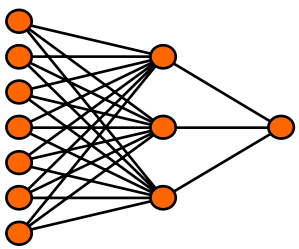
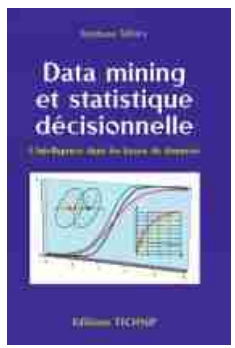
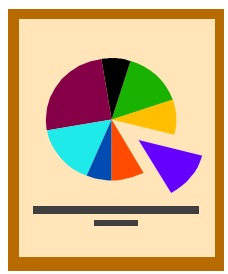
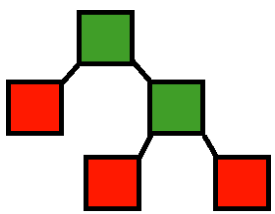
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









Plan du cours

- Qu'est-ce que le data mining ?
- À quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs - Consulting
- *L'analyse et la préparation des données : l'analyse factorielle*
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le web mining
- Le text mining

L'analyse factorielle des correspondances

L'analyse des correspondances

- L'**analyse factorielle des correspondances** (AFC) (Guttman, 1941 – Benzécri, 1973) offre une visualisation en 2 dimensions des tableaux de contingence :
  - deux modalités liées positivement (sureffectif) sont proches
  - deux modalités liées négativement (sous-effectif) sont opposées
  - on utilise une métrique (du χ^2) pondérant ces sur- ou sous-effectifs par l'inverse de la fréquence globale d'une modalité (non croisée avec la modalité de l'autre variable)
  - les + fortes oppositions sont sur l'axe horizontal
  - les modalités non liées aux autres sont au centre
- L'**analyse des correspondances multiples** (ACM) s'applique à plus de 2 variables catégorielles
  - en utilisant le tableau de contingence multiple (tableau de Burt)
  - ou le tableau disjonctif

Tableaux utilisés en ACM

- Tableau de contingence multiple (= tableau de Burt)
  - les lignes et les colonnes correspondent aux modalités des variables
  - croisement des 2 mêmes modalités : nb d'individus possédant cette modalité
  - croisement de 2 modalités différentes, appartenant à des variables différentes : nb d'individus possédant à la fois la $1^{ère}$ et la 2^{de} modalité
- Tableau disjonctif
  - une ligne par individu
  - une colonne par modalité x_{jk} de variable X_j
  - croisement de la i^e ligne et de la colonne correspondant à la modalité x_{jk} : 1 si $X_j(i^e \text{ individu}) = x_{jk}$, et 0 sinon

Intérêt de l'AFC et l'ACM 1/2

- L'AFC et l'ACM s'appuient sur un instrument puissant de détection des groupes d'individus et des individus isolés : l'œil. La perception visuelle est inégalable pour la reconnaissance des formes. Examiner un plan factoriel est :
  - + efficace que d'examiner tous les plans (x,y) des variables d'origine
  - + rapide que dépouiller tous les tableaux de contingence
- Représentation simultanée des individus et des modalités
  - sur un même plan
- Possibilité de visualiser certaines variables (« supplémentaires ») sans les prendre en compte dans le calcul des correspondances :
  - variables qu'on veut lier aux variables actives mais pas lier entre elles
  - ou variables qu'on veut expliquer par les variables actives

Intérêt de l'AFC et l'ACM 2/2

- Détection des liaisons entre les variables
  - même non linéaires (liaisons de d° > 1)
  - pas forcément visibles sur un tableau de contingence
- Filtrage des fluctuations aléatoires des données
  - en remplaçant les variables d'origine par les 1ers axes factoriels
  - utile avant une classification ou un réseau de neurones
- Transformation de var. qualitatives en var. quantitatives
  - en remplaçant les variables qualitatives par les coordonnées de leurs modalités sur les axes factoriels
  - transformation utilisée dans l'analyse discriminante DISQUAL
- Permet de traiter simultanément les variables quantitatives (en les discrétisant) et qualitatives

AFC : signification des valeurs propres

- L'inertie du nuage de points mesure la dispersion du nuage et la liaison entre les 2 variables. On a :
  - inertie x effectif = χ^2 du tableau
  - inertie = somme des valeurs propres $\neq 0$
- Si on a k valeurs propres proches de 1 => le nuage de points (resp. tableau) est scindé en k+1 groupes (blocs)
- Si toutes les valeurs propres sont $\approx$ 1, chaque modalité de V_1 correspond (presque) à une seule modalité de V_2
- Si V_1 et V_2 sont partiellement corrélées (« vont dans le même sens »), le tableau peut être réordonné en un tableau avec des 0 partout sauf dans une bande entourant la diagonale, et le nuage de points a une forme parabolique dite « en fer à cheval » (effet Guttman)

Représentation graphique 1/2

- Propriété de l'analyse des correspondances : on peut superposer individus et modalités sur le même graphique
  - le point représentant une modalité est le barycentre des individus possédant cette modalité
- Le centre correspond au profil moyen pour les 2 variables
- Une modalité A est loin du centre si la distribution des modalités de l'autre variable est très différente dans l'ensemble des individus vérifiant A et dans l'ensemble de tous les individus

Représentation graphique 2/2

- Proximité de 2 modalités A_1 et A_2 de la même variable
  - si l'autre variable a la même distribution restreinte à A_1 et A_2
- Proximité de 2 modalités A et B de 2 variables différentes
  - si sureffectif de l'intersection AxB
  - NB : ceci est souvent vrai mais pas toujours
- Avant d'interpréter la proximité de 2 modalités, s'assurer qu'elles sont bien représentées sur le plan factoriel

Interprétation des résultats 1/2

- Sélectionner les axes factoriels dont la valeur propre :
  - est ≠ de 0 et 1 (critère obligatoire)
  - est > 1/p (p = nb de variables)
  - correspond à un coude dans le diagramme des valeurs propres
  - on dépasse rarement les 5 premiers axes
- Coordonnée d'une modalité sur un axe
  - on s'intéresse parfois aux modalités dont la coordonnée > racine carrée (valeur propre de l'axe)
- Contribution d'une modalité à un axe
  - % de l'inertie de la modalité dans l'inertie de l'axe
  - on explique un axe par ses modalités à forte contribution

Interprétation des résultats 2/2

- Cosinus² (qualité de la projection sur un axe)
  - % pris par l'axe dans la dispersion de la modalité
  - plus cos² est proche de 1, meilleure est la représentation d'une modalité sur un axe
  - il faut n'apprécier la proximité de 2 modalités sur un axe que si elles ont un cos² assez grand sur cet axe
  - la somme des cos² sur l'ensemble des axes vaut 1
- Variables supplémentaires
  - ne contribuent pas aux axes
  - mais leurs cos² peuvent être analysés

Les 2 types de fichiers traités par la proc CORRESP de SAS 1/2

- Type 1 : le fichier contient des données « brutes »
  - format le plus courant dans les bases de données
    - chaque ligne du fichier correspond à un individu
    - chaque colonne correspond à une variable catégorielle caractérisant les individus
    - chaque intersection « ligne x colonne » contient la modalité de la variable pour l'individu
  - utiliser l'instruction « TABLES »
    - qui crée un tableau de contingence, de Burt ou disjonctif selon les options
    - puis effectue l'AFC ou l'ACM

Les 2 types de fichiers traités par la proc CORRESP de SAS 2/2

- Type 2 : le fichier contient des données « matricielles »
  - les données sont déjà sous forme d'un tableau de contingence, de Burt ou disjonctif
    - chaque ligne du fichier correspond à une ligne du tableau
    - chaque colonne correspond à une modalité A_i spécifiée dans l'instruction « VAR »
    - chaque intersection « ligne x colonne » contient l'effectif d'une cellule du tableau
      - cette intersection est une donnée numérique ≥ 0
      - si < 0 ou manquante $\Rightarrow$ la ligne n'entre pas dans l'analyse
  - utiliser l'instruction « VAR »
    - qui effectue directement l'AFC ou l'ACM
  - les lignes Y sont spécifiées par l'instruction ID Y
    - sauf pour le tableau de Burt qui est symétrique

Traitement des données matricielles

- VAR A_1 … A_m;
- ID Y;
  - avec un tableau de Burt $\Rightarrow$ utiliser l'option MCA et spécifier le nb de variables par NVARS = n
  - avec un tableau disjonctif ou de contingence, on ne spécifie pas d'option (ni MCA ni BINARY)
  - pour un tableau disjonctif ou de contingence, les lignes sont identifiées par la variable Y
    - Y = variable en ligne pour un tableau de contingence
    - Y = label pour un tableau disjonctif
  - pas d'instruction ID Y possible pour un tableau de Burt
    - car dans ce cas les lignes et les colonnes sont les mêmes et sont toutes spécifiées dans l'instruction VAR
- Variables supplémentaires
  - à énumérer dans l'instruction SUPPLEMENTARY
  - à énumérer aussi dans l'instruction VAR

Exemple

Nom	**Sexe**	**État**
M. Dupont	Homme	Marié
Mme Dupont	Femme	Marié
M. Durand	Homme	Célibataire
M. Dubois	Homme	Marié
Mme Dubois	Femme	Marié
Mme Martin	Femme	Veuf

Exemple de syntaxe SAS sur tableau de Burt (donné)

- `proc corresp data=sasuser.tableau mca nvars=2 out=sortie all;`
- `var homme femme marié célibataire veuf;`
- `run;`

	Femme	Homme	Marié	Célibataire	Veuf
Femme	3	0	2	0	1
Homme	0	3	2	1	0
Marié	2	2	4	0	0
Célibataire	0	1	0	1	0
Veuf	1	0	0	0	1

- NVARS = 2 indique que les 5 modalités du tableau de Burt proviennent ici de 2 variables
- Attention : les modalités doivent être listées dans le même ordre que dans le fichier

Contenu du fichier en sortie

- Le fichier OUT contient des résultats (inertie, coordonnées sur les axes factoriels, cos²…) sur :
  - les modalités de variables suivant la commande VAR (_type_ = « VAR » dans le fichier)
  - les lignes correspondant à l'instruction ID (_type_ = « OBS »), sauf pour le tableau de Burt qui n'a pas d'enregistrements de _type_ = « OBS »
  - seul le tableau disjonctif (ID LABEL) permet d'obtenir des résultats sur les individus
  - mais les résultats sur les modalités (_type_ = « VAR ») sont les mêmes avec un tableau disjonctif ou de Burt
  - les coordonnées contenues dans le fichier OUT permettent de représenter sur un plan factoriel les modalités et éventuellement (si tableau disjonctif) les individus

Exemple de syntaxe SAS sur tableau disjonctif complet (donné)

- `proc corresp data=sasuser.tableau out=sortie all;`
- `var homme femme marié célibataire veuf;`
- `id label;`
- `run;`

	Femme	Homme	Marié	Célibataire	Veuf
1	0	1	1	0	0
2	1	0	1	0	0
3	0	1	0	1	0
4	0	1	1	0	0
5	1	0	1	0	0
6	1	0	0	0	1

Exemple de syntaxe SAS sur tableau de contingence (donné)

```
proc corresp data=sasuser.tableau
out=sortie all;
var marié célibataire veuf;
id sexe;
run;
```

	marié	célibataire	veuf
femme	2	0	1
homme	2	1	0

Traitement des données brutes

- TABLES X_1 … X_m , Y_1 … Y_n;
  - crée et analyse le tableau de contingence simple des variables X_1 … X_m (en lignes) par Y_1 … Y_n (en colonnes)
  - ne pas oublier la virgule séparant les variables !
  - cas particulier de l'AFC : TABLES X , Y
- TABLES X_1 … X_m;
  - crée et analyse le tableau de Burt de X_1 … X_m (option « MCA » : multiple correspondence analysis)
  - ou crée et analyse le tableau disjonctif (option « BINARY »)
- Variables X_i et Y_i catégorielles
- Variables supplémentaires
  - à énumérer dans l'instruction SUPPLEMENTARY
  - à énumérer aussi dans l'instruction TABLES

Exemple de syntaxe SAS sur tableau de Burt ou disjonctif (créé)

- **proc corresp** data=sasuser.enquete mca dimens=**5** out=sortie all;
- tables sexe age frequence rayon satisfaction;
- supplementary satisfaction;
- **run**;

mca : tableau de Burt

binary : tableau disjonctif

Fichier contenant pour les modalités en colonnes (_type_=« var ») et éventuellement en lignes (_type_=« obs ») du tableau étudié :
-leur inertie
-leurs coordonnées sur les axes
-leurs contributions aux axes
-leurs cos²

nb d'axes demandés (2 par défaut)

- %plotit(data=sortie,datatype=corresp, plotvars=dim2 dim1)

all : tout afficher - short : affichage minimal

observed : affichage du tableau analysé

Contenu du fichier en sortie

- Le fichier OUT contient des résultats (inertie, coordonnées sur les axes factoriels, cos²…) sur :
  - les modalités de variables suivant la commande TABLES, à l'exception de celles à gauche de la virgule le cas échéant (_type_ = « VAR » dans le fichier)
  - les individus en cas de tableau disjonctif (option BINARY) (_type_ = « OBS » dans le fichier)
  - les modalités des variables en ligne (à gauche de la virgule) en cas de tableau de contingence (_type_ = « OBS »)
  - pas d'enregistrement de _type_ = « OBS » en cas de tableau de Burt (option MCA)
  - mais les résultats sur les modalités (_type_ = « VAR ») sont les mêmes avec un tableau disjonctif ou de Burt
  - les coordonnées contenues dans le fichier OUT permettent de représenter sur un plan factoriel les modalités et éventuellement (si tableau disjonctif) les individus

Diagrammes des valeurs propres

Inertia and Chi-Square Decomposition					
Singular Value	**Principal Inertia**	**Chi-Square**	**Percent**	**Cumulative Percent**	2 4 6 8 10 ----+----+----+----+----+---
0.56420	0.31833	759.10	8.49	8.49	**********
0.56043	0.31408	748.97	8.38	16.86	**********
0.53721	0.28860	688.21	7.70	24.56	**********
0.52911	0.27996	667.61	7.47	32.03	**********
0.52401	0.27459	654.80	7.32	39.35	**********
0.51099	0.26111	622.66	6.96	46.31	**********
0.50151	0.25151	599.77	6.71	53.02	**********
0.49404	0.24408	582.04	6.51	59.53	**********
0.49052	0.24061	573.77	6.42	65.94	**********
0.48516	0.23538	561.31	6.28	72.22	**********
0.47999	0.23039	549.40	6.14	78.36	**********
0.47222	0.22299	531.75	5.95	84.31	**********
0.46499	0.21622	515.61	5.77	90.08	**********
0.44178	0.19517	465.41	5.20	95.28	**********
0.42071	0.17700	422.09	4.72	100.00	**********
Total	3.75000	8942.50	100.00		
Degrees of Freedom = 324					

Coordonnées des modalités sur les axes

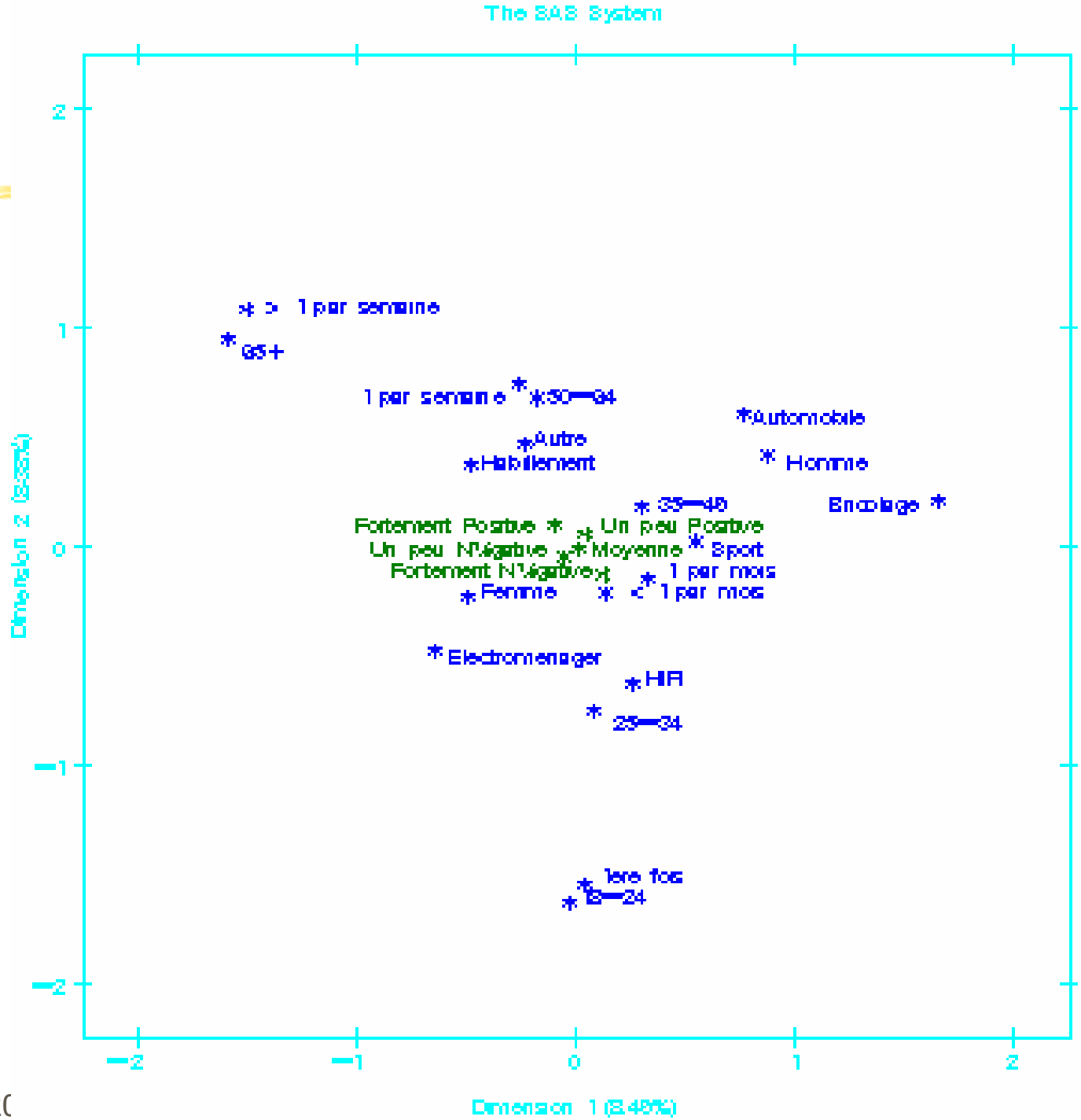
Column Coordinates					
	Dim1	**Dim2**	**Dim3**	**Dim4**	**Dim5**
Femme	-0.4925	-0.2347	-0.1592	-0.1355	-0.1049
Homme	0.8789	0.4189	0.2842	0.2418	0.1872
18-24	-0.0312	-1.6313	0.0627	1.0440	1.5625
25-34	0.0824	-0.7573	0.6018	-0.1037	-0.7778
35-49	0.2980	0.1770	-0.5109	-0.0344	-0.2147
50-64	-0.1814	0.6814	0.2733	-0.5462	0.6680
65+	-1.5907	0.9484	-0.0621	1.6670	-0.6844
1 par mois	0.3271	-0.1452	-0.7581	0.1949	-0.0720
1 par semaine	-0.2645	0.7421	-0.0498	-0.8008	0.4807
1ere fois	0.0447	-1.5486	0.3944	0.1858	0.6363
< 1 par mois	0.1341	-0.2131	0.7872	0.0211	-0.6013
> 1 par semaine	-1.5009	1.0865	0.5446	1.8132	0.1506
Automobile	0.7695	0.6039	-0.1511	0.6080	-0.2946
Autre	-0.2313	0.4674	0.5169	-0.6090	-1.1817
Bricolage	1.6517	0.2073	-0.5205	-0.1879	0.4933
Electromenager	-0.6415	-0.4791	-0.0306	-0.7177	0.5409
HIFI	0.2628	-0.6295	-0.7936	0.0237	-0.6002
Habillement	-0.4837	0.3713	-0.4197	0.9038	0.2133
Sport	0.5439	0.0242	1.7960	0.5286	0.6428

Contribution des modalités aux axes

Partial Contributions to Inertia for the Column Points					
	Dim1	**Dim2**	**Dim3**	**Dim4**	**Dim5**
Femme	0.1221	0.0281	0.0141	0.0105	0.0064
Homme	0.2179	0.0502	0.0251	0.0188	0.0115
18-24	0.0001	0.1674	0.0003	0.0769	0.1757
25-34	0.0012	0.0996	0.0685	0.0021	0.1202
35-49	0.0276	0.0099	0.0894	0.0004	0.0166
50-64	0.0065	0.0933	0.0163	0.0673	0.1026
65+	0.1093	0.0394	0.0002	0.1364	0.0234
1 par mois	0.0290	0.0058	0.1720	0.0117	0.0016
1 par semaine	0.0134	0.1070	0.0005	0.1397	0.0513
1ere fois	0.0001	0.1705	0.0120	0.0028	0.0329
< 1 par mois	0.0037	0.0095	0.1411	0.0001	0.0866
> 1 par semaine	0.1034	0.0549	0.0150	0.1715	0.0012
Automobile	0.0384	0.0239	0.0016	0.0272	0.0065
Autre	0.0061	0.0254	0.0338	0.0484	0.1857
Bricolage	0.1730	0.0028	0.0190	0.0025	0.0179
Electromenager	0.0811	0.0458	0.0002	0.1154	0.0668
HIFI	0.0077	0.0450	0.0778	0.0001	0.0468
Habillement	0.0360	0.0215	0.0299	0.1429	0.0081
Sport	0.0236	0.0000	0.2833	0.0253	0.0381

Qualité des projections sur les axes

Squared Cosines for the Column Points					
	Dim1	**Dim2**	**Dim3**	**Dim4**	**Dim5**
Femme	0.4328	0.0983	0.0452	0.0328	0.0196
Homme	0.4328	0.0983	0.0452	0.0328	0.0196
18-24	0.0001	0.2284	0.0003	0.0935	0.2095
25-34	0.0019	0.1601	0.1011	0.0030	0.1689
35-49	0.0580	0.0205	0.1706	0.0008	0.0301
50-64	0.0111	0.1569	0.0252	0.1008	0.1508
65+	0.1472	0.0523	0.0002	0.1617	0.0272
1 par mois	0.0564	0.0111	0.3032	0.0200	0.0027
1 par semaine	0.0226	0.1777	0.0008	0.2069	0.0746
1ere fois	0.0002	0.2353	0.0153	0.0034	0.0397
< 1 par mois	0.0064	0.0162	0.2210	0.0002	0.1290
> 1 par semaine	0.1398	0.0732	0.0184	0.2040	0.0014
Automobile	0.0532	0.0328	0.0021	0.0332	0.0078
Autre	0.0092	0.0374	0.0457	0.0634	0.2388
Bricolage	0.2397	0.0038	0.0238	0.0031	0.0214
Electromenager	0.1378	0.0769	0.0003	0.1725	0.0980
HIFI	0.0115	0.0659	0.1047	0.0001	0.0599
Habillement	0.0570	0.0336	0.0429	0.1990	0.0111
Sport	0.0334	0.0001	0.3639	0.0315	0.0466


The SAS System
2
1
0
−1
−2
−2
−1
0
1
2
Dimension 2 (33.54%)
Dimension 1 (51.48%)
> 1 par semaine
65+
1 par semaine
30—34
Autre
Habillement
Automobile
Homme
Bricolage
35—45
Fortement Positive
Un peu Positive
Un peu Négative
Moyenne
Sport
Fortement Négative
1 par mois
Femme
1 par mois
Electroménager
HiFi
25—34
1ere fois
18—24

Autres fonctionnalités de la proc CORRESP

- Gestion des valeurs manquantes (MISSING)
- Standardisation des variables (PROFILE)
- Création de croisements de modalités de variables (CROSS)
  - exemple : age * situation de famille
- Pondérations d'observations (WEIGHT)
  - impossible à utiliser avec VAR et MCA (le fichier en entrée est un tableau de Burt)
  - une observation avec un poids < 0 est une observation supplémentaire

La préparation des données :

Analyse en composantes principales

But de l'ACP (analyse en composantes principales)

- À partir de *n* variables initiales continues, construire *m* ($\leq$ *n*) autres variables, appelées *composantes principales*, combinaisons linéaires des variables initiales, telles que :
  - les CP sont ordonnées selon l'information (variance) qu'elles restituent, la 1ère étant celle qui restitue le plus d'information
  - on sait quelle part d'information restitue chaque CP, et des critères permettent de décider combien de CP il est pertinent de conserver
  - les CP sont des vecteurs indépendants, c'est-à-dire des variables non corrélées entre elles
  - on a une inégalité stricte $m < n$ s'il existe des relations linéaires entre les variables initiales
- Origine : Karl Pearson (1901) – Harold Hotelling (1933)

Intérêt de l'ACP

- Représentation assez fidèle des individus d'une population en 2 ou 3 dimensions
- Localisation des grandes masses d'individus
- Détection des individus exceptionnels et d'éventuels groupes isolés d'individus
- Détection des liaisons entre les variables
- Outil de réduction des dimensions d'un problème
  - diminuer le nombre de variables étudiées sans perdre beaucoup d'information
  - utile avant un réseau de neurones ou une classification
- L'ACP sur les indicatrices de variables nominales conduit aux mêmes résultats que l'ACM sur ces var. nominales
  - ACM ⇔ ACP quand les variables sont binaires

Métrique dans l'espace des individus

- Métrique euclidienne
  - $d(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2$
- Métrique « inverse des variances »
  - $d(x,y) = ((x_1 - y_1) / \sigma_1)^2 + ((x_2 - y_2) / \sigma_2)^2 + \ldots + ((x_n - y_n) / \sigma_n)^2$
  - avec σ_i = écart-type de la $i^{ème}$ variable
- Avec cette nouvelle métrique, qui revient à réduire les variables, la distance entre deux individus ne dépend plus de l'unité de mesure, et les variables les plus dispersées ne sont pas avantagées

Obtention des composantes principales 1/2

- Les composantes principales : obtenues en exprimant les variables initiales, non selon les axes d'origine de l'espace des individus, mais selon de nouveaux axes, les *axes principaux*, qui sont les vecteurs propres de la matrice
  - des covariances, lorsque c'est la métrique euclidienne qui a été choisie dans l'espace des individus
  - des corrélations lorsque les unités de mesure ne sont pas les mêmes pour toutes les variables et que l'on a choisi la métrique « inverse des variances »

$$cov(X,Y) = \sigma_X . \sigma_Y . r_{XY}$$

$$M_{corr} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ \dots & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}$$

Obtention des composantes principales 2/2

- Si on a des données hétérogènes, avec des ordres de grandeur différents, diagonaliser la matrice des corrélations
  - sinon, diagonaliser la matrice des covariances
- Les composantes principales sont classées par variance décroissante
- Variance d'une composante principale = valeur propre correspondante de la matrice diagonalisée
- Trace de cette matrice = Σ_valeurs propres = Σ_variances des composantes principales = Σ_variances des variables d'origine = inertie du nuage d'individus
  - = nombre de variables pour la matrice des corrélations

Exemple de syntaxe SAS

- `proc princomp data=sasuser.etude_de_cas out=coord outstat=stat;`
- `var age ancienete revenus nbproduits nbachats nbpoints evolconsom utilcredit;`
- `weight poids;`
- `run;`
- `proc transpose data=stat out=sortie;run;`
- `proc plot data=sortie; plot prin2*prin1=_name_ $ _name_;run;`
- Si la variable de l'instruction WEIGHT est ≤ 0 pour un individu, il n'est pas pris en compte dans les analyses, mais ses composantes principales sont calculées, ce qui permet de le représenter sur les graphiques comme point supplémentaire (illustratif)

pour utiliser la matrice des covariances : ajouter l'option « COV »

Contenu du fichier OUTSTAT

Obs	_TYPE_	_NAME_	age	anciennete	revenus	nbproduits	nbachats	nbpoints	evolconsom	utilcredit
1	MEAN		38.34	12.24	1882.08	9.23	38.99	1.51	1.14	32.88
2	STD		9.51	7.04	1529.57	4.43	22.66	0.88	0.32	112.92
3	N		5942.00	5942.00	5942.00	5942.00	5942.00	5942.00	5942.00	5942.00
4	CORR	age	1.00	0.38	0.09	0.06	-0.09	-0.08	-0.11	0.01
5	CORR	anciennete	0.38	1.00	0.16	0.22	0.15	0.08	-0.16	0.02
6	CORR	revenus	0.09	0.16	1.00	0.45	0.58	0.36	-0.04	0.16
7	CORR	nbproduits	0.06	0.22	0.45	1.00	0.52	0.62	0.06	0.21
8	CORR	nbachats	-0.09	0.15	0.58	0.52	1.00	0.54	0.03	0.15
9	CORR	nbpoints	-0.08	0.08	0.36	0.62	0.54	1.00	0.12	0.28
10	CORR	evolconsom	-0.11	-0.16	-0.04	0.06	0.03	0.12	1.00	0.00
11	CORR	utilcredit	0.01	0.02	0.16	0.21	0.15	0.28	0.00	1.00
12	EIGENVAL		2.70	1.48	0.94	0.93	0.67	0.57	0.39	0.32
13	SCORE	Prin1	0.04	0.18	0.44	0.50	0.49	0.48	0.03	0.23
14	SCORE	Prin2	0.64	0.61	0.08	0.00	-0.09	-0.19	-0.41	-0.08
15	SCORE	Prin3	0.26	0.19	-0.04	0.12	0.00	0.03	0.77	-0.54
16	SCORE	Prin4	0.31	0.06	-0.27	0.02	-0.30	0.11	0.38	0.75
17	SCORE	Prin5	0.33	-0.39	0.65	-0.31	0.13	-0.38	0.18	0.15
18	SCORE	Prin6	-0.49	0.62	0.12	-0.34	0.22	-0.31	0.23	0.22
19	SCORE	Prin7	0.26	-0.05	-0.32	-0.62	0.54	0.39	-0.03	-0.02
20	SCORE	Prin8	0.11	-0.14	-0.43	0.39	0.55	-0.57	0.01	0.10

Valeurs propres

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.69628418	1.21286011	0.3370	0.3370
2	1.48342407	0.54475789	0.1854	0.5225
3	0.93866619	0.00785658	0.1173	0.6398
4	0.93080961	0.26102963	0.1164	0.7561
5	0.66977998	0.10079635	0.0837	0.8399
6	0.56898363	0.17882698	0.0711	0.9110
7	0.39015665	0.06826097	0.0488	0.9598
8	0.32189568		0.0402	1.0000

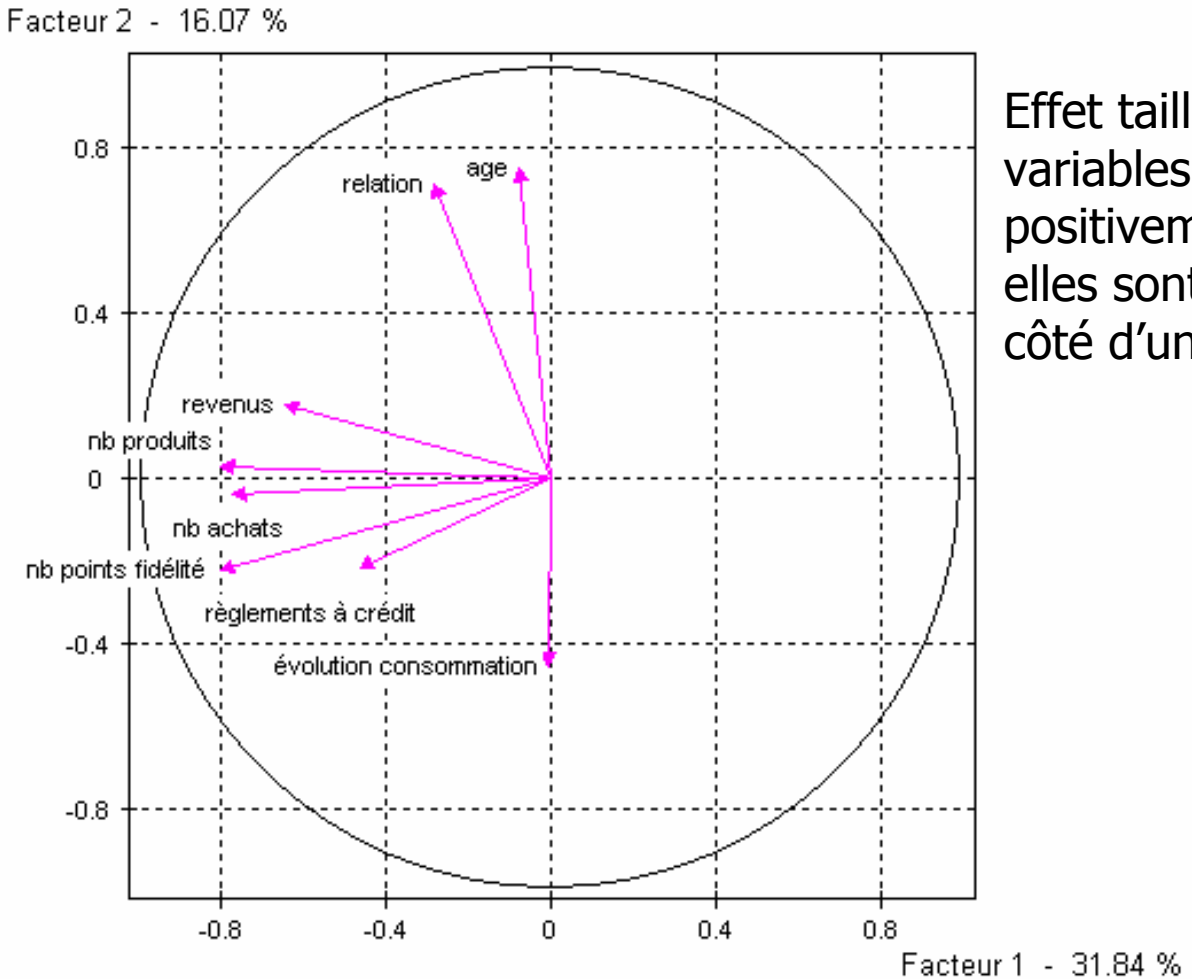
Coefficients des composantes principales

Eigenvectors									
		Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**
age	âge	0.035722	0.637571	0.258608	0.314857	0.332034	-.485766	0.260887	0.109258
anciennete	ancienneté client	0.182484	0.605982	0.187981	0.057860	-.389761	0.622209	-.049299	-.138971
revenus	revenus du client	0.443030	0.080052	-.040509	-.273237	0.651810	0.117833	-.318435	-.425293
nbproduits	nb produits	0.495244	0.001202	0.117938	0.022297	-.306377	-.337970	-.616976	0.389326
nbachats	nb achats	0.491426	-.093575	0.003426	-.304492	0.130701	0.220890	0.542087	0.545237
nbpoints	nb points fidélité	0.478177	-.191761	0.030312	0.110554	-.381114	-.308847	0.389827	-.573440
evolconsom	évolution consommation	0.029868	-.408820	0.770733	0.384360	0.184274	0.234905	-.032310	0.005507
utilcredit	règlements à crédit	0.229862	-.084950	-.535980	0.754779	0.151093	0.222308	-.024057	0.100762

Interprétation des résultats

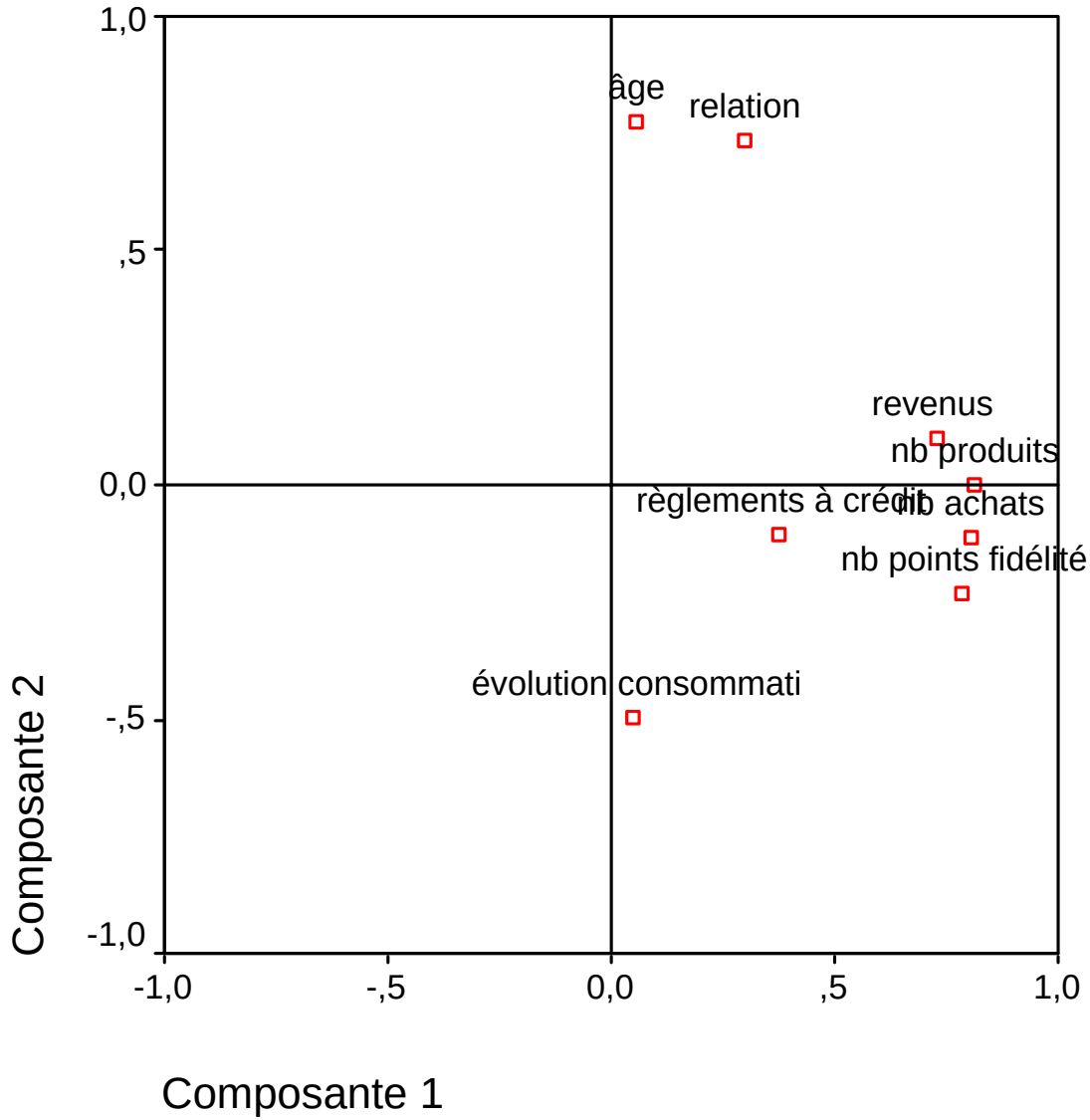
- *Critère de Kaiser* : avec la matrice des corrélations (= sur des données centrées réduites), conserver les axes correspondant aux valeurs propres > 1
- Calculer les valeurs cumulées successives $\lambda_1/\Sigma_i\lambda_i$, $(\lambda_1+\lambda_2)/\Sigma_i\lambda_i$... pour voir quelle proportion de la somme des variances $\Sigma_i\lambda_i$ (c.a.d quelle part de l'inertie totale) est restituée par les p premiers axes principaux
  - plus les variables sont nombreuses et moins elles sont corrélées, plus est faible la part d'inertie restituée par les p premiers axes
- Avec la matrice des corrélations, la coordonnée d'une variable V sur un axe P = le coefficient de corrélation de V et P. Une variable n'est bien représentée sur un plan principal (= 2 axes principaux) que si ses coordonnées sont élevées $\Rightarrow$ si elle est proche du cercle des corrélations

ACP avec SPAD



Effet taille : si toutes les variables sont corrélées positivement entre elles ⇒ elles sont toutes du même côté d'un axe factoriel

ACP avec SPSS



Pièges à éviter

- La proximité de 2 variables est mesurée par le cosinus de l'angle qu'elles forment avec l'origine dans $\mathbf{R}^n$
  - dans $\mathbf{R}^2$, cette proximité n'a de sens que si les variables sont bien représentées sur le plan factoriel, i.e. sont proches du cercle des corrélations
- Le 1er plan principal n'est pas toujours le seul intéressant
- Ne pas superposer l'espace des individus et celui des variables
- Eviter qu'un seul individu (ou un petit groupe d'individus) ait une trop forte contribution aux 1ers axes (> 25 %)
  - si c'est le cas : lui affecter un poids = 0 (et un poids = 1 aux autres individus) pour le transformer en point illustratif grâce à l'instruction WEIGHT

Variables supplémentaires

- Variables ne servant pas à la construction des axes principaux, mais représentées sur les plans principaux
- Variables qualitatives ou quantitatives
- Objectif :
  - variables que l'on veut lier aux variables actives mais pas lier entre elles
  - des variables que l'on veut expliquer par les variables actives
  - variables que l'on veut utiliser pour conforter l'interprétation des axes sans faire appel à des variables ayant servi à les déterminer

Variantes de l'ACP

- Rappel sur le but de l'ACP standard
  - maximiser la variance restituée sur le 1er facteur
  - ⇔ maximiser sa valeur propre = somme des carrés des coefficients de corrélation des variables avec ce facteur
- Inconvénient
  - les variables vont toutes + ou - dans le sens du 1er facteur
- Rotation des facteurs :
  - on pivote les facteurs, en respectant ou non l'orthogonalité, en remplaçant le critère ci-dessus par un autre (selon les méthodes)
  - le but est de faciliter l'interprétation
  - la variance expliquée totale ne change pas après rotation (le sous-espace de projection est le même)
    - mais sa répartition change

Les différentes rotations possibles

- Rotation orthogonale : facteurs non corrélés
  - meilleure interprétation des facteurs
  - ACP varimax
  - ACP quartimax
  - ACP equamax : compromis entre varimax et quartimax
- Rotation oblique : facteurs corrélés
  - les facteurs ne sont plus orthogonaux
  - valeurs propres + fortes $\Rightarrow$ + forte corrélation des facteurs avec les variables
  - mais interprétation plus difficile
  - ACP oblimin
  - ACP promax (+ rapide $\Rightarrow$ utilisé sur de gros volumes)
- Dans SAS/STAT : proc FACTOR
  - + générale, mais + complexe et - rapide que PRINCOMP

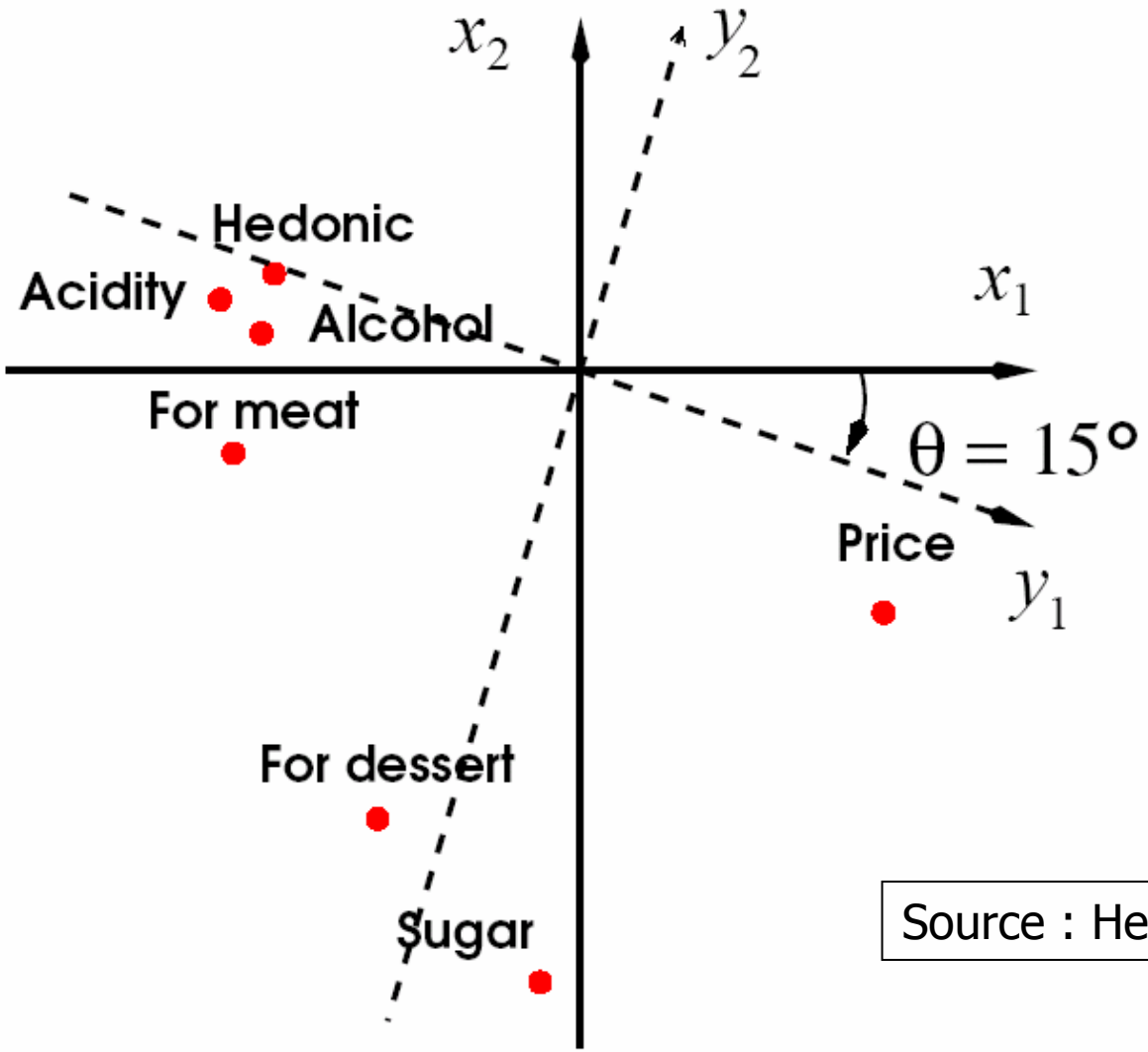
ACP varimax

- Pour chaque facteur
  - on calcule ses coefficients de corrélation avec l'ensemble des variables
  - puis on calcule la variance de ces coefficients de corrélation
  - on pivote le facteur de façon à maximiser cette variance
  - différence avec l'ACP standard, où on maximise la somme des carrés des coefficients de corrélation et non leur variance
- Chaque facteur est fortement corrélé à quelques variables et faiblement corrélé aux autres

ACP varimax

- Chaque variable est identifiée à un (ou un petit nombre) facteur
  - les variables ne vont plus seulement dans le sens du 1er axe
  - elles sont bien séparées par axe
  - les axes sont facilement interprétables
- Variante de l'ACP la plus utilisée

Exemple de rotation VARIMAX



Source : Hervé Abdi

ACP quartimax

- Pour chaque variable
  - on calcule ses coefficients de corrélation avec l'ensemble des facteurs
  - puis on calcule la variance de ces coefficients
  - on pivote les facteurs de façon à maximiser cette variance
- Chaque variable est fortement corrélée à quelques facteurs et faiblement corrélée aux autres
- Minimisation du nombre de facteurs nécessaires pour expliquer chaque variable
- Variante utilisée dans la classification VARCLUS de variables (implémentée dans SAS/STAT)

Autre variante : l'ACP sur variables catégorielles

- Transforme de façon optimale les variables qualitatives en variables numériques
  - de façon à optimiser les propriétés de la matrice des corrélations ou des covariances
    - notamment maximiser l'importance des 1ères valeurs propres
- Dans SAS/STAT : proc PRINQUAL
- Par rapport à l'analyse des correspondances (proc CORRESP) :
  - (=) on traite des variables qualitatives
  - ($\neq$) on travaille sur la matrice des corrélations et non sur les effectifs de tableaux de contingence
  - ($\neq$) il s'agit d'une ACP sur des var. qualitatives transformées chacune en 1 var. numérique, et non, comme l'ACM, d'une ACP sur les *k* indicatrices de chaque variable qualitative