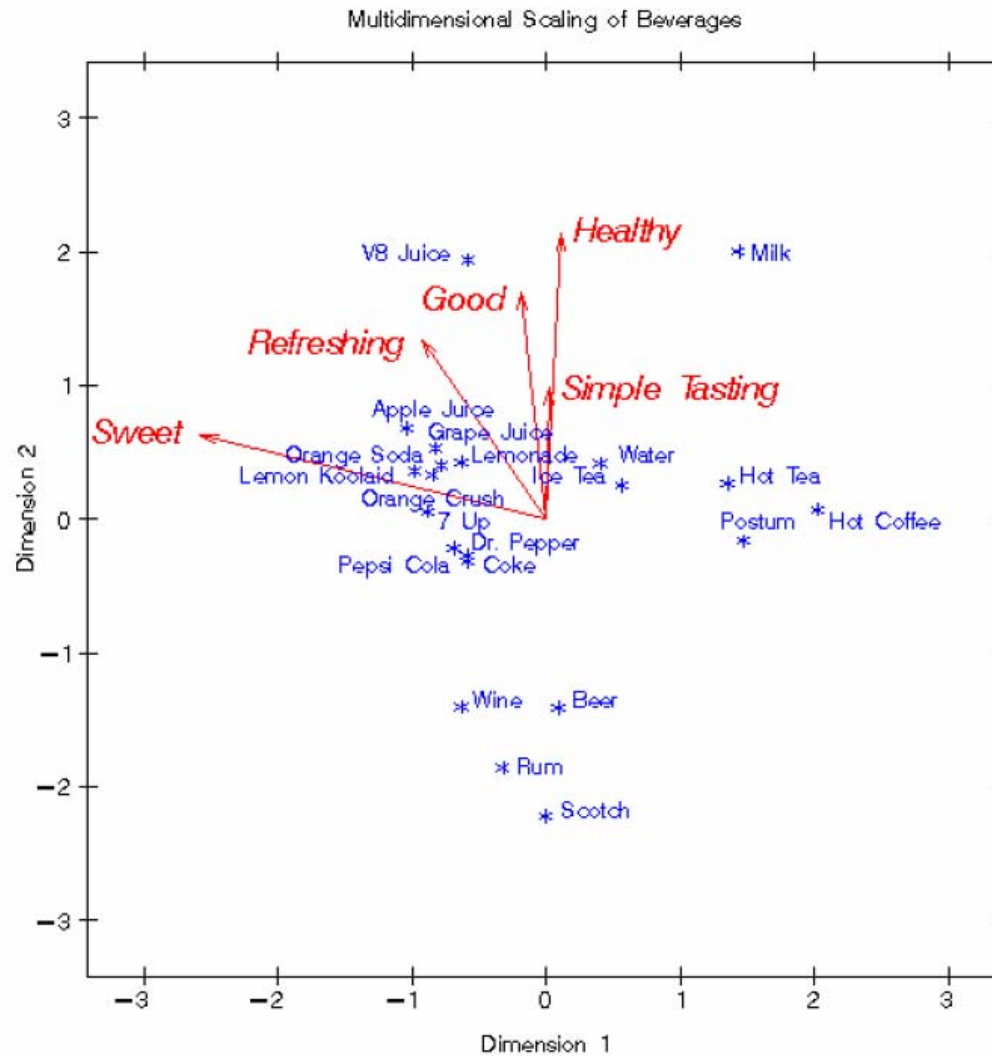
- Algorithme:
  - On part d'une configuration euclidienne.
  - On calcule les disparités  $f(d_{ij})$  et le stress par régression monotone
  - On modifie la configuration à l'aide d'une méthode de gradient en déplaçant les points pour diminuer le stress
  - Etc.

- Diagramme de Shepard



- Nécessite de connaître  $p$ : solutions non emboîtées
- Approximation en plus ou en moins alors qu'avec Torgerson approximation par en dessous

- Possibilité de rajouter des variables explicatives (voir cartographie des préférences)



# 3. Cas de $q$ tableaux de distances

- Le modèle INDSCAL (Individual Differences In Scaling)

$$\left(d_{ij}^k\right)^2 \cong \sum_{l=1}^r m_l^k \left(x_i^l - x_j^l\right)^2$$

- Configuration unique avec métriques diagonales différentes
- Passage aux produits scalaires

$$w_{ij}^k = \sum_{l=1}^r m_l^k a_i^l b_j^l + \varepsilon$$

- Estimation alternée : on fixe  $m$  et  $a$  et on estime  $b$ , puis  $a$  à  $m$  et  $b$  fixés, puis  $m$  à  $a$  et  $b$  fixés etc.

# • Exemple (SensoMineR)

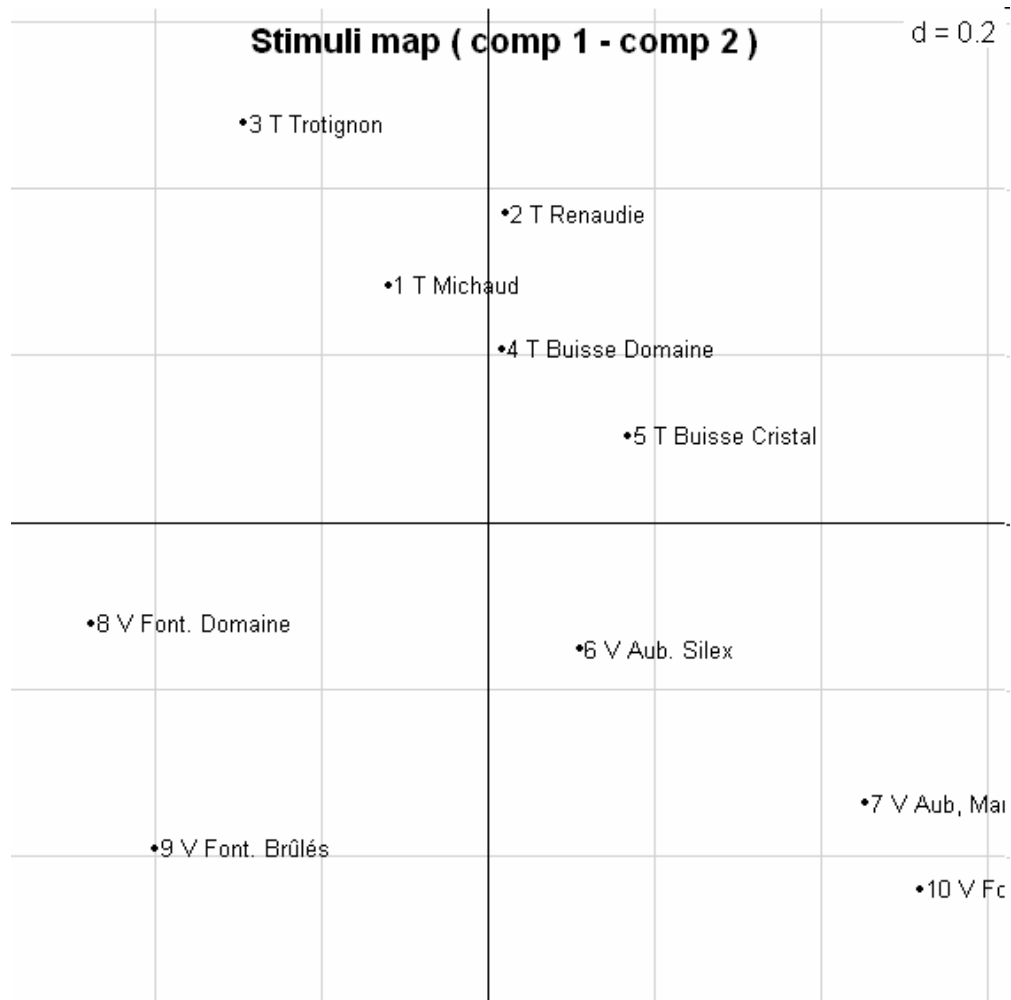
- 10 different French wines evaluated by 11 panelists. Two data sets: napping.don and napping.words.
  - 1) napping.don: panelists were asked to position the wines on a tablecloth of dimension (60,40)
  - 2) napping.words: panelists were asked to describe each wine using their own word list

Wines	X1	Y1	X2	Y2	X3	Y3		X10	Y10	X11	Y11
T Michaud	43.0	29.5	12.5	15.5	48.0	15.0		25.5	10.0	25.5	30.0
T Buisse Domaine	18.0	20.0	19.0	22.0	31.0	9.5		7.5	13.0	55.0	8.5
T Buisse Cristal	17.0	22.0	24.0	30.0	34.5	31.0	...	34.5	13.0	17.0	31.5
V Font. Domaine	56.0	3.0	23.0	20.0	4.5	5.0		7.5	18.5	43.5	17.0
V Font. Brûlés	42.5	4.5	22.0	26.0	8.0	6.5		18.5	19.0	56.0	19.0
V Font Coteaux	1.5	38.5	8.0	23.0	54.0	36.0		44.5	19.0	35.5	11.5

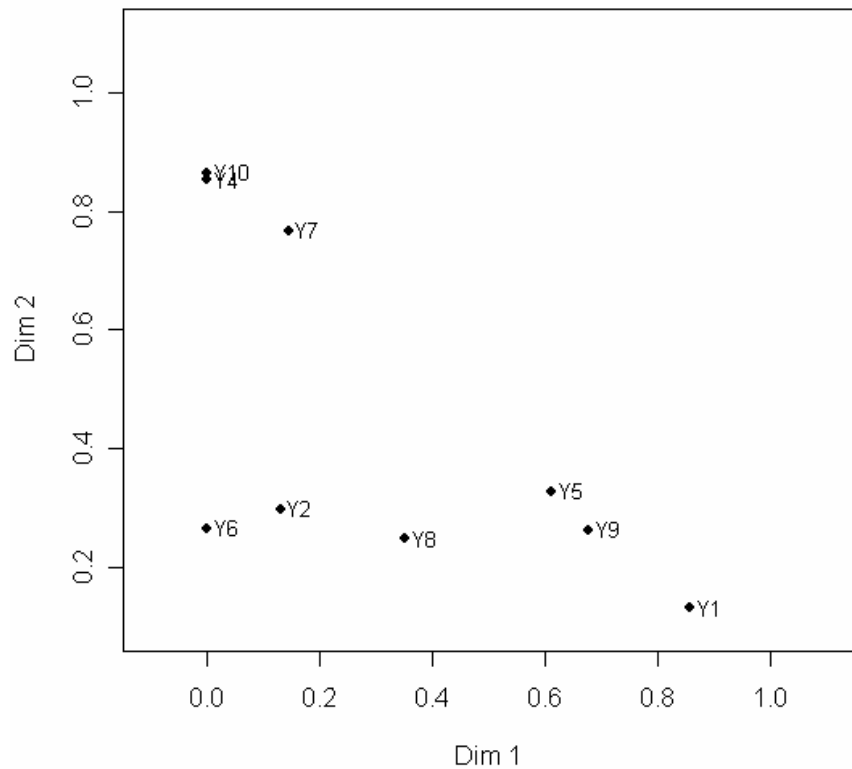
Tab1: napping.don

Wines	Wood	Liqueur like	Fresh-Sharp	Fruity	Discrete	Intense	Grilled bread	Floral	Bitterness	Green
T Michaud	1	0	3	2	0	2	0	2	1	3
T Buisse Domaine	0	0	3	2	4	0	0	0	0	1
T Buisse Cristal	3	0	3	2	0	...	1	0	0	0
V Font. Domaine	0	1	0	4	1	0	0	1	1	0
V Font. Brûlés	2	0	0	2	1	1	0	0	2	0
V Font Coteaux	7	0	0	1	0	1	4	2	3	

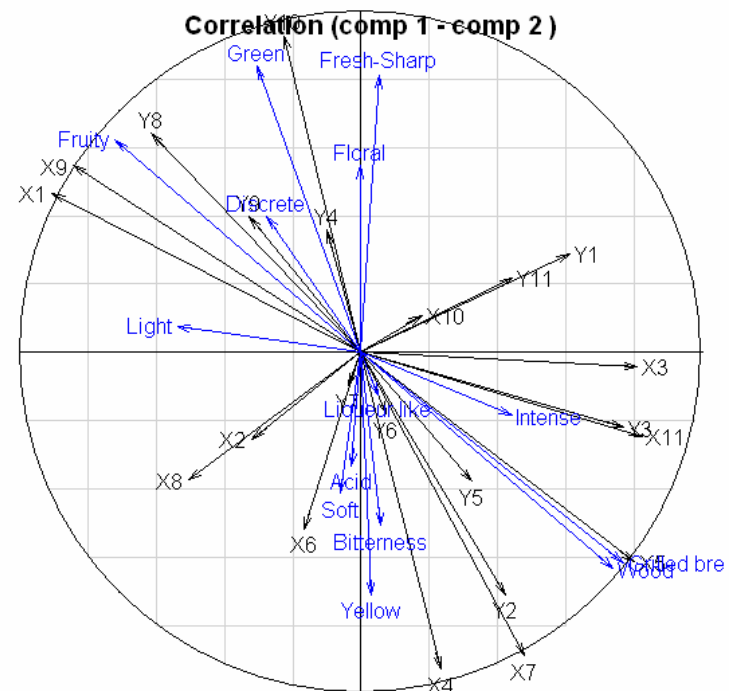
Tab2: napping.words



**Weight representation ( comp 1 - comp 2 )**



**Correlation (comp 1 - comp 2)**



- Modèle IDIOSCAL (Individual Differences In Orientation and Scaling)
  - Métriques  $M_k$  quelconque
  - Solution analytique

$$W_k = X M_k X'$$

$$W = \frac{1}{n} \sum_{k=1}^q W_k = X \left( \frac{1}{n} \sum_{k=1}^q M_k \right) X' = X X'$$

$$\text{car on peut prendre } \frac{1}{n} \sum_{k=1}^q M_k = I$$

- $X$  s'obtient par une méthode classique de Torgerson sur  $W$

$$W_k = X M_k X'$$

$$M_k = (X'X)^{-1} X' W_k X (X'X)^{-1}$$

# 4. Cartographie des préférences

- Analyser les préférences des consommateurs
- Applications en marketing et analyse sensorielle:
  - Ex. : relier les préférences des consommateurs aux caractéristiques physico-chimiques et/ou sensorielles d'un produit
  - Visualiser ces relations sur une carte «facilement » lisible

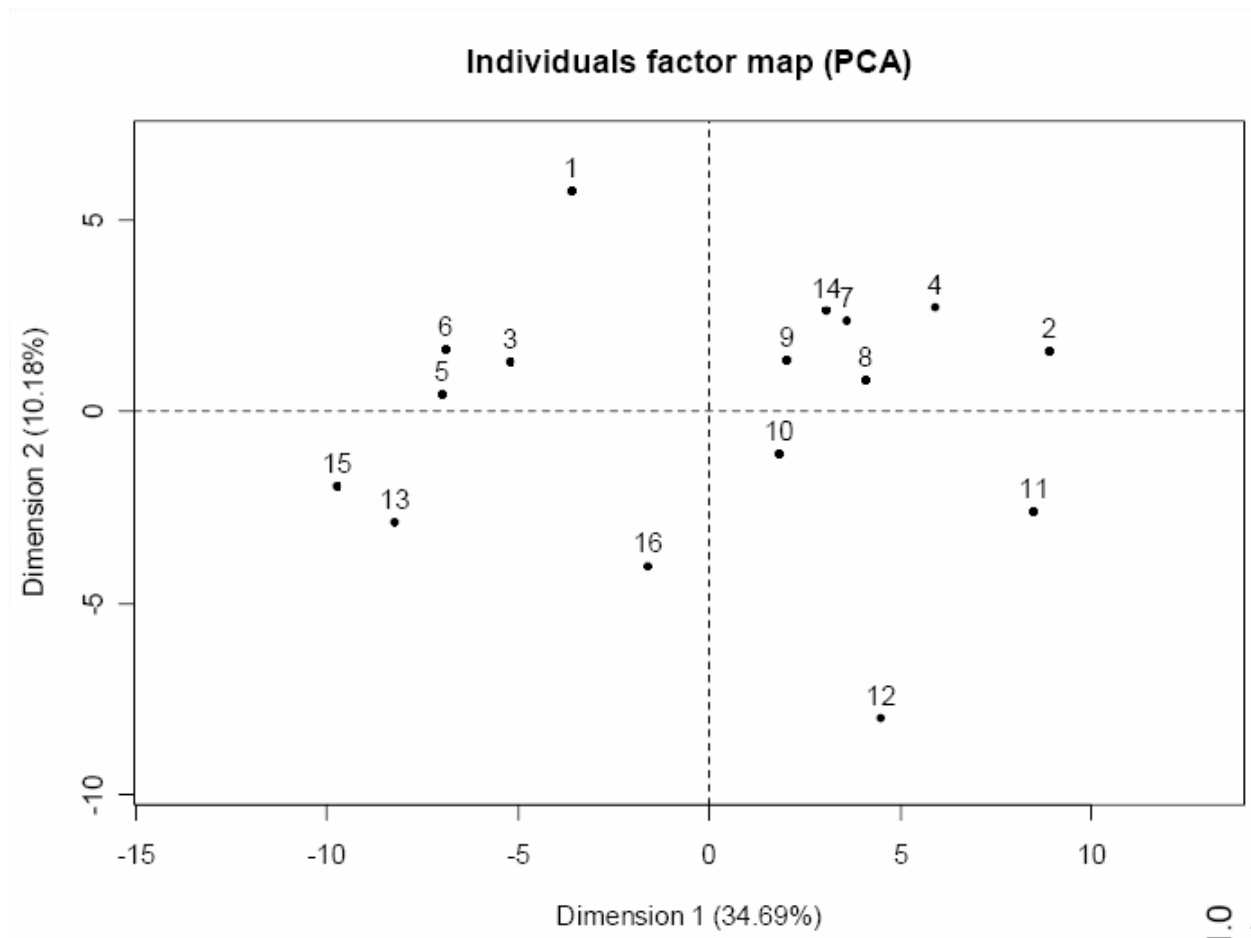
- $n$  produits,  $p$  consommateurs,  
 $q$  caractéristiques
  - Tableau  $X$  de notes  $(n,p)$
  - Tableau  $Y$  de caractéristiques  $(n,q)$

Exemple (cf SensoMineR)

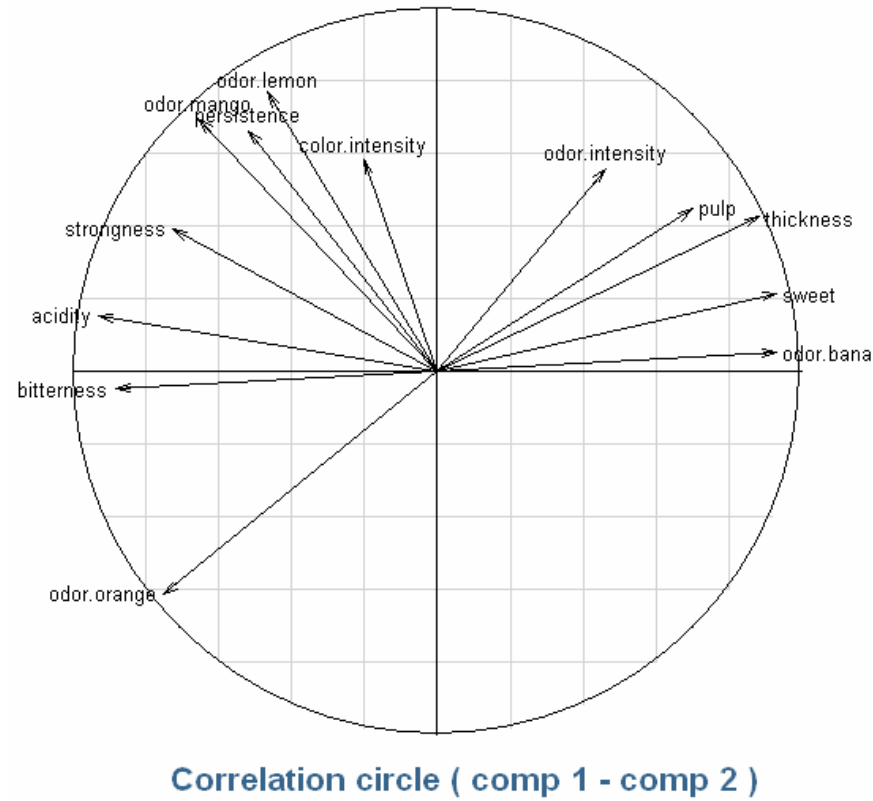
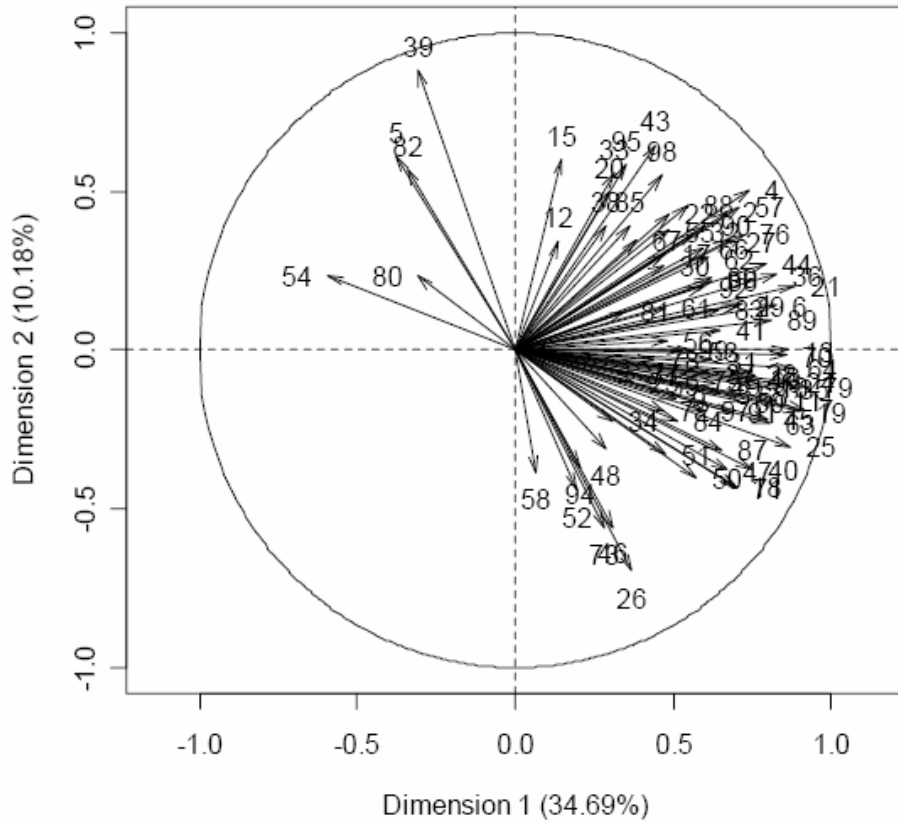
16 cocktails, jugés par 100 consommateurs  
et décrits par 13 variables sensorielles  
(fournies par des experts)

# 4.1 Cartographie « interne » ou MDPREF

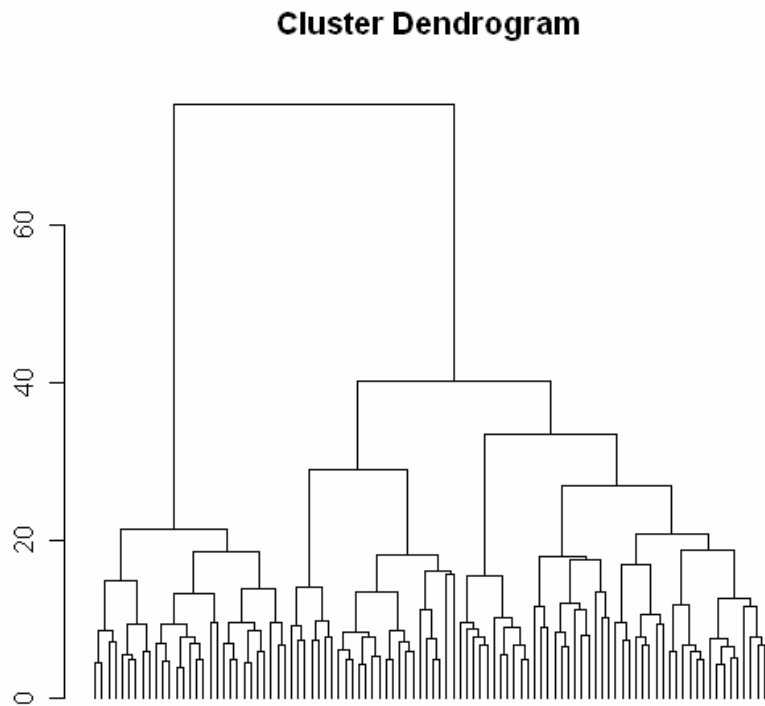
- ACP de X avec projection en éléments supplémentaires des colonnes de Y



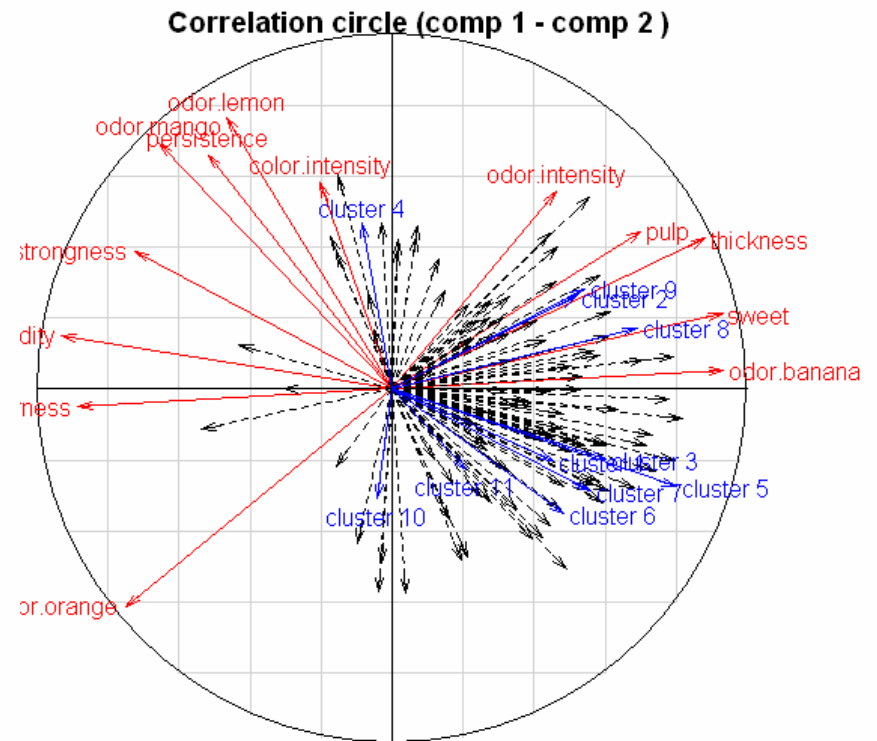
**Variables factor map (PCA)**



- Avec classification des consommateurs



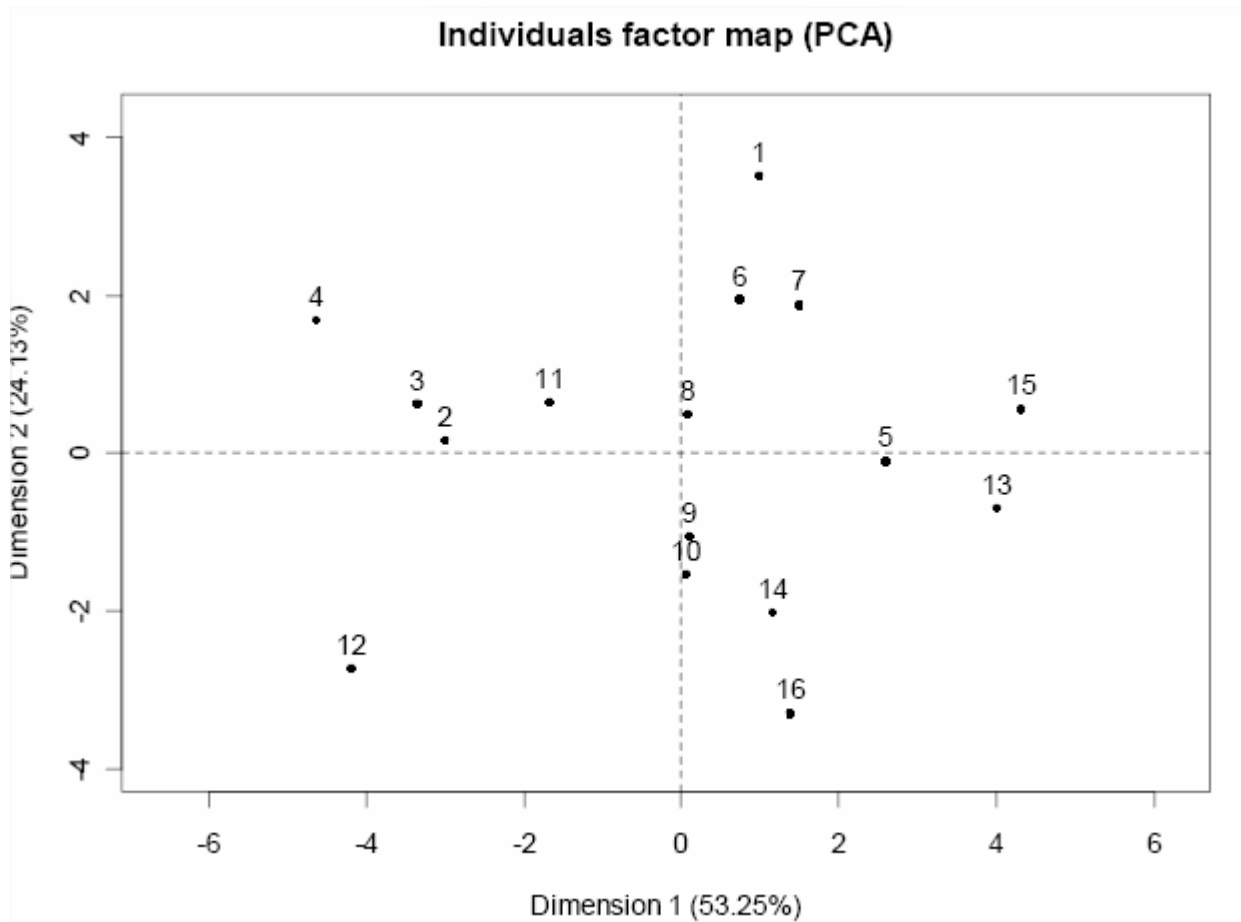
Panelists

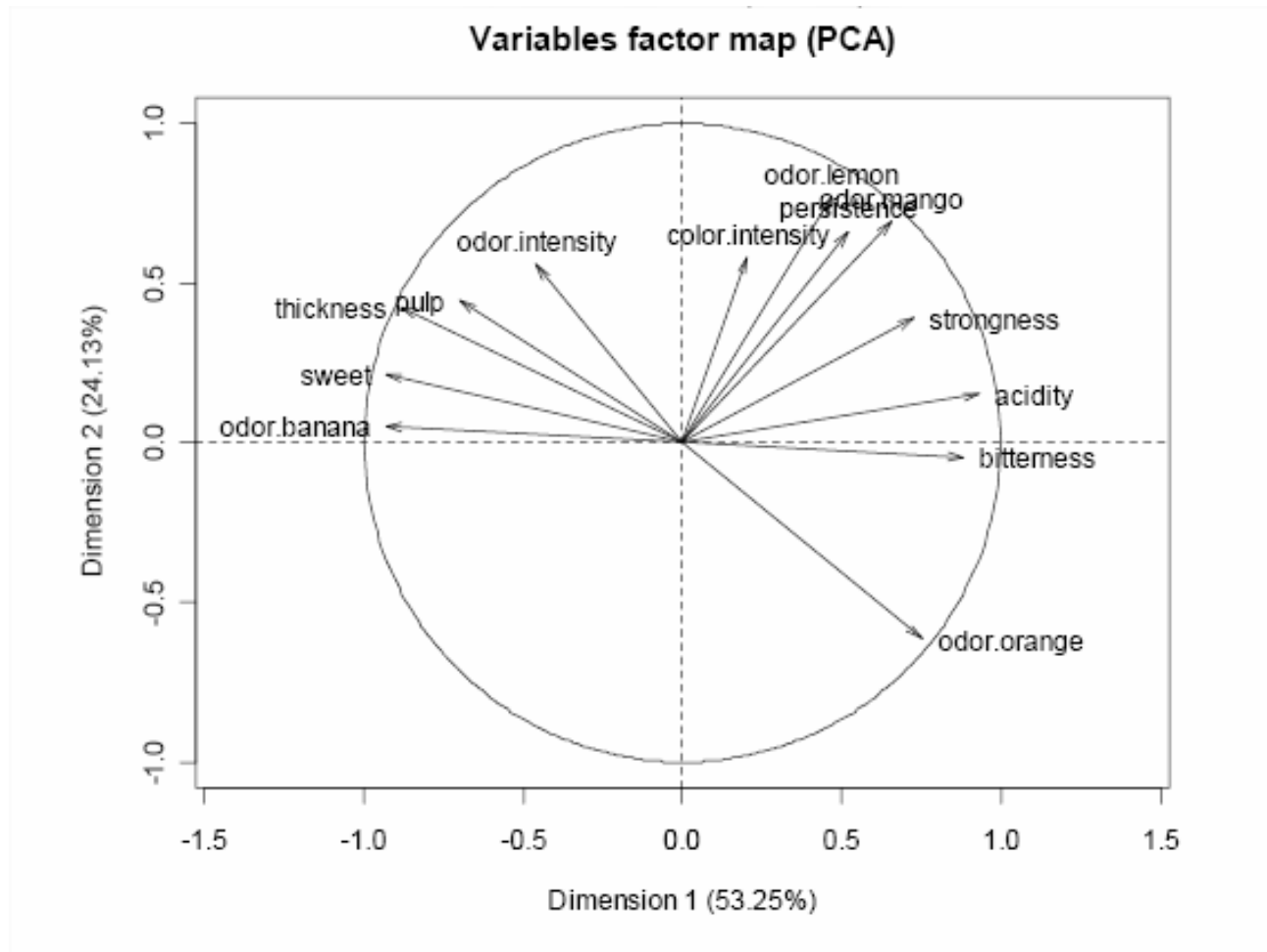


## 4.2 Cartographie « externe » ou PREFMAP

- ACP de Y (descripteurs en variables actives)
- Régression des préférences des consommateurs sur les axes de l'ACP

[http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/20286\\_preference.pdf](http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/20286_preference.pdf)





- Modélisation des préférences de j expliquée par les deux premières composantes principales de Y
  - modèle linéaire ou vectoriel  $x_j = m + ac_1 + bc_2$

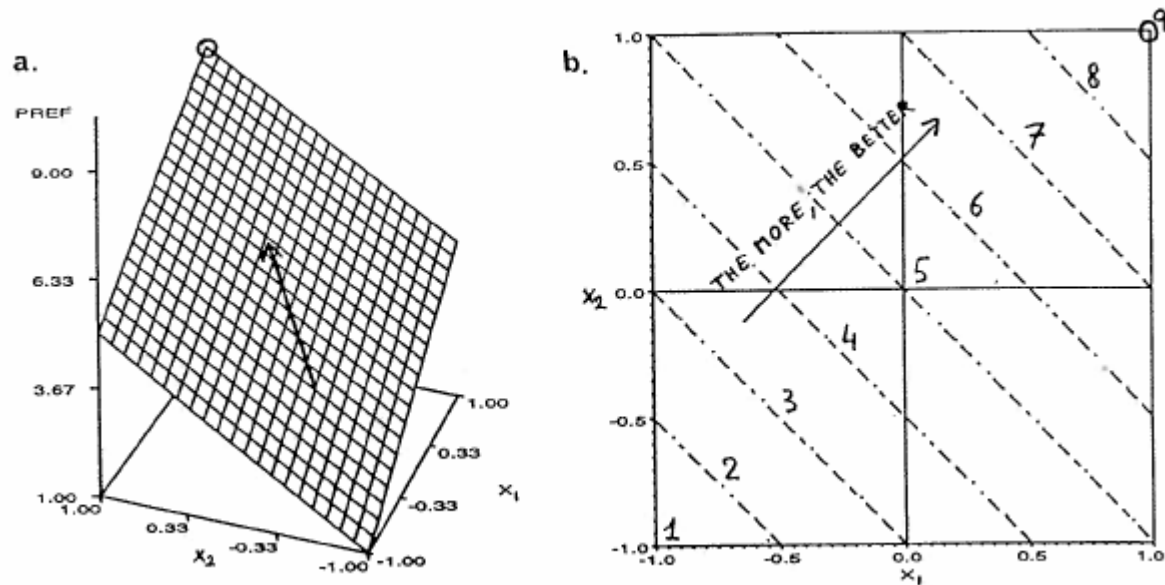


Figure 1 : Vectorial model  $PREF = 5 + 2X_1 + 2X_2$  ; a. response surface, b. isocontours

Schlich, 2007

– Modèle linéaire pas toujours adapté si le produit idéal est au milieu : *ni trop ni trop peu sucré.*

– modèles circulaires ou elliptiques « point idéal »

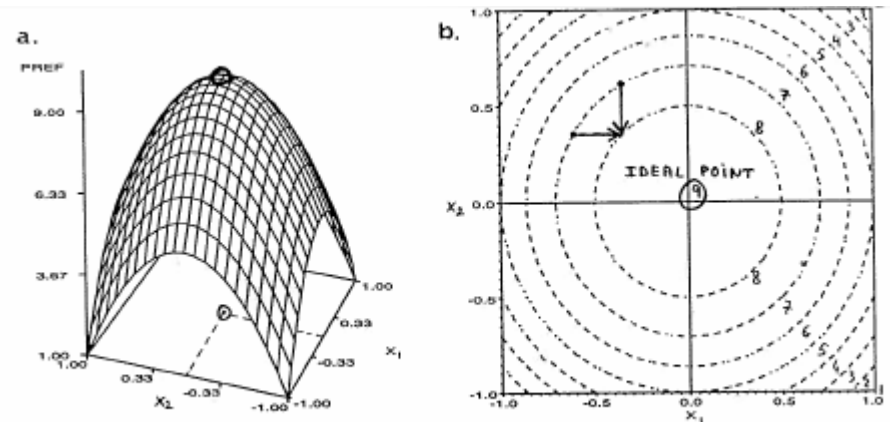


Figure 2 : Circular model  $PREF = 9 - 4X_1^2 - 4X_2^2$ ; a. response surface, b. isocontours

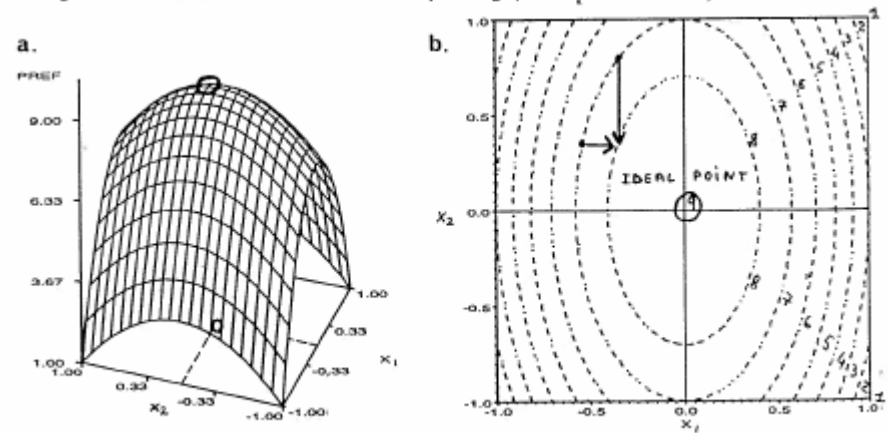
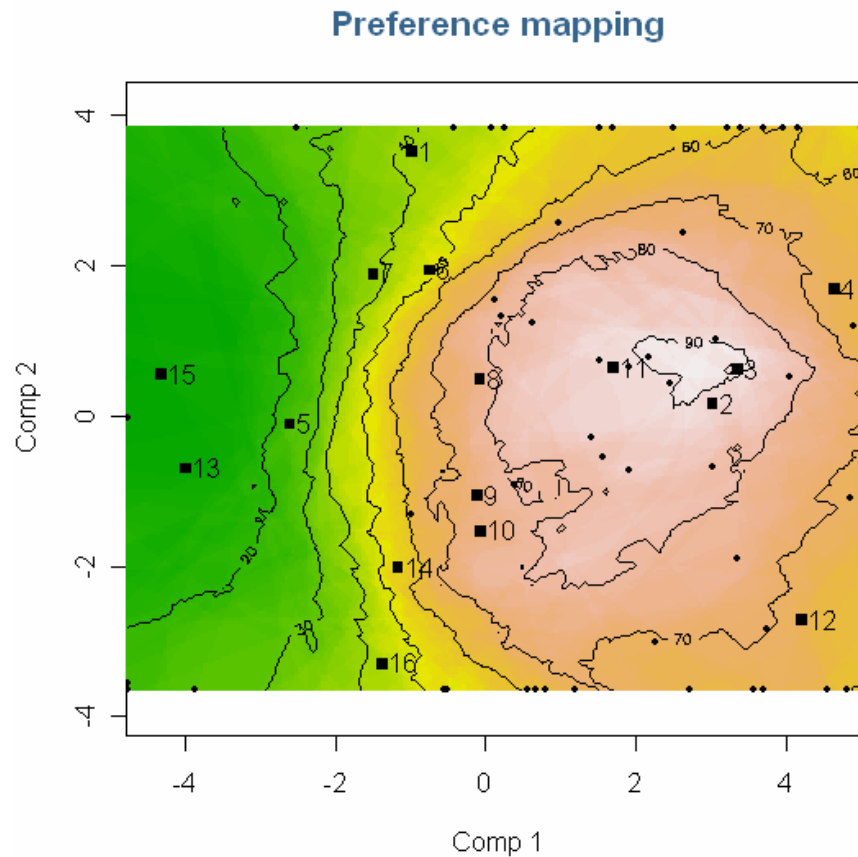


Figure 3 : Elliptical model  $PREF = 9 - 6X_1^2 - 2X_2^2$ ; a. response surface, b. isocontours

- Cumul des préférences: pour chaque point on estime le pourcentage de consommateurs qui préfèrent ce point



# Références

- Aubigny (d') G. (2003): Positionnement multidimensionnel et quantification vectorielle , in *Traitement du signal et de l'image. Analyse des Données* Govaert G. (dir.) Hermès, 105-150.
- Cailliez F. (1983): The Analytical Solution of the Additive Constant Problem , *Psychometrika*, 48, 305-308
- [Desbois D. \(2005\): Une introduction au positionnement multidimensionnel. \*Modulad\*, 32, 1-28](#)
- Kruskal J.B., Wish M.: (1984) Multidimensional Scaling, *Quantitative Applications in the Social Sciences*, 11, Sage University Paper.
- [Kuhfeld W. \(2010\): Marketing Research Methods in SAS, \*MR2010 report\*, SAS Institute](#)
- [Schlich,P, \(2007\): De la sensometrie à l'analyse des préférences, HDR, Univ.Bourgogne](#)
- <http://sensominer.free.fr>