le cnam

Le Multidimensional Scaling et la cartographie des préférences

Gilbert Saporta
Conservatoire National des Arts et Métiers
http://cedric.cnam.fr/~saporta

Avril 2014

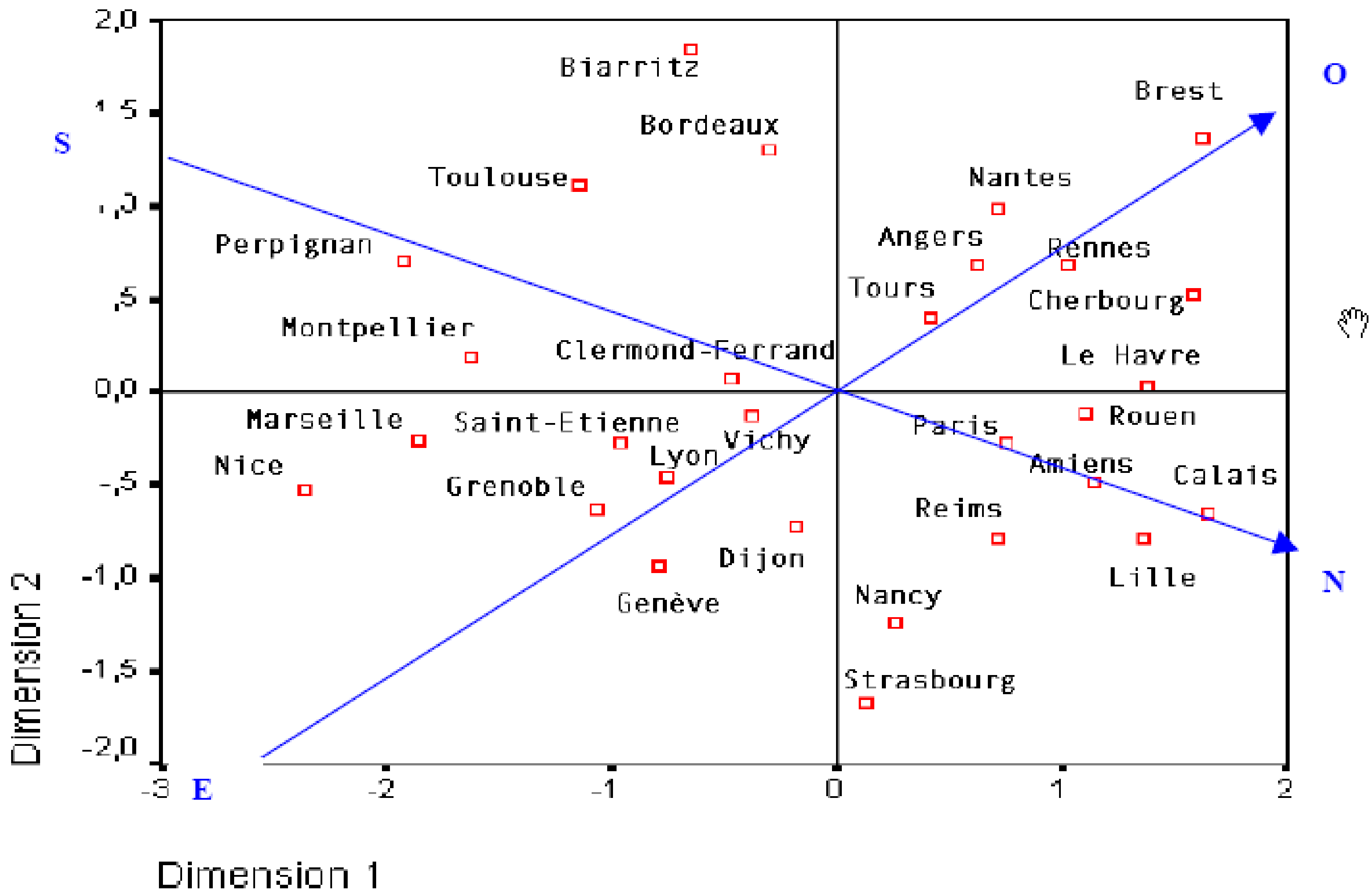
Multidimensional scaling

- 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

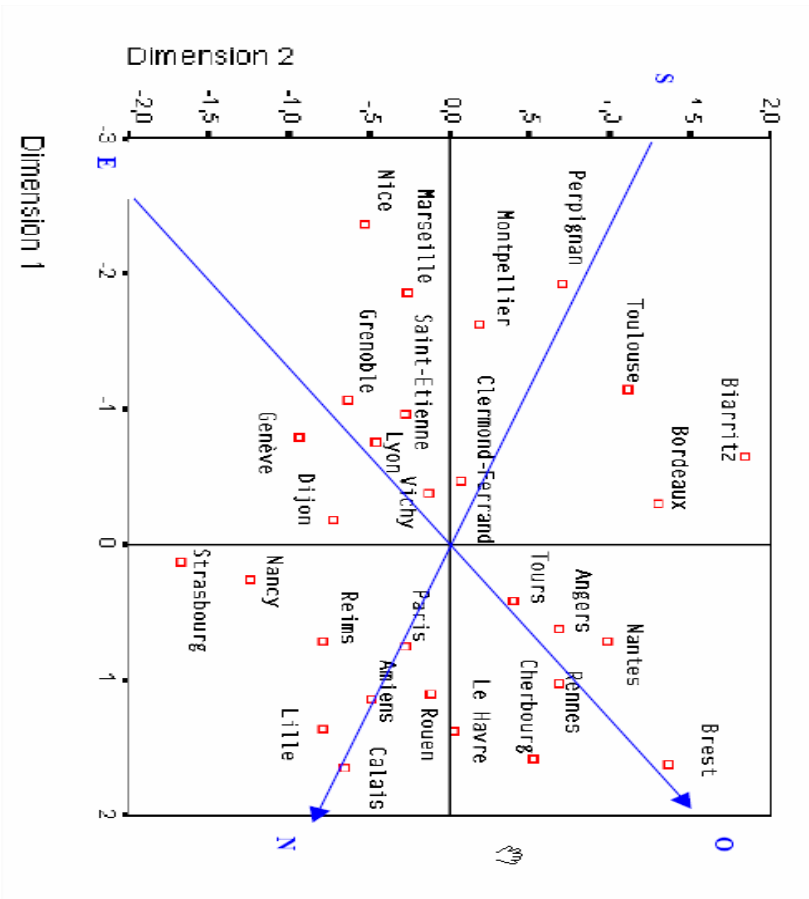
- Exemple: reconstituer une carte connaissant le tableau des distances entre n villes de France

Tableau 1: la France en automobile	Villes	Amiens	Angers	Biarritz	Bordeaux	Brest	Calais	Cherbourg	Clermond-Ferrand	Dijon
Villes		V01	V02	V03	V04	V05	V06	V07	V08	V09
Amiens	v01	0	399	862	679	619	159	366	524	417
Angers	v02	399	0	518	335	371	494	290	402	498
Biarritz	v03	862	518	0	183	817	997	808	555	815
Bordeaux	v04	679	335	183	0	634	814	625	369	632
Brest	v05	619	371	817	634	0	714	402	752	812
Calais	v06	159	494	997	814	714	0	461	672	556
Cherbourg	v07	366	290	808	625	402	461	0	647	661
Clermond-Ferrand	v08	524	402	555	369	752	672	647	0	280
Dijon	v09	417	498	815	632	812	556	661	280	0

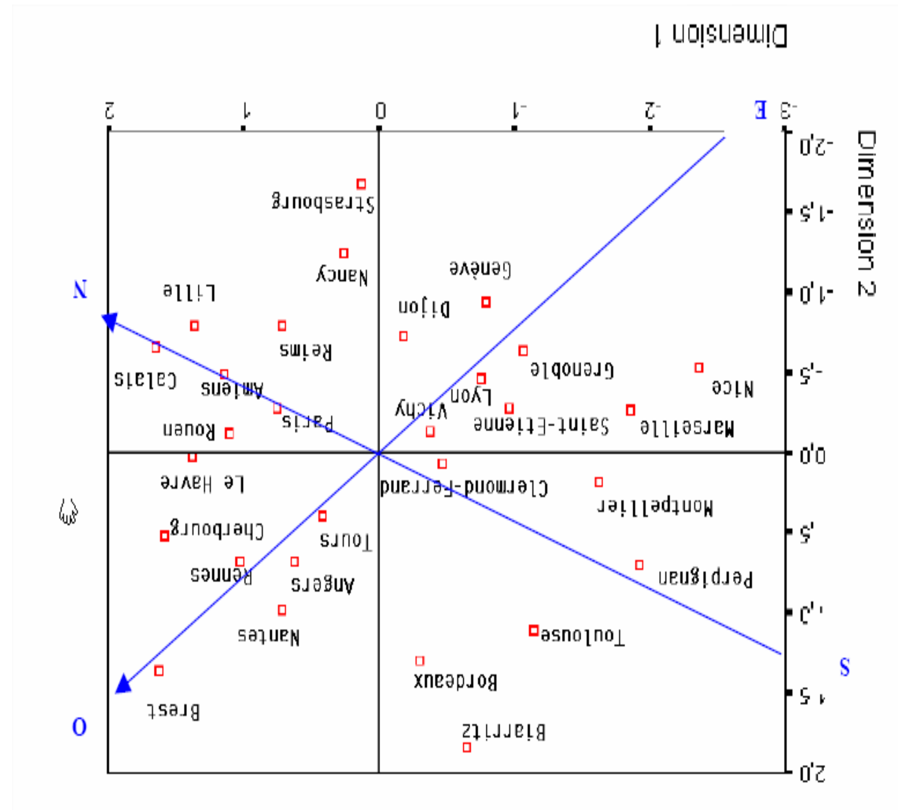
Desbois, 2005


2,0
1,5
1,0
,5
0,0
-,5
-1,0
-1,5
-2,0
-3
-2
-1
0
1
2
Dimension 2
Dimension 1
S
O
E
N
Biarritz
Bordeaux
Toulouse
Perpignan
Montpellier
Clermond-Ferrand
Marseille
Saint-Etienne
Vichy
Lyon
Nice
Grenoble
Dijon
Genève
Brest
Nantes
Angers
Rennes
Tours
Cherbourg
Le Havre
Paris
Rouen
Amiens
Calais
Reims
Lille
Nancy
Strasbourg

- Mais aussi:



Dimension 2
Dimension 1
E
S
N
O
Nice
Marseille
Montpellier
Perpignan
Toulouse
Biarritz
Bordeaux
Grenoble
Saint-Etienne
Clermond-Ferrand
Genève
Lyon
Vichy
Dijon
Strasbourg
Nancy
Reims
Paris
Tours
Angers
Nantes
Amiens
Rouen
Le Havre
Cherbourg
Rennes
Brest
Lille
Calais



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

- Axiomes:

$$\left.\begin{array}{l} \left.\begin{array}{l} d_{ij} \geq 0 \\ d_{ii} = 0 \\ d_{ij} = d_{ji} \end{array}\right\} \text{dissimilarité} \\ d_{ij} \leq d_{ik} + d_{kj} \end{array}\right\} \text{distance}$$

$$d_{ij} = \left(\left(\mathbf{e_i} - \mathbf{e_j}\right)'\mathbf{M}\left(\mathbf{e_i} - \mathbf{e_j}\right)\right)^{\frac{1}{2}}$$ Distance euclidienne

- **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

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

- 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 $\frac{1}{n}\sum_j w_{ij} = < e_i ; \frac{1}{n}\sum_j e_j >$

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

- Matriciellement
  - Opérateur de centrage

$$\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{11'}$$

  - Double centrage en lignes et en colonnes

$$\mathbf{W} = -\frac{1}{2}\mathbf{ADA}$$

- Les coordonnées sur les axes principaux sont données 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 λ négatif

2. La méthode de la constante additive

- Si d n'est pas euclidienne. En ajoutant c² à tous les carrés de distance, on peut la rendre euclidienne

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

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

$$\mathbf{W}_c = -\frac{1}{2}\mathbf{A}\begin{pmatrix} 0 & c^2 & c^2 & c^2 \\ c^2 & 0 & c^2 & c^2 \\ c^2 & c^2 & 0 & c^2 \\ c^2 & c^2 & c^2 & 0 \end{pmatrix}\mathbf{A} = -\frac{c^2}{2}\mathbf{A}\left(\mathbf{11'\text{-}I}\right)\mathbf{A}$$

$$\textit{puisque} \quad \mathbf{11'} = n(\mathbf{I\text{ - }A})$$

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

- Les vecteurs propres de W_δ 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 δ_{ij} soient proches d'une fonction monotone des dissimilarités d_{ij}
- La méthode de Kruskal de minimisation du STRESS

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

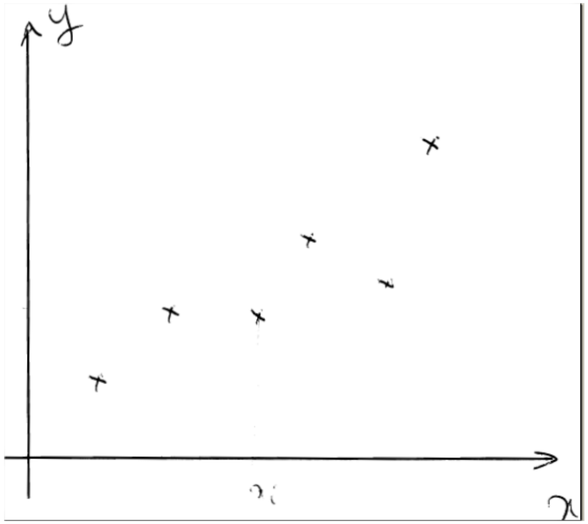
- f transformation monotone (on ne garde que l'information portée par l'ordre des dissimilarités)

- 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.

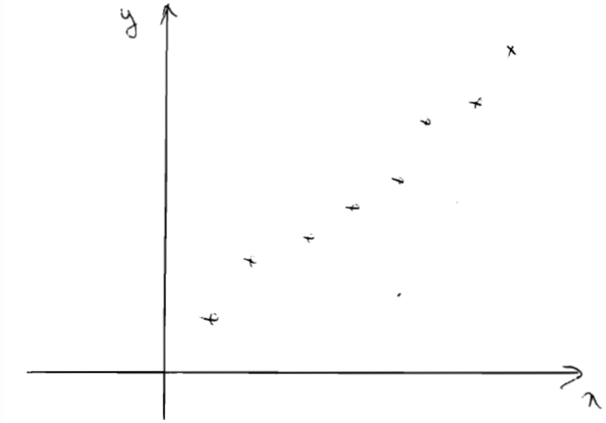
La régression monotone simple

- Régression simple sur variable transformée

$$\min_T \sum_{i=1}^{n} \left(y_i - T(x_i) \right)^2$$



Si relation monotone T(x)=y



- Algorithme de Kruskal:
  - Si $y_2<y_1$ on pose $T(x_1)=T(x_2)= (y_1+y_2)/2$ et on continue avec $T(x_3)=y_3$ si $y_3>(y_1+y_2)/2$
  - Sinon on pose $T(x_1)=T(x_2)=T(x_3)= (y_1+y_2+ y_3)/3$
  - Etc.
- Présentation matricielle:

$$\mathbf{y}=\begin{pmatrix}1&0&0&0\\1&1&0&0\\1&1&1&0\\1&1&1&1\end{pmatrix}\begin{pmatrix}b_1\\b_2\\.\\b_n\end{pmatrix}+\mathbf{e}=\begin{pmatrix}T(x_1)\\T(x_2)\\.\\T(x_n)\end{pmatrix}+\mathbf{e}$$

Régression multiple à coefficients positifs sauf la constante

$$\hat{\mathbf{y}}=b_1\mathbf{1}+\sum_{j=2}b_j\mathbf{z}_j$$

- Un algorithme général de régression sous contraintes de positivité:

Régression descendante avec élimination des variables à coefficients négatifs et itération.

- Diagramme de Shepard

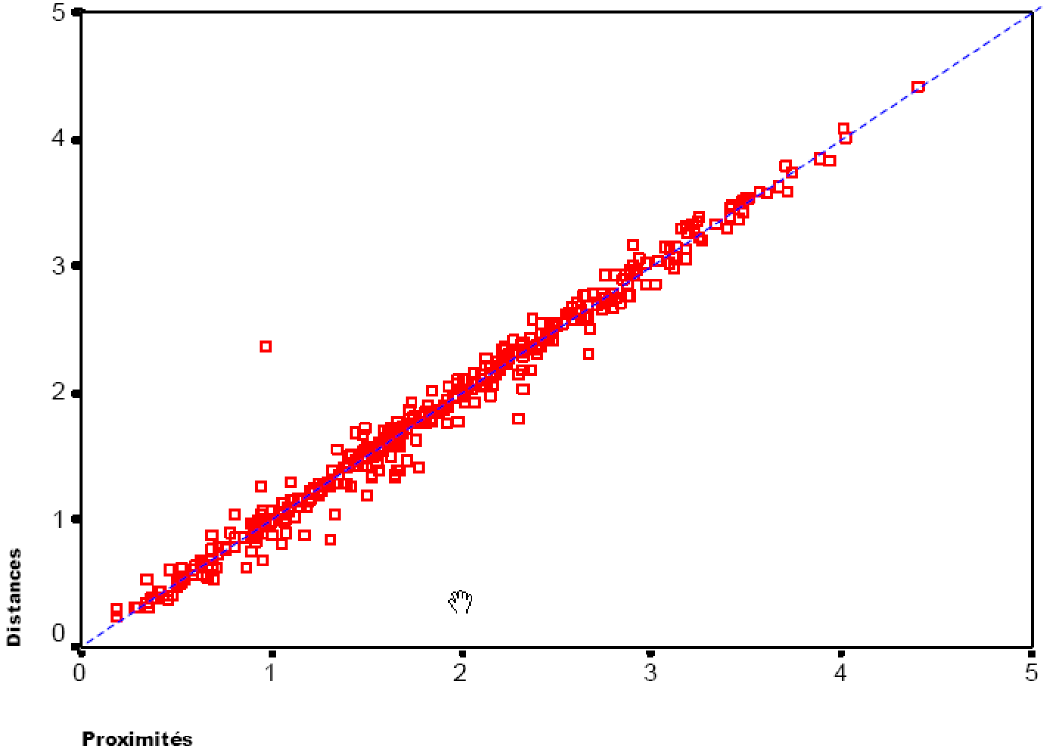
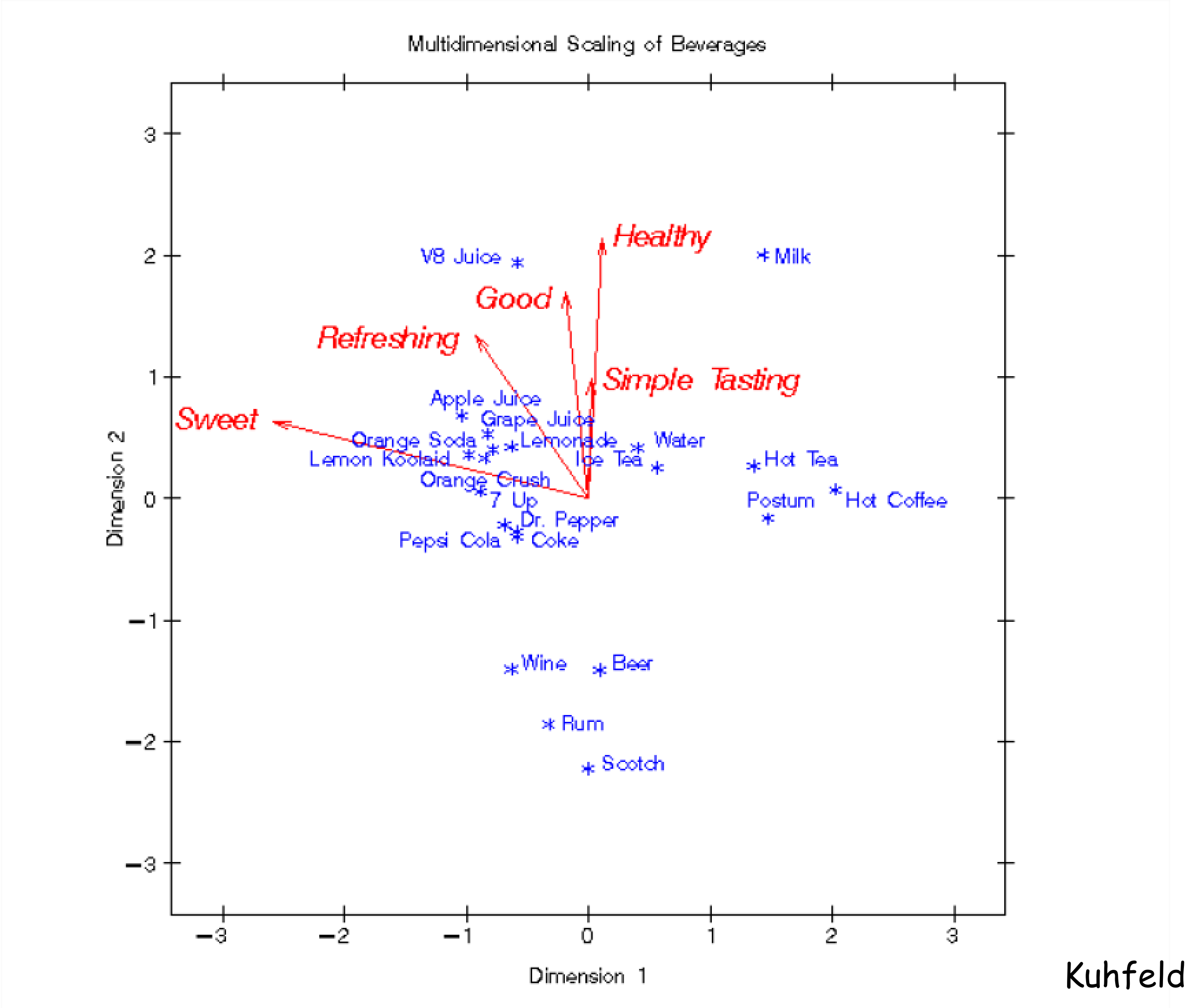


Figure 4 : le diagramme de Shepard permettant de visualiser la relation entre distances et disparités.

Desbois, 2005

- Nécessite de connaitre p: solutions non emboitées
- Approximation en plus ou en moins alors qu’avec Torgerson approximation par en dessous

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



Kuhfeld, 2010

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)

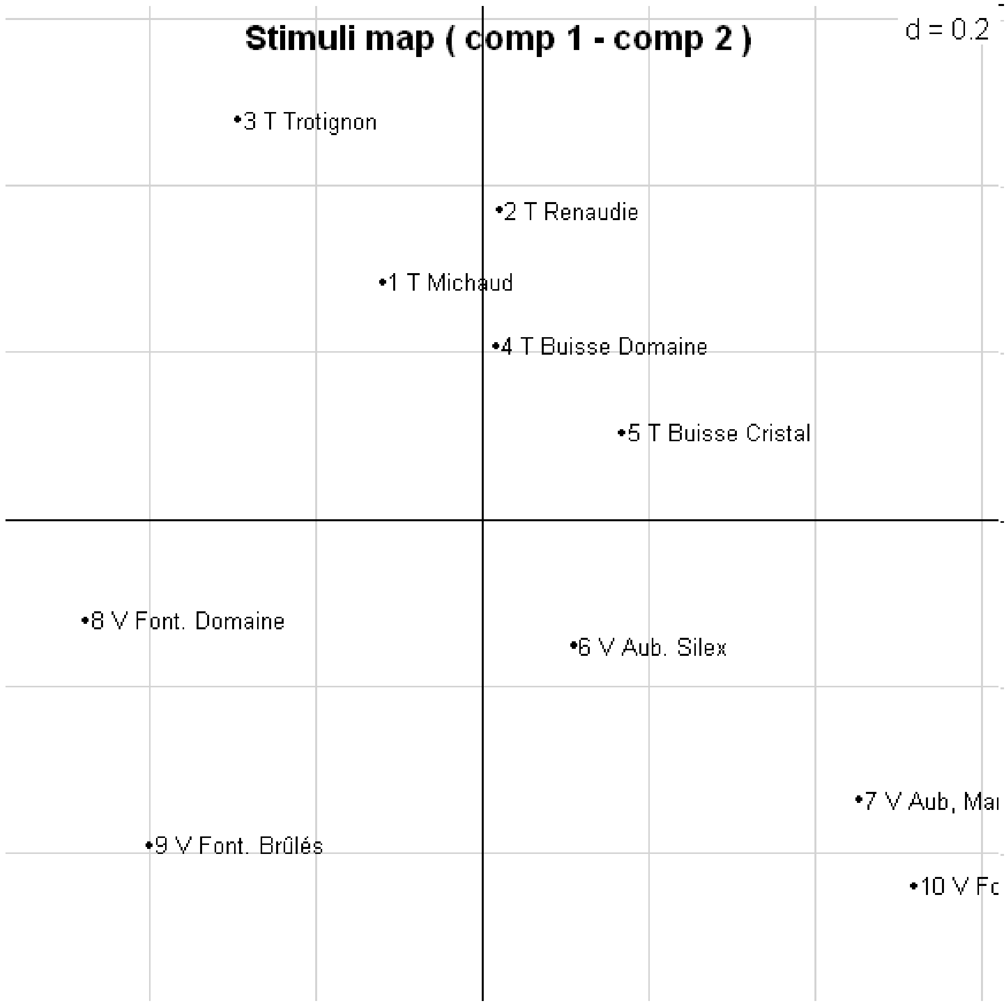
- 10 different French wines evaluated by 11 panelists. Two data sets: napping.don and napping.words.
  - 1) napping.don: panelists were asked to position the wines on a tableclothe of dimension (60,40)
  - 2) napping.words: panelists were asked to describe each wine using their own word list

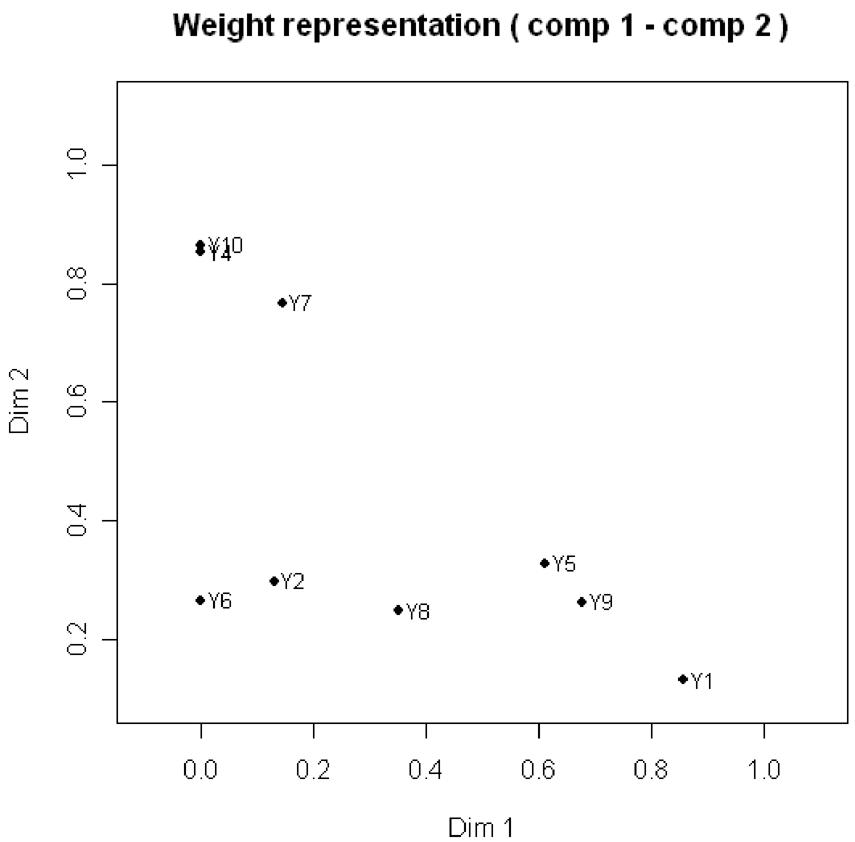
Wines	X1	Y1	X2	Y2	X3	Y3		X10	Y10	X11	Y11
T Michaud	43.0	29.5	12.5	15.5	48.0	15.0		25.5	10.0	25.5	30.0
T Buisse Domaine	18.0	20.0	19.0	22.0	31.0	9.5		7.5	13.0	55.0	8.5
T Buisse Cristal	17.0	22.0	24.0	30.0	34.5	31.0	...	34.5	13.0	17.0	31.5
V Font. Domaine	56.0	3.0	23.0	20.0	4.5	5.0		7.5	18.5	43.5	17.0
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

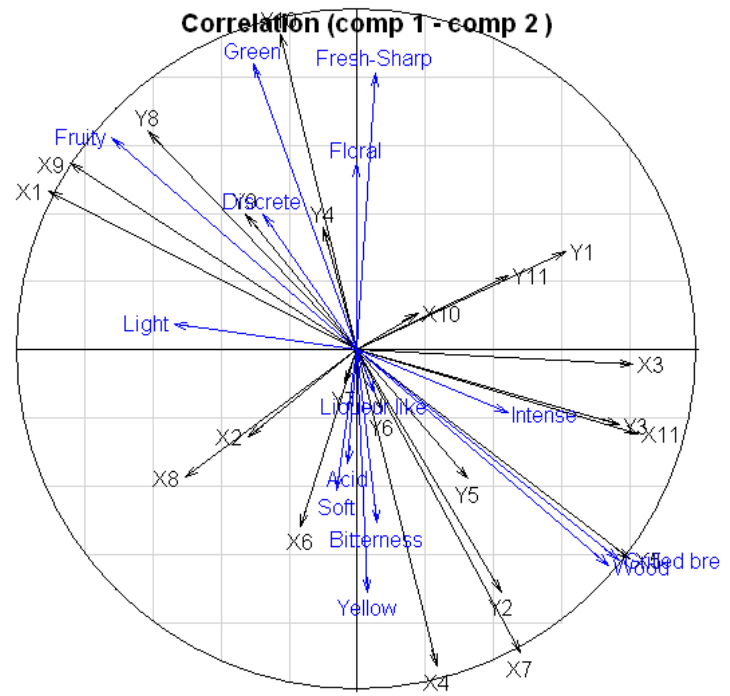
Tab1: napping.don

Wines	Wood	Liqueur like	Fresh-Sharp	Fruity	Discrete		Intense	Grilled bread	Floral	Bitterness	Gree
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	3	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


Stimuli map (comp 1 - comp 2)
d = 0.2
3 T Trotignon
2 T Renaudie
1 T Michaud
4 T Buisse Domaine
5 T Buisse Cristal
8 V Font. Domaine
6 V Aub. Silex
7 V Aub, Ma
9 V Font. Brûlés
10 V Fo


Weight representation (comp 1 - comp 2)
Dim 1
Dim 2
Y10
Y4
Y7
Y6
Y2
Y8
Y5
Y9
Y1


Correlation (comp 1 - comp 2)
Green
Fresh-Sharp
Fruity
Floral
Discrete
Light
Intense
Acid
Soft
Bitterness
Yellow
Y8
X9
X1
Y4
Y1
Y11
X10
X3
X11
X2
X8
X6
Y6
Y5
X7
X4

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

$$\mathbf{W_k} = \mathbf{X M_k X'}$$

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

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

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

$$\mathbf{W_k} = \mathbf{X M_k X'}$$

$$\mathbf{M_k} = \left(\mathbf{X'X}\right)^{\mathbf{-1}} \mathbf{X'W_k X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}$$

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

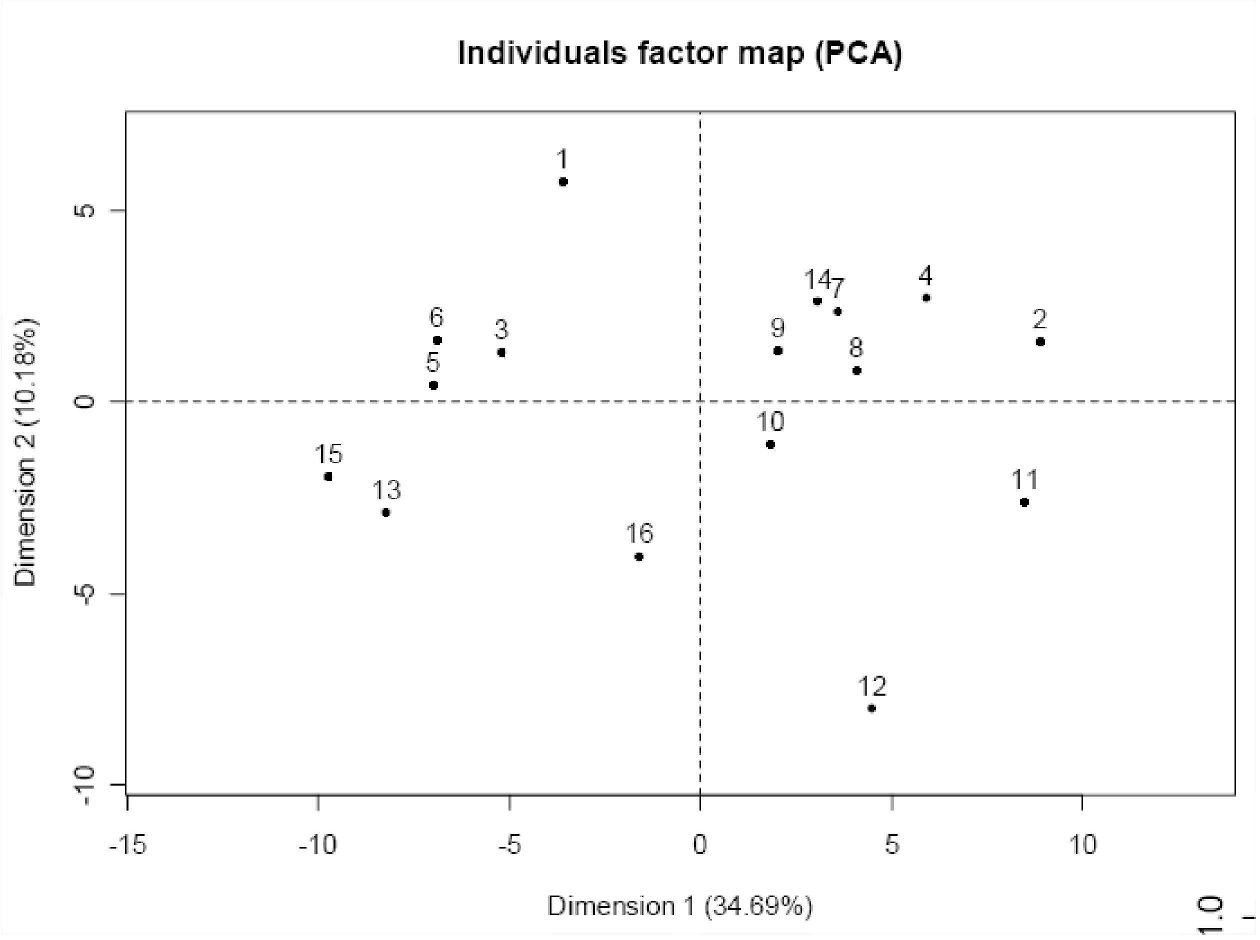
- n produits, p consommateurs, q caractéristiques
  - Tableau X de notes (n,p)
  - Tableau Y de caractéristiques (n,q)

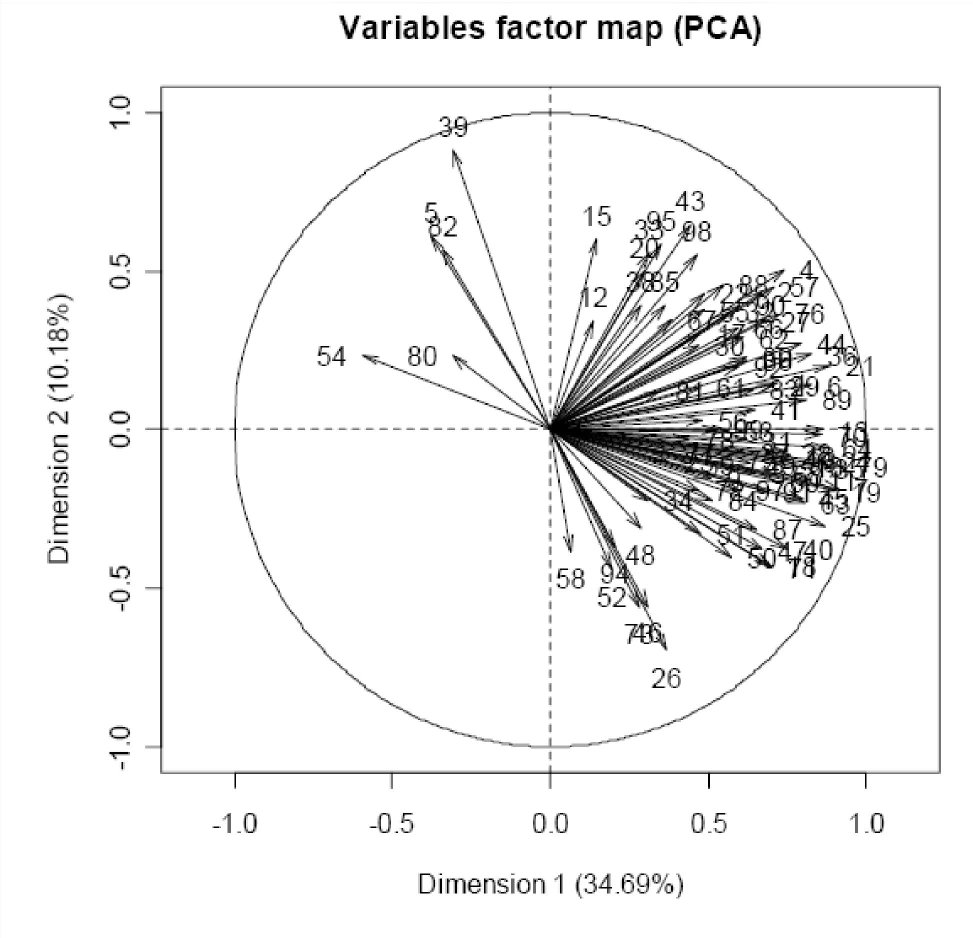
Exemple (cf SensoMineR)

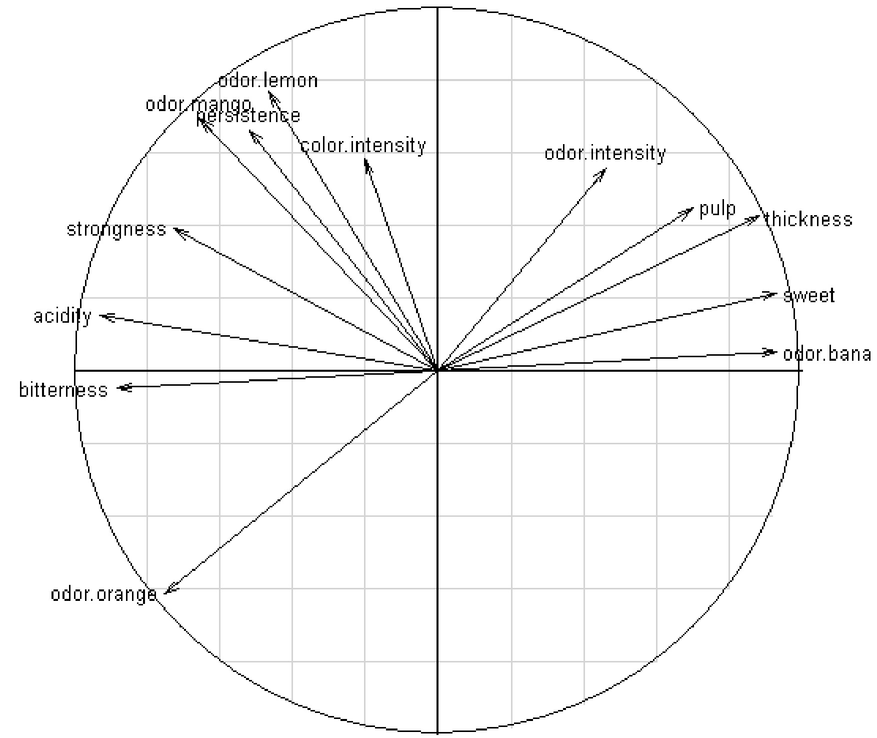
16 cocktails, jugés par 100 consommateurs et décrits par 13 variables sensorielles (fournies par des experts)

4.1 Cartographie « interne » ou MDPREF

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

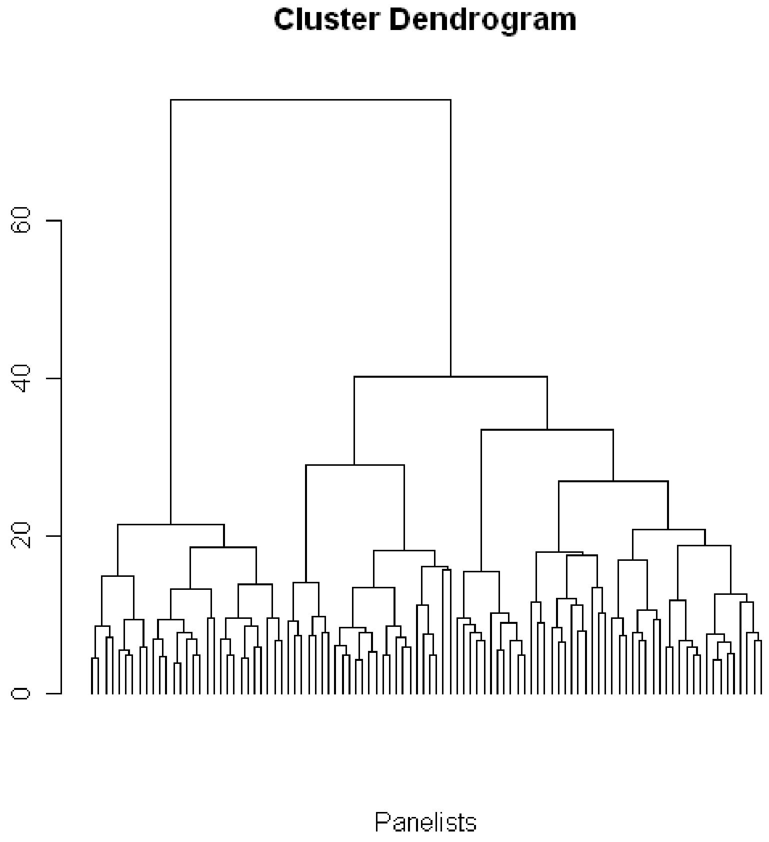


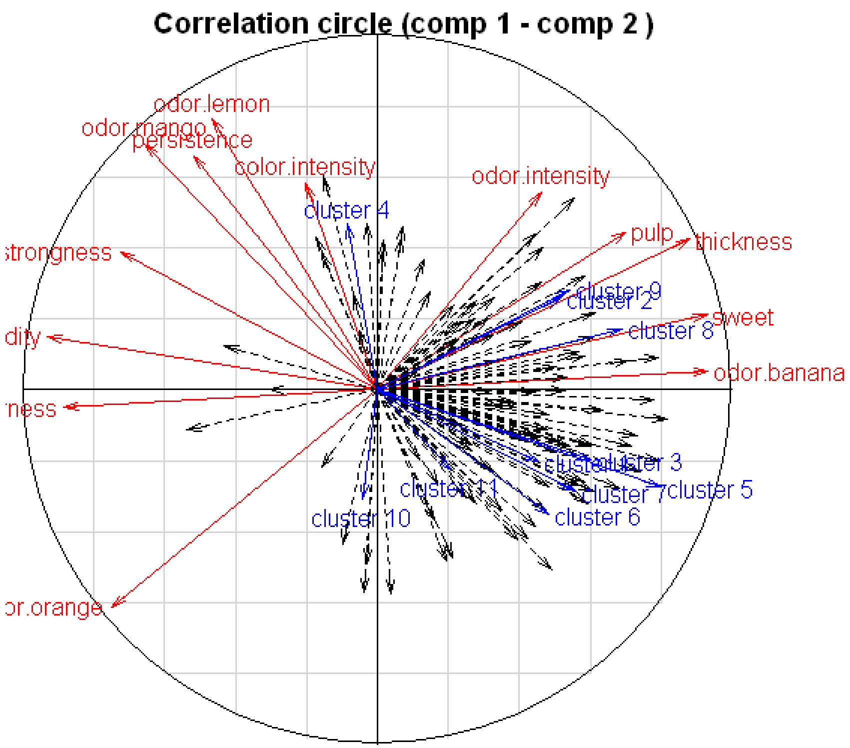




Correlation circle (comp 1 - comp 2)

- Avec classification des consommateurs

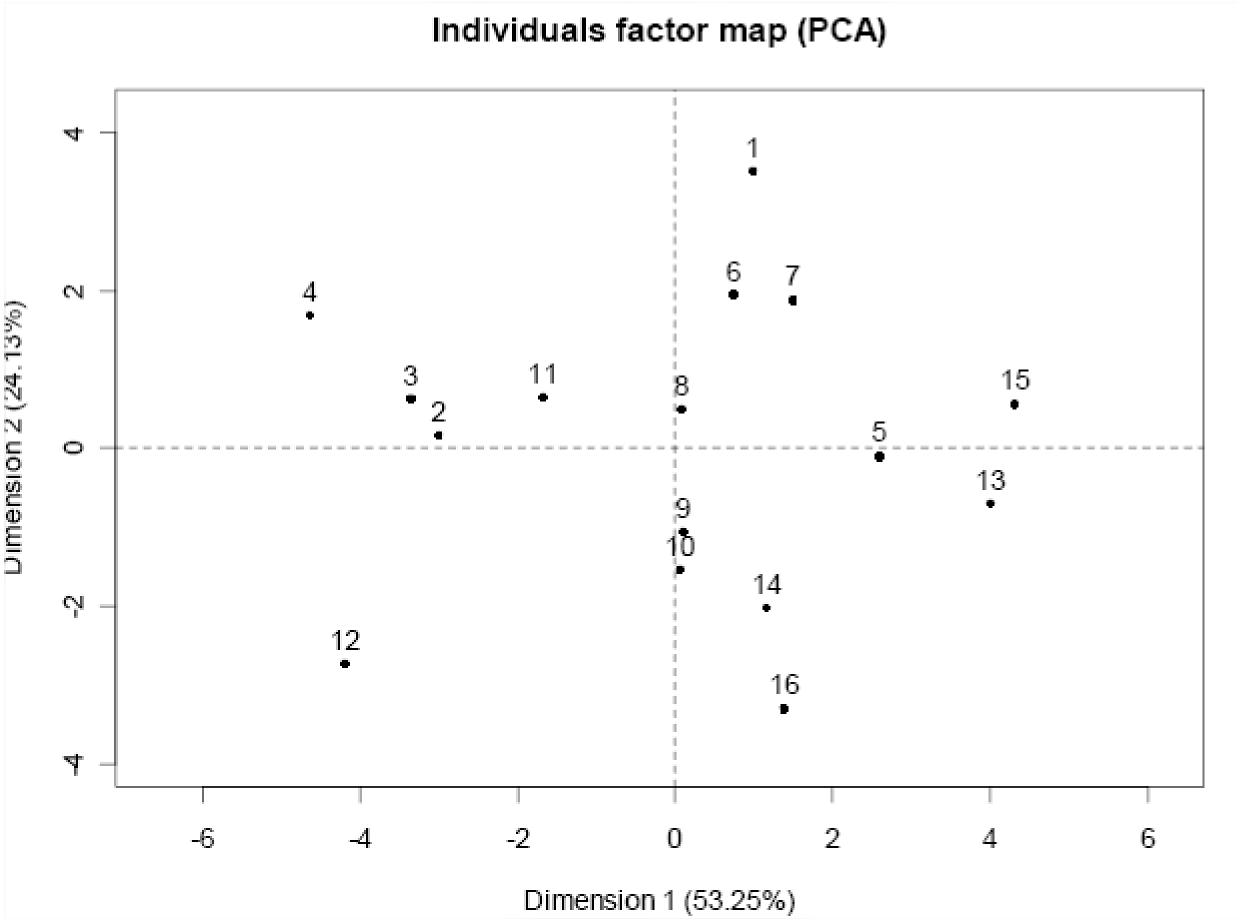


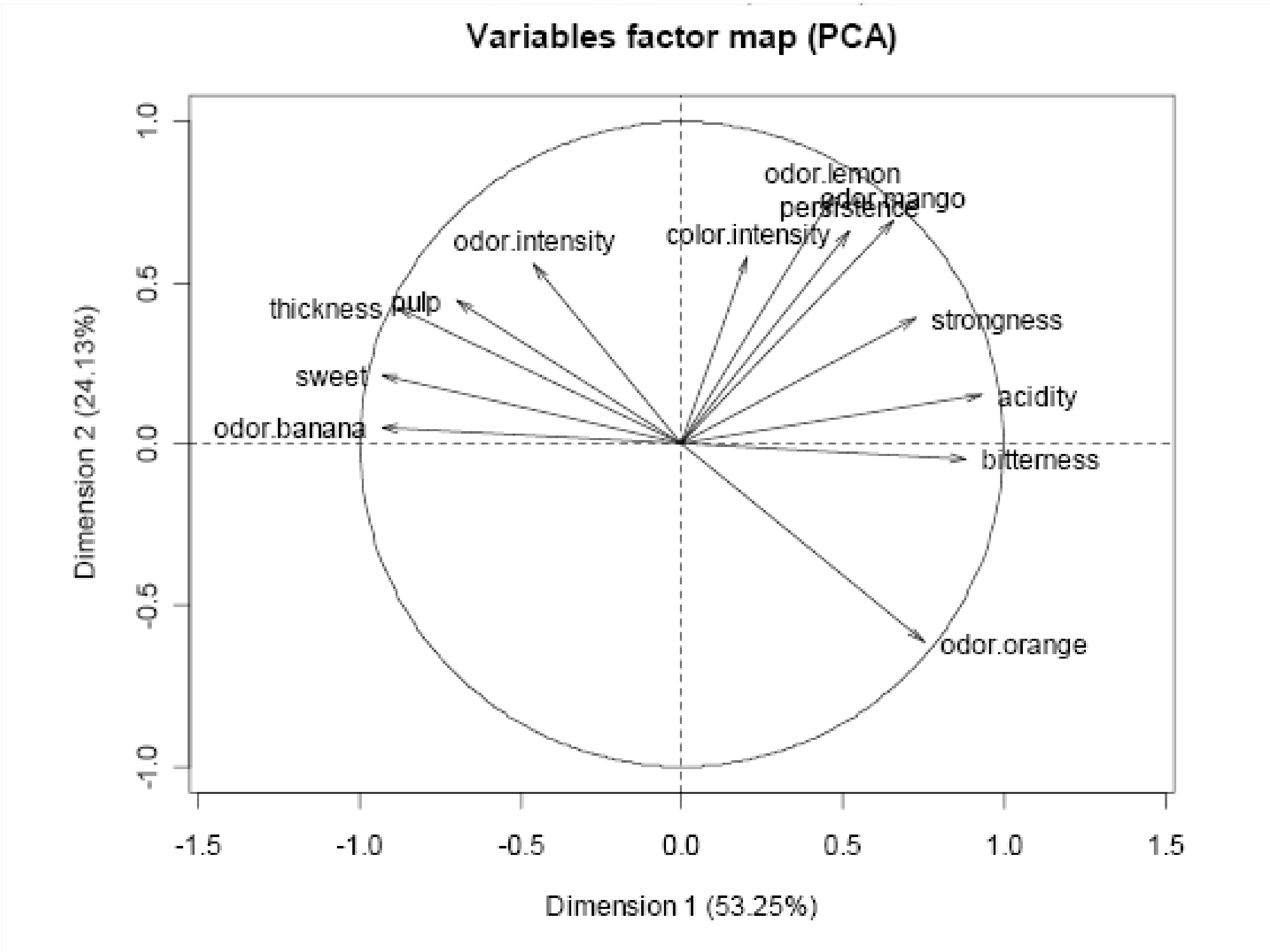


4.2 Cartographie « externe » ou PREFMAP

- 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


Individuals factor map (PCA)
Dimension 1 (53.25%)
Dimension 2 (24.13%)
-6
-4
-2
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


Variables factor map (PCA)
odor.lemon
odor.mango
persistence
color.intensity
odor.intensity
thickness
pulp
strongness
sweet
acidity
odor.banana
bitterness
odor.orange
Dimension 2 (24.13%)
Dimension 1 (53.25%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

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

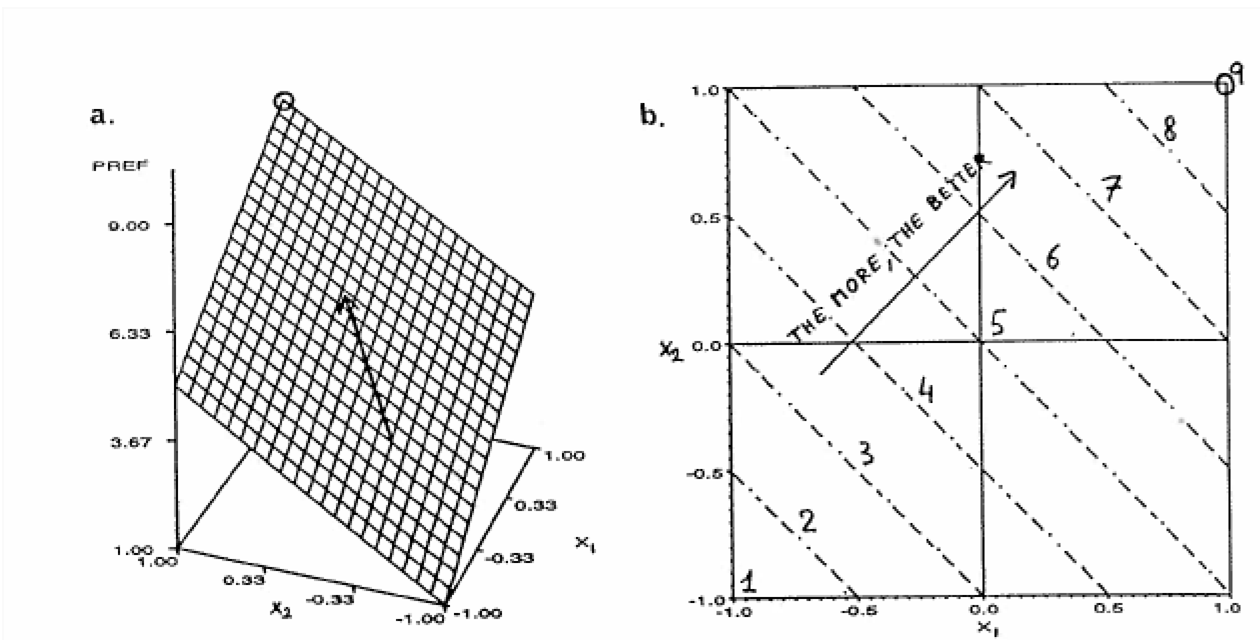


Figure 1 : Vectorial model PREF = 5 + $2X_1$ + $2X_2$; a. response surface, b. isocontours

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»

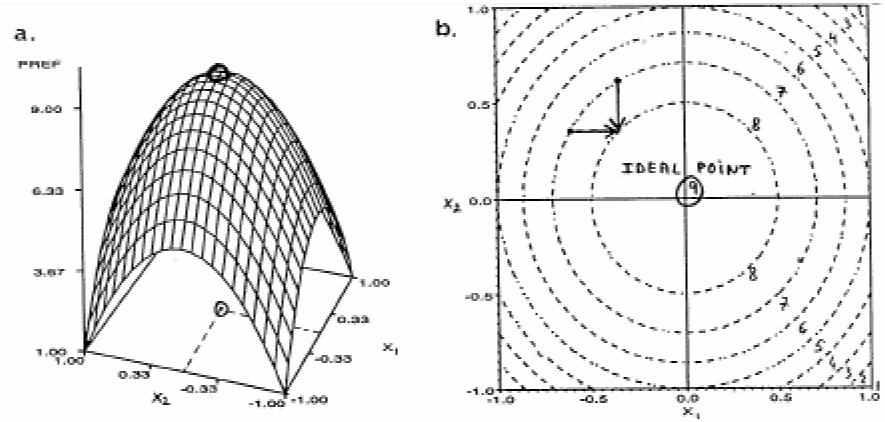


Figure 2 : Circular model PREF = 9 - $4X_1^2$ - $4X_2^2$; a. response surface, b. isocontours

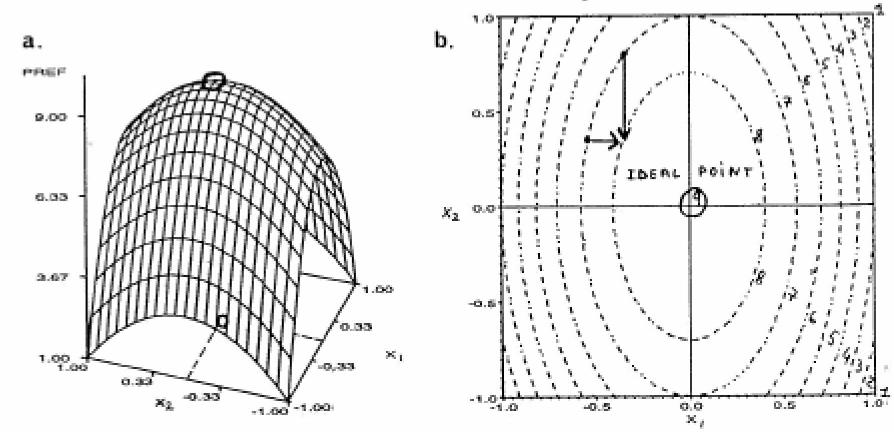
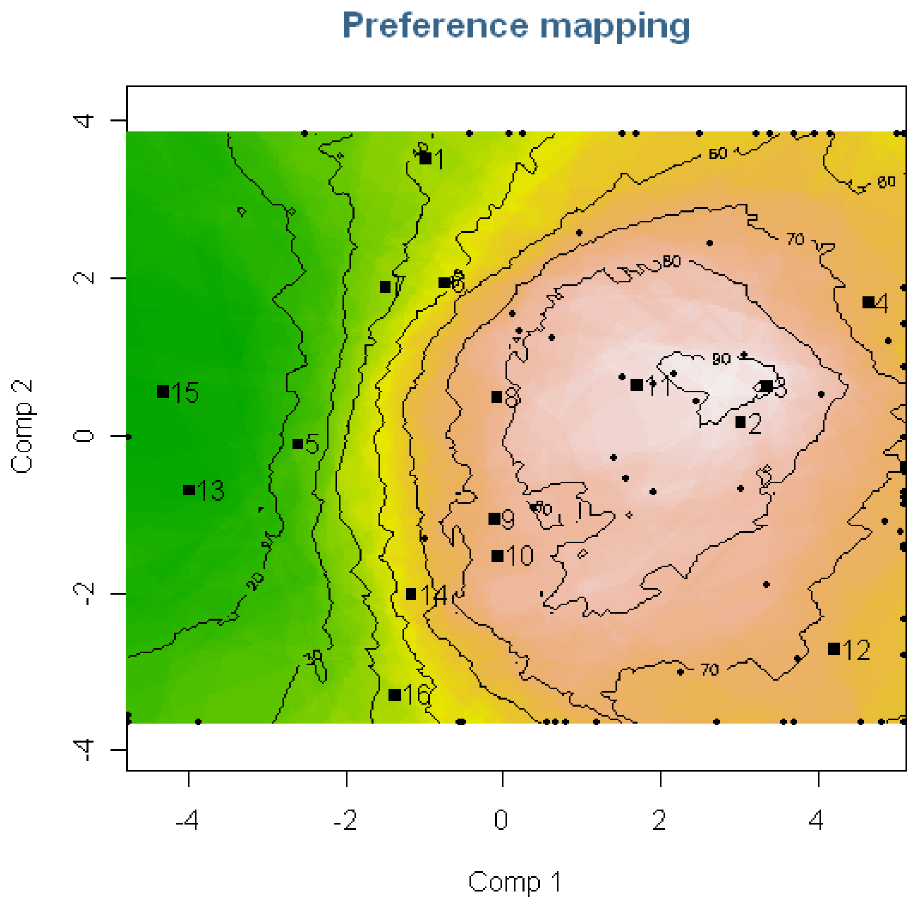


Figure 3 : Elliptical model PREF = 9 - $6X_1^2$ - $2X_2^2$; a. response surface, b. isocontours

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



Références

- d'Aubigny G. (2009). The Analysis of Proximity Data, in Govaert G. (Ed.) *Data Analysis*. Chapter 4, pp. 93-145. John Wiley and sons,
- 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