Analyse en Composantes Principales

Pr Roch Giorgi

 roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr

Objectif

- Étudier simultanément un nombre important de variables quantitatives
  - ✓ 2 variables quantitatives ⇒ nuage de points (espace de dimension 2)
  - ✓ n variables ⇒ espace de dimension n !
- Méthode pour obtenir un résumé « pertinent » des données initiales
  - ✓ Travail sur les intercorrélations entre les variables pour en extraire des dimensions plus globales
  - ✓ Revenir à un espace de dimension réduite

Exemple de Présentation (1)

- Enquête sur les consommateurs de bière sur leurs motivations d'achat (Wuensch, 2005)
- n=220
- 7 variables (cotées sur une échelle de 1 à 100)
  - ✓ Coût faible d'un pack de 6 bières
  - ✓ Quantité de la bouteille
  - ✓ Haut degré d'alcool
  - ✓ Prestige de la marque
  - ✓ Couleur de la bière
  - ✓ Arôme
  - ✓ Goût

Exemple de Présentation (2)

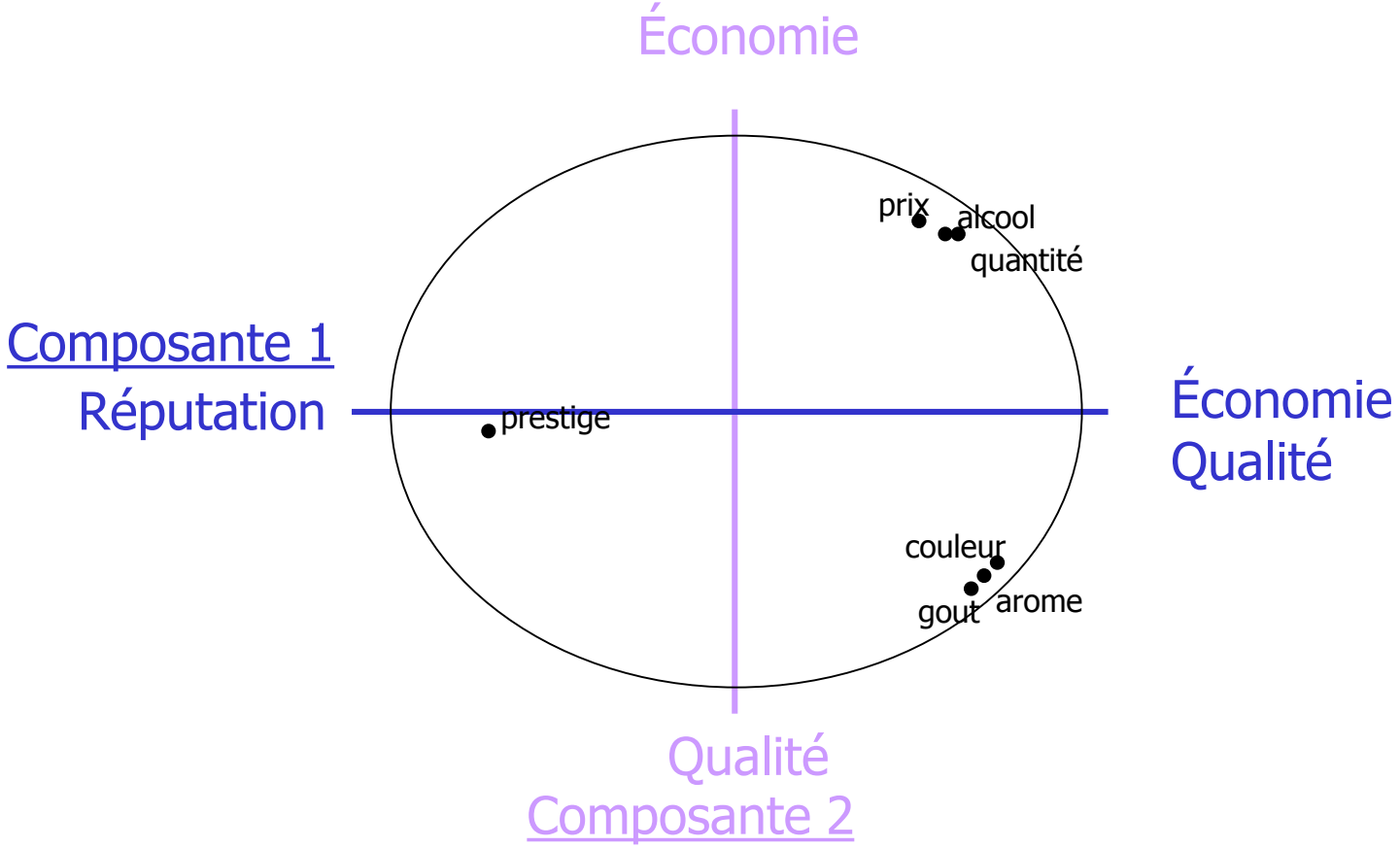
- Résultats univariés

	N	Minimum	Maximum	Moyenne	Ecart type
Prix	220	0	100	47,25	34,264
Quantité	220	0	100	43,50	33,733
Alcool	220	10	100	46,50	32,249
Prestige	220	0	100	48,25	24,204
Couleur	220	0	95	51,00	26,875
Arôme	220	0	90	44,75	25,821
Goût	220	25	100	67,25	24,111
N valide (listwise)	220				

- Comment ces informations sont liées entre elles ?
- Quel est le profil des consommateurs de bière ?
- Qu'est-ce qui motive leurs achats ?

Exemple de Présentation (3)

- De 7 variables => 2 composantes (à interpréter selon les variables)



Principe (1)

- Analyse reposant sur la dispersion des données observées (variance observée)
- Utilisation de la matrice des variances-covariances (ou des corrélations)
- Procédure mathématique pour « extraire » des composantes (transformations mathématiques des variables initiales) expliquant la structure des données

$$\text{Composante}_c = \text{coef}_1.V_1 + \text{coef}_2.V_2 + ... + \text{coef}_k.V_k$$

où les coefficients sont à estimer

Principe (2)

- Algorithme ayant 2 contraintes
  - ✓ La première composantes doit maximiser la proportion de variance expliquée dans les variables initiales (V_1, ..., V_k)

$$C_1 = \hat{a}_1^1 V_1 + \hat{a}_2^1 V_2 + \ldots + \hat{a}_k^1 V_k$$

  - ✓ Extraction des composantes suivantes indépendantes des précédentes

$$C_2 = \hat{a}_1^2 V_1 + \hat{a}_2^2 V_2 + \ldots + \hat{a}_k^2 V_k$$

$$\vdots$$

$$C_k = \hat{a}_1^k V_1 + \hat{a}_2^k V_2 + \ldots + \hat{a}_k^k V_k$$

Principe (3)

- La proportion de variance expliquée diminue avec le nombre de composantes
- La proportion de variance totale cumulée pourra atteindre 100 %
  - ✓ Nombre de composantes extraites = nombre de variables
- Critères statistiques pour déterminer le nombre de composantes pertinentes à extraire
- Réalisation de graphiques dans l'espace de dimension défini par le nombre de composantes extraites
  - ✓ Représentation des distances euclidiennes inter individus
  - ✓ Représentation des corrélations inter variables

Exemple de Présentation (suite)

- Enquête sur les consommateurs de bière sur leurs motivations d'achat (Wuensch, 2005)
- n=220
- 7 variables (cotées sur une échelle de 1 à 100)
  - ✓ Coût faible d'un pack de 6 bières
  - ✓ Quantité de la bouteille
  - ✓ Haut degré d'alcool
  - ✓ Prestige de la marque
  - ✓ Couleur de la bière
  - ✓ Arôme
  - ✓ Goût

Résultats Univariés

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Prix	220	0	100	47,25	34,264
Quantité	220	0	100	43,50	33,733
Alcool	220	10	100	46,50	32,249
Prestige	220	0	100	48,25	24,204
Couleur	220	0	95	51,00	26,875
Arôme	220	0	90	44,75	25,821
Goût	220	25	100	67,25	24,111
N valide (listwise)	220				

Matrice de Corrélation (1)

Matrice de corrélation

		Prix	Quantité	Alcool	Prestige	Couleur	Arôme	Goût
Corrélation	Prix	1,000	,832	,767	-,406	,018	-,046	-,064
	Quantité	,832	1,000	,904	-,392	,179	,098	,026
	Alcool	,767	,904	1,000	-,463	,072	,044	,012
	Prestige	-,406	-,392	-,463	1,000	-,372	-,443	-,443
	Couleur	,018	,179	,072	-,372	1,000	,909	,903
	Arôme	-,046	,098	,044	-,443	,909	1,000	,870
	Goût	-,064	,026	,012	-,443	,903	,870	1,000

- Coefficients de corrélation linéaire des variables prises 2 à 2
  - ✓ Diagonale ne comporte que des 1
  - ✓ Matrice symétrique par rapport à la diagonale
- Coefficients
  - ✓ Positifs ou négatifs
  - ✓ Faibles ou forts
  - ✓ De –1 à +1
- Un certain degré d'intercorrélation est nécessaire pour extraire une composante correspondant à une fonction linéaire des variables initiales

Matrice de Corrélation (2)

- La matrice ne doit pas être singulière
  - ✓ Matrice singulière
    - Au moins une variable est parfaitement corrélée avec une autre variable ou avec une combinaison de plusieurs variables (ex : total des scores)
    - Déterminant = 0
  - ✓ Son déterminant doit être ≠ de 0 (> 0,00001)
- La matrice ne doit pas être la matrice d'identité
  - ✓ Matrice d'identité
    - Matrice ne comportant que des 0 sauf des 1 sur la diagonale
    - Déterminant = 1
  - ✓ Son déterminant doit être ≠ de 1
  - ✓ Testé par le test de sphéricité de Barlett

Matrice de Corrélation (3)

- Chaque variable doit être en relation avec les autres variables
  - ✓ Inspection visuelle de la matrice de corrélation
  - ✓ Mesure d'adéquation de l'échantillonnage de Kaiser-Meyer-Olkin
    - Calculées pour chacune des variables et pour la matrice globale
    - Doivent être ≥ 0,5

	Indice d'adéquation de KMO
Couleur	0,779
Arôme	0,550
Prestige	0,630
Goût	0,763
Prix	0,590
Alcool	0,801
Quantité	0,676
Global	0,665

Extraction des Composantes Principales (1)

- Critère de Kaiser
  - ✓ Repose sur la notion de la valeur propre (eigenvalue) ⇔ variance de chaque CP

	Valeurs propres initiales		
Composante	Total	% de la variance	% cumulés
1	3,313	47,327	47,327
2	2,616	37,369	84,696
3	0,575	8,209	92,905
4	0,24	3,427	96,332
5	0,134	1,921	98,252
6	0,085	1,221	99,473
7	0,037	0,527	100
Total	7	100	

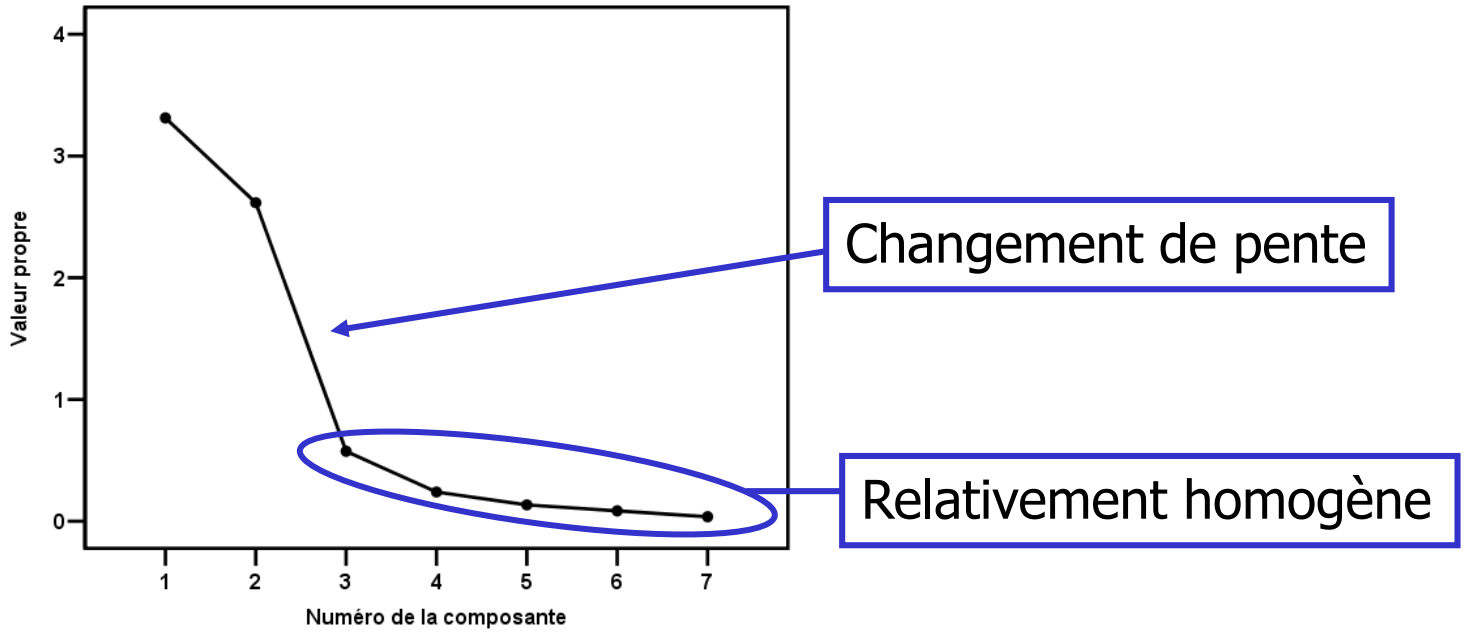
Réduire les données de 7 variables à 2 composantes permet de rendre compte de 84,7% de la variance initiale

Chaque variable à 1 unité de variance

Ne pas retenir les composantes dont la VP < 1

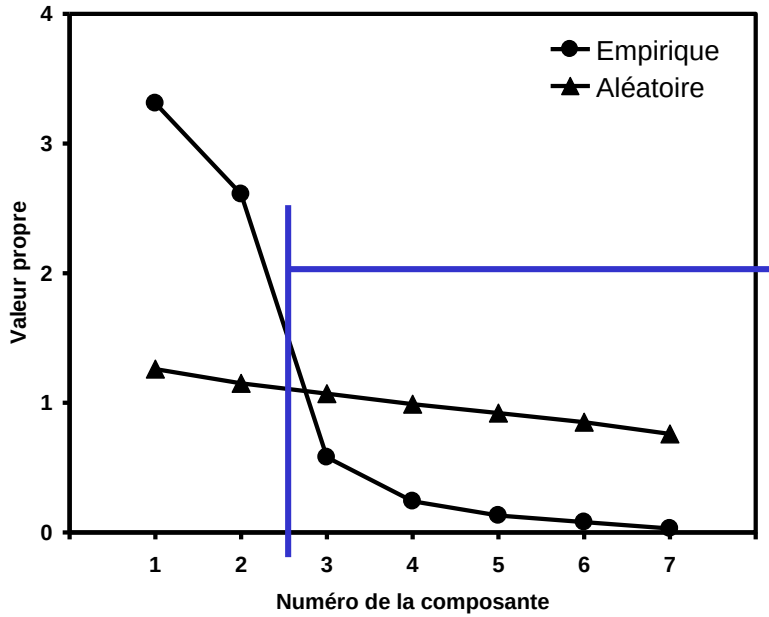
Extraction des Composantes Principales (2)

- Test d'accumulation de variance de Cattel (scree test)
  - ✓ Critère graphique
  - ✓ Arrêt de l'extraction des composantes déterminé par le changement de pente



Extraction des Composantes Principales (3)

- Analyse parallèle de Horn
  - ✓ Possibilité de découvrir par chance une composante pouvant expliquer une proportion de variance
  - ✓ Comparaison des VP empiriques et celles obtenues sur une matrice de corrélation générée au hasard en fonction du nombre d'individus et du nombre de variables initiales



Composantes dont les variances sont supérieures à celles obtenues par chance

Extraction des Composantes Principales (4)

- Facilité à interpréter les facteurs à extraire
  - ✓ Difficultés d'interprétation si trop de composantes
  - ✓ Mais une surestimation du nombre de composantes est moins dommageable qu'une sous-estimation

- Confrontation des méthodes pour décider

Matrice des Composantes (1)

- Extraction de 2 composantes (dans notre exemple)
  - ✓ Orthogonales, donc indépendantes
- Matrice des poids des composantes
- Correspond aux coefficients de corrélation entre les variables et les composantes

$$V_i = \hat{b}_i^1 C_1 + \hat{b}_i^2 C_2 + U_i$$

Où U est la variance non explicable pour aucune des composantes

Matrice des Composantes (2)

	Composante	
	1	2
Couleur	0,760	-0,576
Arôme	0,736	-0,614
Prestige	-0,735	-0,071
Goût	0,710	-0,646
Prix	0,550	0,734
Alcool	0,632	0,699
Quantité	0,667	0,675

$\Rightarrow$ Prix = $0{,}550.C_1 + 0{,}734.C_2$

↗ de 1 unité de $C_1 \Rightarrow$ ↗ de 0,550 dans le prix
↗ de 1 unité de $C_2 \Rightarrow$ ↗ de 0,734 dans le prix

Corrélation(C_1, Prix) = 0,550
Variance commune(C_1, Prix) = $(0{,}550)^2$=30,23%
Variance commune(C_2, Prix) = $(0{,}734)^2$=54,18%
84,41% de la variable prix expliquée par C_1 et C_2

Proportion de Variance Commune

	Composante	
	1	2
Couleur	0,760	-0,576
Arôme	0,736	-0,614
Prestige	-0,735	-0,071
Goût	0,710	-0,646
Prix	0,550	0,734
Alcool	0,632	0,699
Quantité	0,667	0,675

Qualité de représentation

	Initial	Extraction
Prix	1	0,842
Quantité	1	0,901
Alcool	1	0,889
Prestige	1	0,546
Couleur	1	0,910
Arôme	1	0,918
Goût	1	0,922

Prix = $(0,550)^2 + (0,734)^2$ = 84,2% de variance commune

Quantité = $(0,667)^2 + (0,675)^2$ = 90,1% de variance commune

Prestige = $(-0,735)^2 + (-0,071)^2$ = 54,6% de variance commune
et donc (1 - 0,546)=45,7% de variance unique

Matrice des Composantes (3)

	Composante	
	1	2
Couleur	0,760	-0,576
Arôme	0,736	-0,614
Prestige	-0,735	-0,071
Goût	0,710	-0,646
Prix	0,550	0,734
Alcool	0,632	0,699
Quantité	0,667	0,675

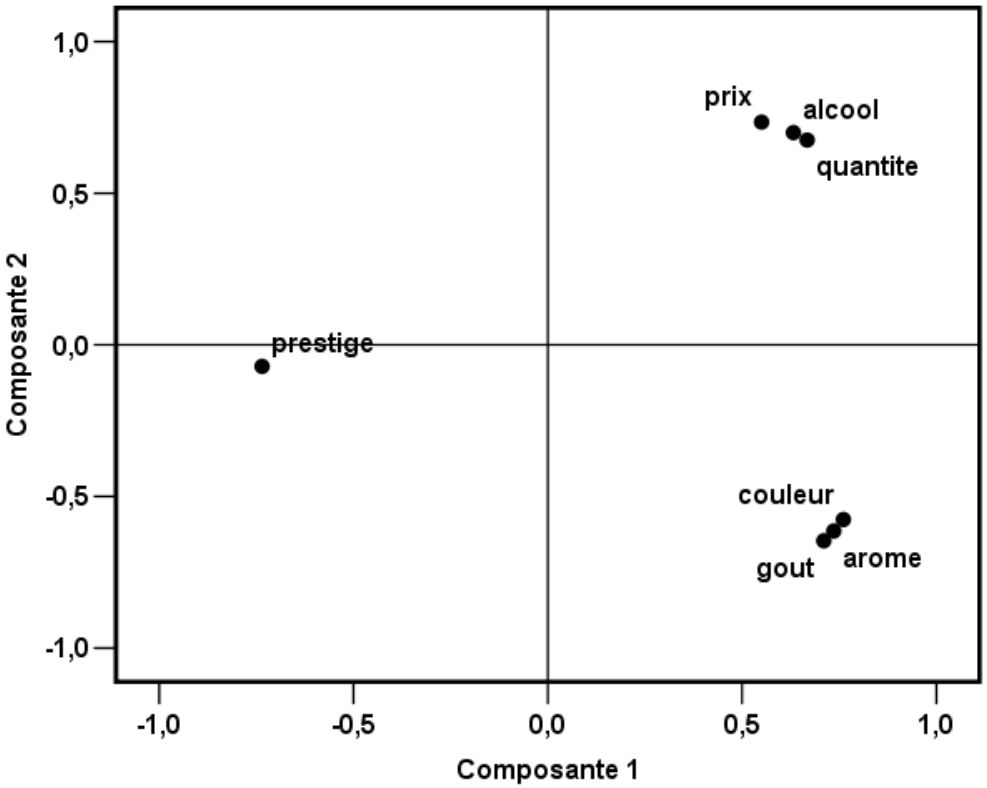
Expression des variables sur chacune des composantes

C_1	Prestige	Les autres
C_2	Couleur Arôme Goût	Prix Alcool Quantité

Graphe des Composantes

Économie

Réputation



Économie
Qualité

Qualité

Rotation des Axes

- Identification des axes souvent délicate
- Plusieurs variables peuvent avoir une pondération importante sur C_1 (inhérent à la technique)
- Transformation de la solution obtenue par une rotation des axes définissant les composantes
  - ✓ Préserve la variance expliquée de chaque variable
  - ✓ Réassignée à des composantes transformées
- Rotation orthogonale
  - ✓ VARIMAX : minimise le nombre de variables ayant une forte corrélation sur C_1
- Rotation oblique
  - ✓ Modifie les angles formés par les axes
  - ✓ Introduit une corrélation entre les composantes

Rotation des Axes : VARIMAX (1)

	Composante	
	1	2
Prix	,550	,734
Quantité	,667	,675
Alcool	,632	,699
Prestige	-,735	-,071
Couleur	,760	-,576
Arôme	,736	-,614
Goût	,710	-,646

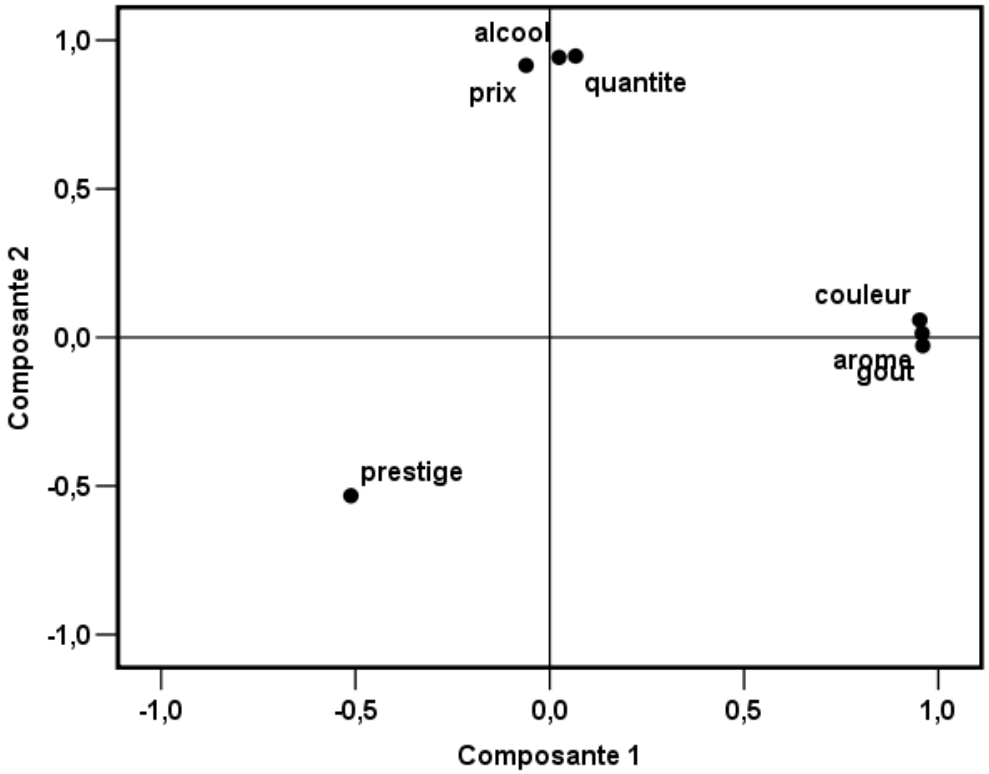
Sans rotation

	Composante	
	1	2
Prix	-,061	,916
Quantité	,066	,947
Alcool	,024	,942
Prestige	-,512	-,533
Couleur	,952	,058
Arôme	,958	,014
Goût	,960	-,028

Rotation VARIMAX

Rotation des Axes : VARIMAX (2)

Buveur



Dégustateur

Remarques

- Nombre de sujets / Nombre de variable
  - ✓ n < 100
  - ✓ 5 à 10 fois plus de sujets que de variables analysées
  - ✓ Rapport surtout fonction du nombre de communalité des variables utilisées et de la détermination des composantes
- Contrôler la matrice de corrélation
- Critères multiples pour décider du nombre de composantes à extraire
- Essayer des rotations, orthogonales puis obliques

Sources

- Baillargeon J. L'analyse en composantes principales (2003) http://www.uqtr.ca/cours/srp-6020/acp/acp.pdf
- Wuensch K.L. Principal component analysis (2004) http://core.ecu.edu/psyc/wuenschk/MV/FA/PCA.doc)
- Besse P., Baccini A. Data mining I - Exploration statistique (2005) http://www.lsp.ups-tlse.fr/Besse/pub/Explo_stat.pdf
- Dunlap W.P. Logiciel pour générer des valeurs propres aléatoires. http://www.tulane.edu/~dunlap/psylib/pa.exe