Rotation orthogonale en ACP de données mixtes. Le package PCAmixdata et une application en sociologie culturelle.

Marie Chavent [1,2], Vanessa Kuentz-Simonet [3]
Zoltan Lakatos [4], Jérôme Saracco [1,2]

[1]IMB, Université de Bordeaux, France
[2]Inria Bordeaux Sud-Ouest, Equipe CQFD, Talence, France
[3]Irstea, UR ADBX, Cestas, France
[4]Université Polytechnique et Economique de Budapest, Hongrie

Plan

Plan

Une ACP de données mixtes

Analyse en Composantes Principales d'un mélange de données quantitatives et qualitatives

- PCAMIX (Kiers, 1991) et AFDM (Pagès, 2004)
- Inclut l'ACP et l'ACM comme cas particuliers
- Fonction `AFDM` dans le package R **FactoMineR**
- Rotation dans la méthode PCAMIX
  $\hookrightarrow$ Réécriture de PCAMIX sous forme d'une Décomposition en Valeurs Singulières

Quelques notations

- Soit $\mathbf{X}_1$ une matrice $n \times p_1$ de données **quantitatives** où n observations sont décrites par p_1 variables quantitatives
- Soit $\mathbf{X}_2$ une matrice $n \times p_2$ de données **qualitatives** où n observations sont décrites par p_2 variables qualitatives
- Soit $p = p_1 + p_2$ le nombre total de variables et m le nombre total de modalités
- Soit k le nombre de composantes issues de PCAMIX

Une première étape de recodage

La procédure pour PCAMIX se déroule de la façon suivante :

1. Recodage de $\mathbf{X}_1$ et $\mathbf{X}_2$:
   - $\mathbf{Z}_1$ est la version standardisée de la matrice quantitative $\mathbf{X}_1$
   - $\mathbf{Z}_2 = \mathbf{JGD}^{-1/2}$ est la version standardisée du tableau disjonctif complet $\mathbf{G}$ associé à la matrice qualitative $\mathbf{X}_2$, où $\mathbf{D}$ est la matrice diagonale des fréquences des modalités et $\mathbf{J} = \mathbf{I} - \mathbf{1}'\mathbf{1}/n$ est l'opérateur de centrage

$\hookrightarrow \mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{Z}_1|\mathbf{Z}_2)$ est la matrice $n \times (p_1 + m)$ d'intérêt

Décomposition en Valeurs Singulières

2. Décomposition en Valeurs Singulières de $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}'$$

$\hookrightarrow$ $\mathbf{F} = \sqrt{n}\mathbf{U}_k$ est la matrice $n \times k$ des scores des composantes principales, où $\mathbf{U}_k$ est la matrice composée des k premières colonnes de $\mathbf{U}$
$\hookrightarrow$ $\mathbf{A} = \mathbf{V}_k\Lambda_k$ est la matrice $(p_1 + m) \times k$ des "loadings" des composantes principales, où $\mathbf{V}_k$ est la matrice composée des k premières colonnes de $\mathbf{V}$ et Λ_k la matrice diagonale des k premières valeurs singulières

Réécriture de la matrice A des loadings

3. Ecrire $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$ avec :
   - $\mathbf{A}_1$ la matrice $p_1 \times k$ des "loadings" (corrélations) des variables quantitatives
   - $\mathbf{DA}_2$ la matrice $m \times k$ des coordonnées des modalités des variables qualitatives sur les composantes principales

$\hookrightarrow$ Cercle des corrélations pour les variables quantitatives

$\hookrightarrow$ Graphique des modalités

Calcul de la matrice C des "squared loadings"

4. Calculer la matrice **C** de dimension $(p_1 + p_2) \times k$ des "squared loadings" :

$$\begin{cases} c_{jl} = a_{jl}^2 & \text{si la variable } j \text{ est quantitative} \\ c_{jl} = \sum_{s \in I_j} a_{sl}^2 & \text{si la variable } j \text{ est qualitative} \end{cases}$$

où I_j est l'ensemble des indices des lignes de **A** associés aux modalités de la variable j

$\hookrightarrow$ c_{jl} est une corrélation au carré si j est quantitative

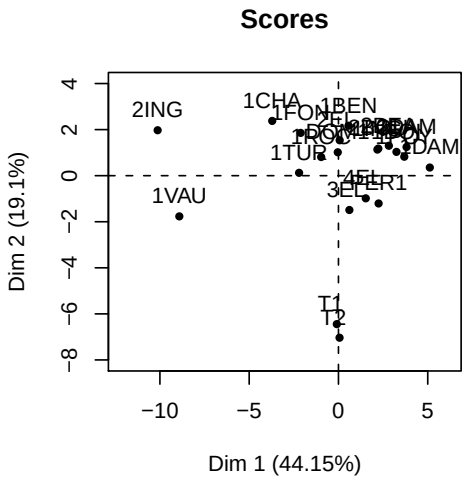
$\hookrightarrow$ c_{jl} est un rapport de corrélation si j est qualitative

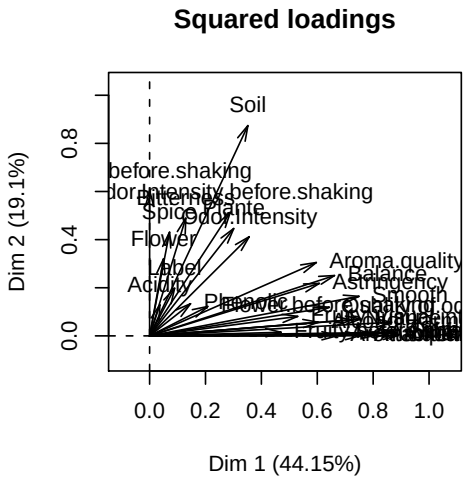
$\hookrightarrow$ Variables quantitatives et qualitatives sur le même graphique

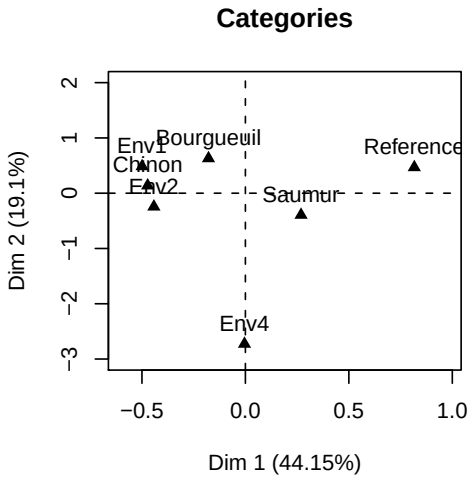
Le package R PCAmixdata

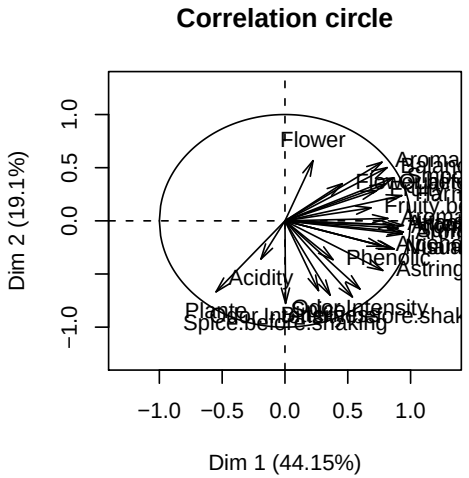
```
> require(PCAmixdata)
> data(wine)
> head(wine[,c(1:4)])
      Label        Soil        Odor.Intensity  Aroma.quality
2EL   Saumur       Env1        3.07            3.00
1CHA  Saumur       Env1        2.96            2.82
1FON  Bourgueuil   Env1        2.85            2.92
1VAU  Chinon       Env2        2.80            2.59
1DAM  Saumur       Reference   3.60            3.42
2BOU  Bourgueuil   Reference   2.85            3.11

> X.quanti <- wine[,c(3:29)]
> X.quali <- wine[,c(1,2)]
> pca <- PCAmix(X.quanti,X.quali,ndim=10)
```

Le package R PCAmixdata









Plan

Rotation en ACP (1/2)

En conservant k composantes principales :

$$\begin{aligned} \mathbf{Z} &\approx \mathbf{U}_k \Lambda_k \mathbf{V}_k' \\ &= \mathbf{F}\mathbf{A}' \\ &= \mathbf{F}\mathbf{T}\mathbf{T}'\mathbf{A}' \\ &= \widetilde{\mathbf{F}}\widetilde{\mathbf{A}}' \end{aligned}$$

où

- $\mathbf{T}$ est une matrice de rotation orthonormale : $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k$
- $\widetilde{\mathbf{F}} = \mathbf{F}\mathbf{T}$ et $\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{T}$ sont les scores et loadings après rotation

Rotation en ACP (2/2)

↪ Faciliter l'interprétation : trouver $\mathbf{T}$ tel que les loadings au carré aient des valeurs élevées (proche de 1) ou faibles (proche de zéro)

↪ La fonction varimax (Kaiser, 1958) :

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{j=1}^{p}(\tilde{a}_{jl}^2)^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{a}_{jl}^2\right)^2$$

↪ Le problème d'optimisation :

$$\begin{aligned} \max_{\mathbf{T}} \quad & f(\mathbf{T}), \\ \text{s.c.} \quad & \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k \end{aligned}$$

Critère de rotation pour PCAMIX

Dans PCAMIX, la fonction varimax s'écrit :

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{j=1}^{p}(\tilde{c}_{jl})^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{c}_{jl}\right)^2$$

où $\tilde{c}_{jl} = \sum_{s\in I_j}\tilde{a}_{sl}^2$ sont ici les loadings au carré après rotation

$\hookrightarrow$ Les loadings au carré après rotation $\tilde{c}_{jl}$ sont les corrélations au carré ou les rapports de corrélation des variables aux scores après rotation dans $\widetilde{\mathbf{F}}$

$\hookrightarrow$ Kiers (1991) donne une formulation matricielle de cette fonction varimax

$\hookrightarrow$ Il propose d'utiliser un algorithme de diagonalisation simultanée de matrices symétriques (De Leeuw et Pruzansky, 1978)

Rotation orthogonale pour PCAMIX

Notre proposition :

- Ecriture de la solution directe pour l'angle optimal de rotation dans la méthode PCAMIX ($k = 2$)
- Proposition d'une procédure itérative pour la rotation varimax lorsque $k > 2$
- Développement du package R **PCAmixdata** avec la fonction "PCArot"

L'angle optimal de rotation ($k = 2$)

Pour $k = 2$

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$\hookrightarrow$ le problème d'optimisation varimax devient non contraint :

$$\max_{\theta \in \mathbb{R}} f(\theta)$$

Rotation planaire dans PCAMIX (1/3)

On démontre que :

$$f(\theta) = f(0) + \frac{\rho}{4p}\left(\cos(4\theta - \psi) - \cos\psi\right)$$

où ρ et ψ sont définis par :

$$\rho = (g^2 + h^2)^{1/2} \quad , \quad \cos\psi = g/\rho \quad , \quad \sin\psi = h/\rho$$

et où g et h sont donnés par ...

Rotation planaire dans PCAMIX (2/3)

... où g et h sont donnés par :

$$g = 2p\sum_{j=1}^{p} u_j v_j - 2\sum_{j=1}^{p} u_j \sum_{j=1}^{p} v_j$$

$$h = p\sum_{j=1}^{p}({u_j}^2 - {v_j}^2) - \left(\sum_{j=1}^{p} u_j\right)^2 + \left(\sum_{j=1}^{p} v_j\right)^2$$

et où u_j et v_j sont définis par :

$$u_j = \sum_{s\in I_j}(a_{s1}^2 - a_{s2}^2) \quad \text{et} \quad v_j = 2\sum_{s\in I_j} a_{s1}a_{s2}$$

Rotation planaire dans PCAMIX (3/3)

$$f(\theta) = f(0) + \frac{\rho}{4p}\left(\cos(4\theta - \psi) - \cos\psi\right)$$

est maximum pour

$$\cos(4\theta - \psi) = 1 \Leftrightarrow 4\theta - \psi = 2k\pi$$

$\hookrightarrow$ les angles optimaux sont :

$$\theta = \frac{\psi}{4} + k\frac{\pi}{2}, \quad k \in \mathbb{Z}$$

Une procédure itérative de rotation pour $k > 2$ (1/2)

1. Initialisation :
   - Calculer $\mathbf{F}$ et $\mathbf{A}$ avec PCAMIX
   - $\widetilde{\mathbf{F}} = \mathbf{F}$ et $\widetilde{\mathbf{A}} = \mathbf{A}$
2. Pour chaque pair de dimensions (l, t) :
   - Calculer $\theta = \psi/4$ avec

$$\psi = \begin{cases} \arccos\left(\dfrac{h}{\sqrt{g^2 + h^2}}\right) & \text{si} \quad g \geq 0 \\ -\arccos\left(\dfrac{h}{\sqrt{g^2 + h^2}}\right) & \text{si} \quad g \leq 0 \end{cases}$$

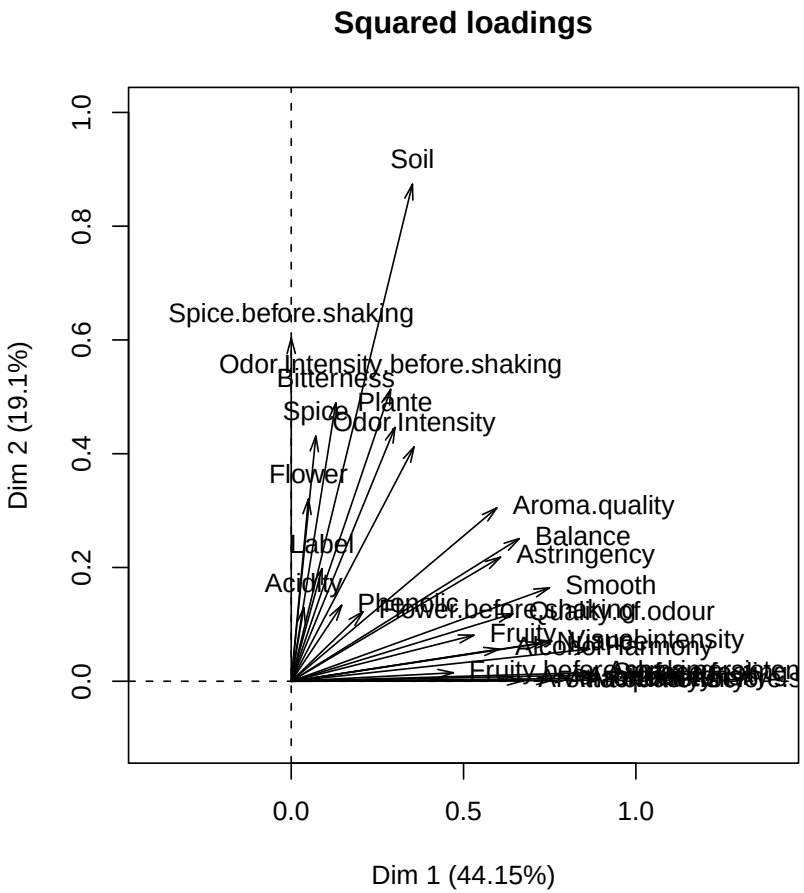
   - ...

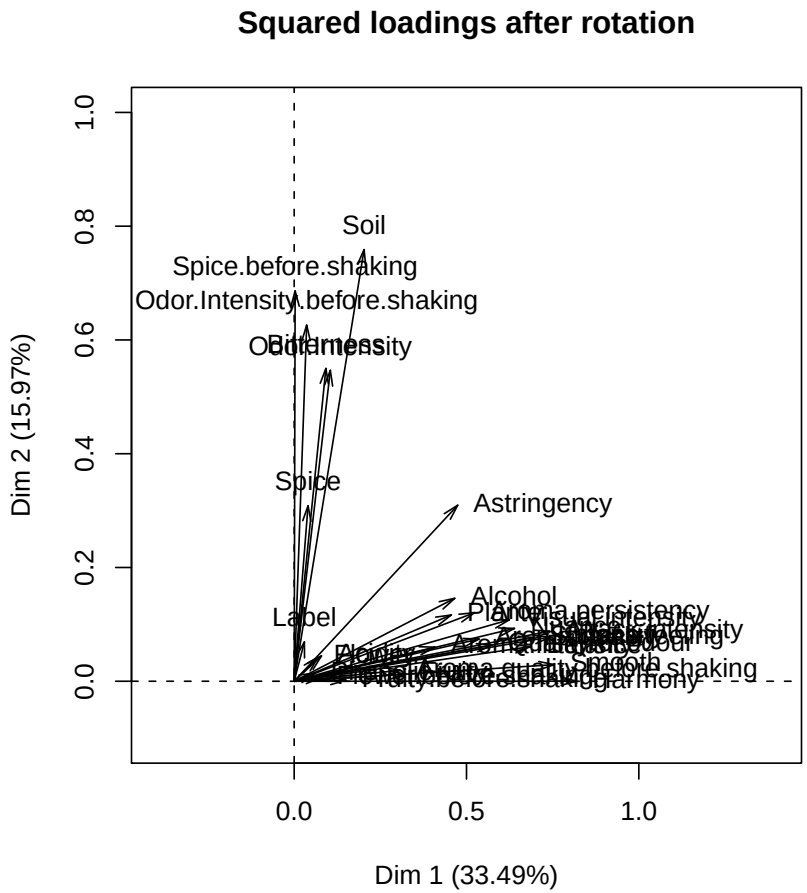
Une procédure itérative de rotation pour $k > 2$ (2/2)

2. Pour chaque pair de dimensions (l, t) :
   - ...
   - $\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
   - Mettre à jour $\widetilde{\mathbf{F}}$ et $\widetilde{\mathbf{A}}$ par rotation de leurs colonnes l et t
3. Répéter l'étape précédente jusqu'à obtenir successivement $k(k-1)/2$ angles θ égaux à zéro

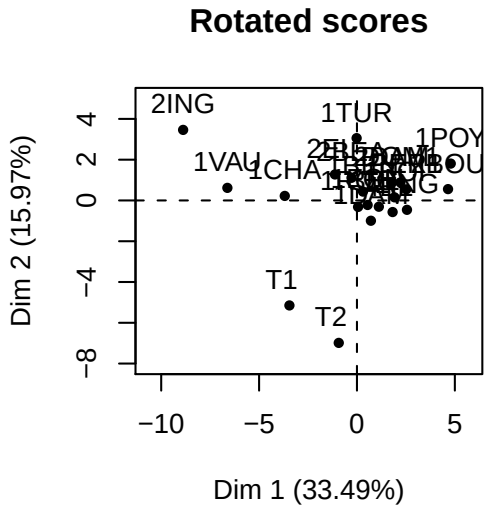
Le package R PCAmixdata

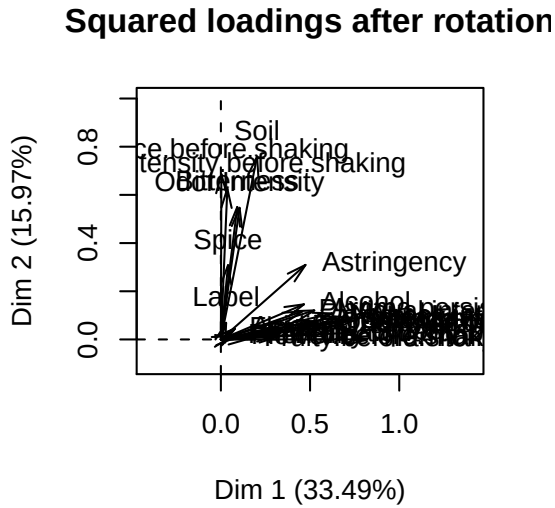
```
> rot<-PCArot(pca,dim=8)
```

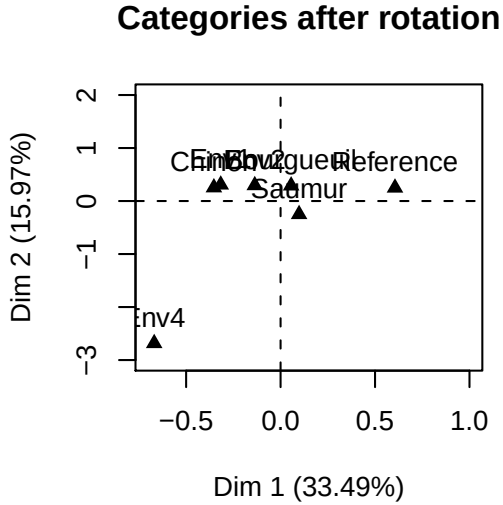


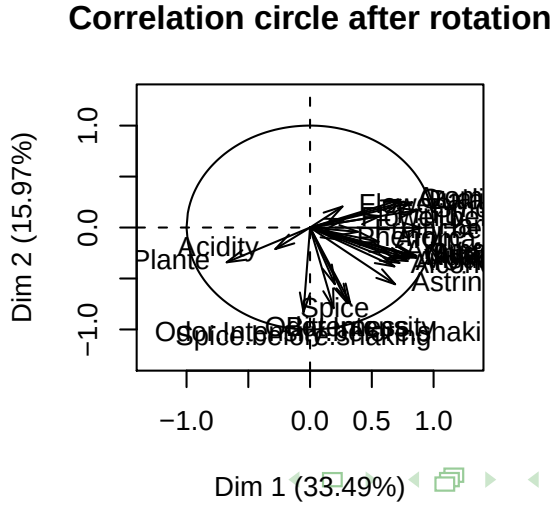


Le package R PCAmixdata









Comparaison avec l'approche de Kiers

- Simulations de jeux de données (variables quantitatives et qualitatives)
- 20 réplications pour chaque couple (n, p)
- Comparaison du temps de calcul médian (en sec) de l'angle optimal entre les 2 approches

		p=10	p=50	p=100	p=200
n=50	Matrix reformulation	0.05	0.12	0.22	0.44
n=50	SVD	0.02	0.06	0.12	0.27
n=100	Matrix reformulation	0.14	0.33	0.56	1.04
n=100	SVD	0.02	0.09	0.17	0.34
n=200	Matrix reformulation	0.55	1.12	1.86	3.38
n=200	SVD	0.02	0.11	0.26	0.53
n=400	Matrix reformulation	2.15	4.32	7.1	12.65
n=400	SVD	0.03	0.16	0.37	0.89
n=800	Matrix reformulation	10.06	19.27	30.54	error
n=800	SVD	0.05	0.25	0.58	1.79

$\hookrightarrow$ Ratio entre les temps de calcul des deux approches : de 2 à 214 fois plus rapide !

Plan

Autour de l'évolution des valeurs culturelles

Travail de thèse de Zoltan Lakatos (2012) :

- Evolution des valeurs culturelles dans les sociétés
- Critique empirique de la thèse sociologique du *post materialism* du politologue américain Ronald Inglehart
- World Values Survey (WVS) : enquête globale sur les valeurs culturelles, initiée et dirigée par Ronald Inglehart (enquêtes individuelles menées au niveau national dans une centaine de pays, par vagues successives depuis 1981)

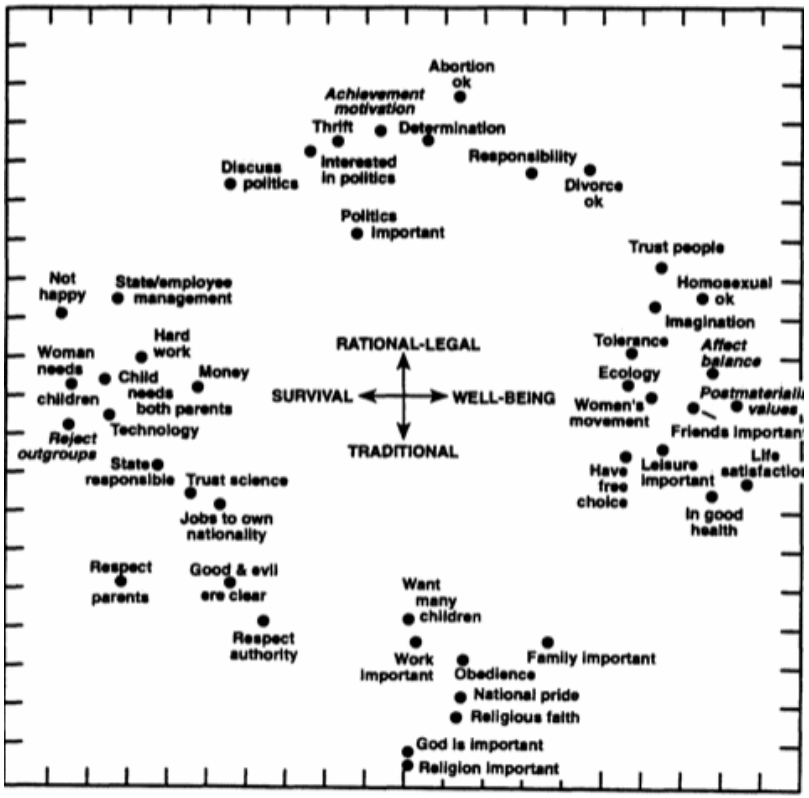
Les résultats de Inglehart

Graphique de l'espace sociologique selon Inglehart :

- Inglehart (1997), Modernization and Postmodernization, p. 82, Figure 3.2.
- Source : vague 1990-93 de l'enquête World Values Survey, 43 "pays".
- Premier plan factoriel d'une ACP réalisée sur les données agrégées au niveau national

Amalgame entre deux notions

Le deuxième axe "valeurs traditionnelles vs. rationalité-laïcité" amalgame l'attitude libertaire et l'activisme citoyen (en haut) et la forte religiosité et l'autoritarisme (en bas) :



Une approche nouvelle

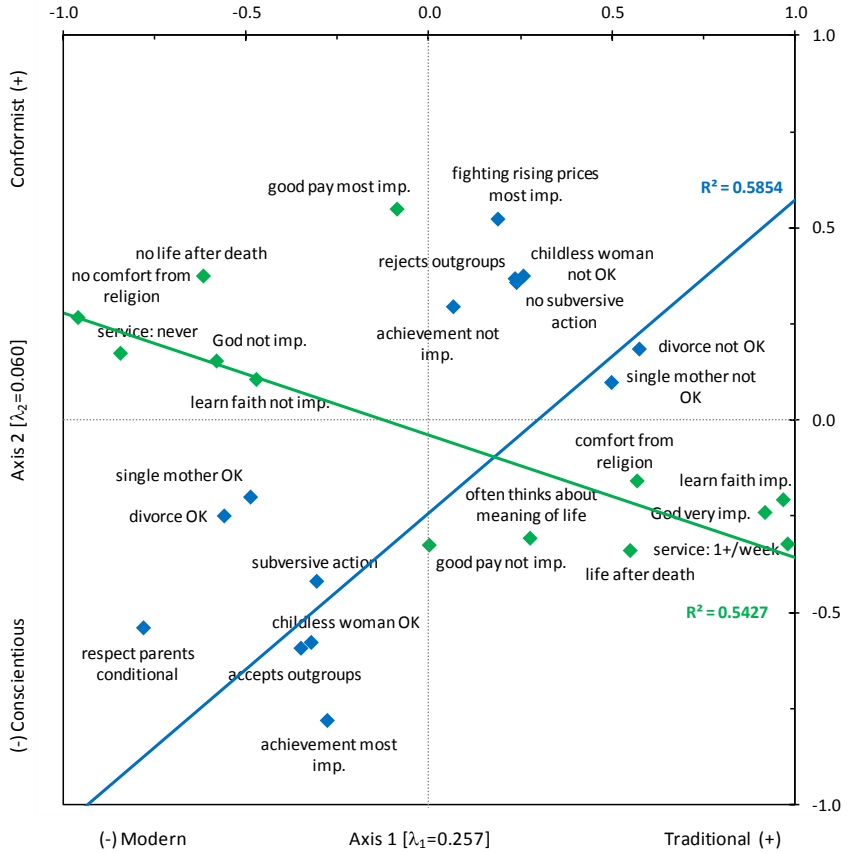
Approche de Zoltan Lakatos (2012) :

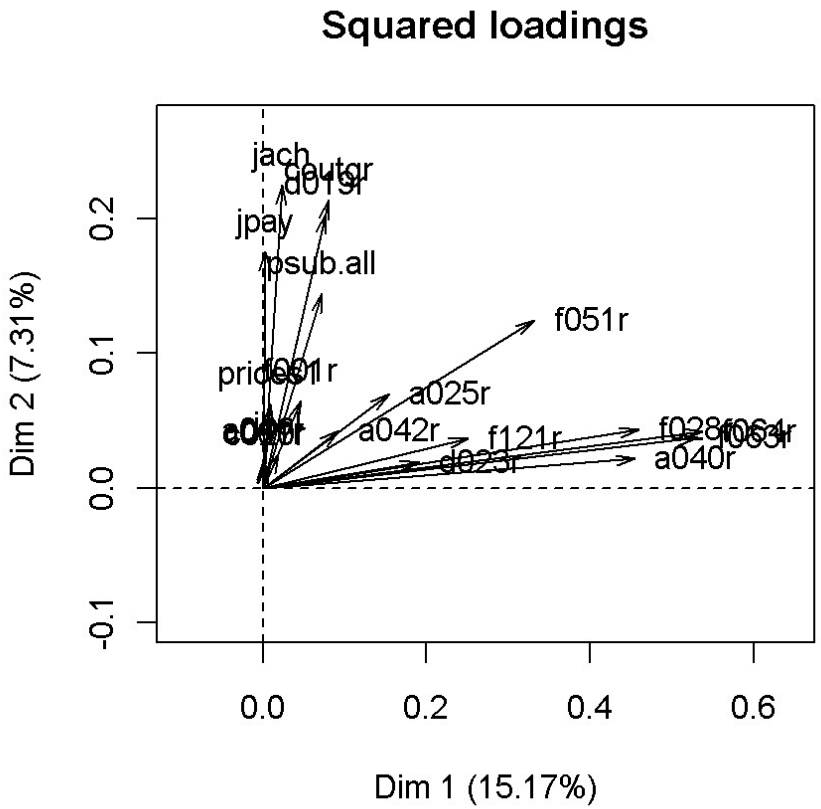
- Enquête WVS de 1981 à 2004 dans 86 pays soit 276870 cas pondérés pour obtenir 1000 cas par pays
- 86000 cas et 20 valeurs culturelles (données manquantes)
- ACM avec rotation (et non pas ACP) sur les données individuelles (et non pas agrégées)
- Package PCAmixdata et CAR de Matlab pour la rotation en AC (van de Velden & Kiers, 2005)

$\Rightarrow$ Identification de deux dimensions distinctes : "religieux vs. laïque " et "autoritaire vs. libertaire"

Résultats de l'ACM

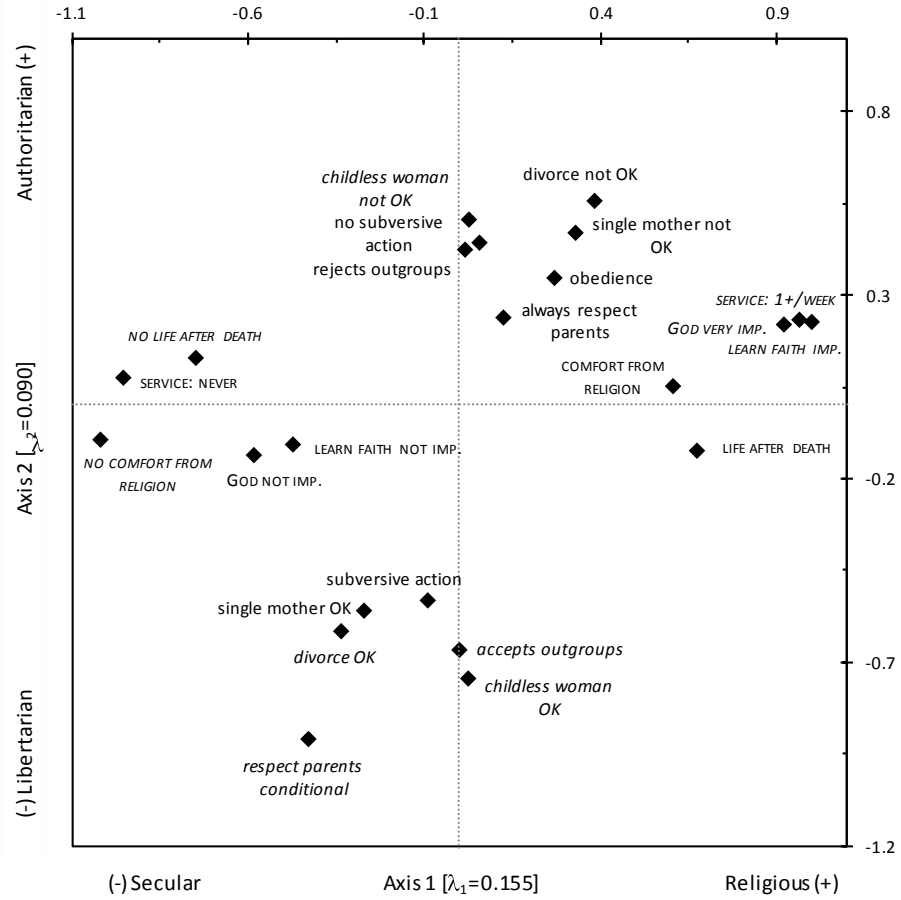
Modalités et variables avant rotation

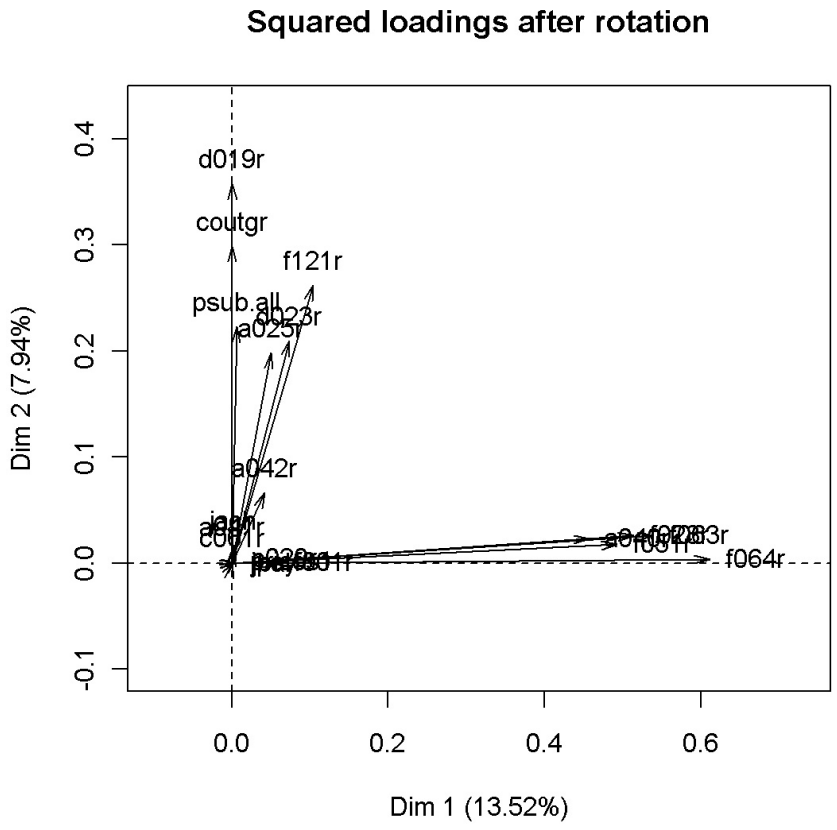




Apport de la rotation (1/2)

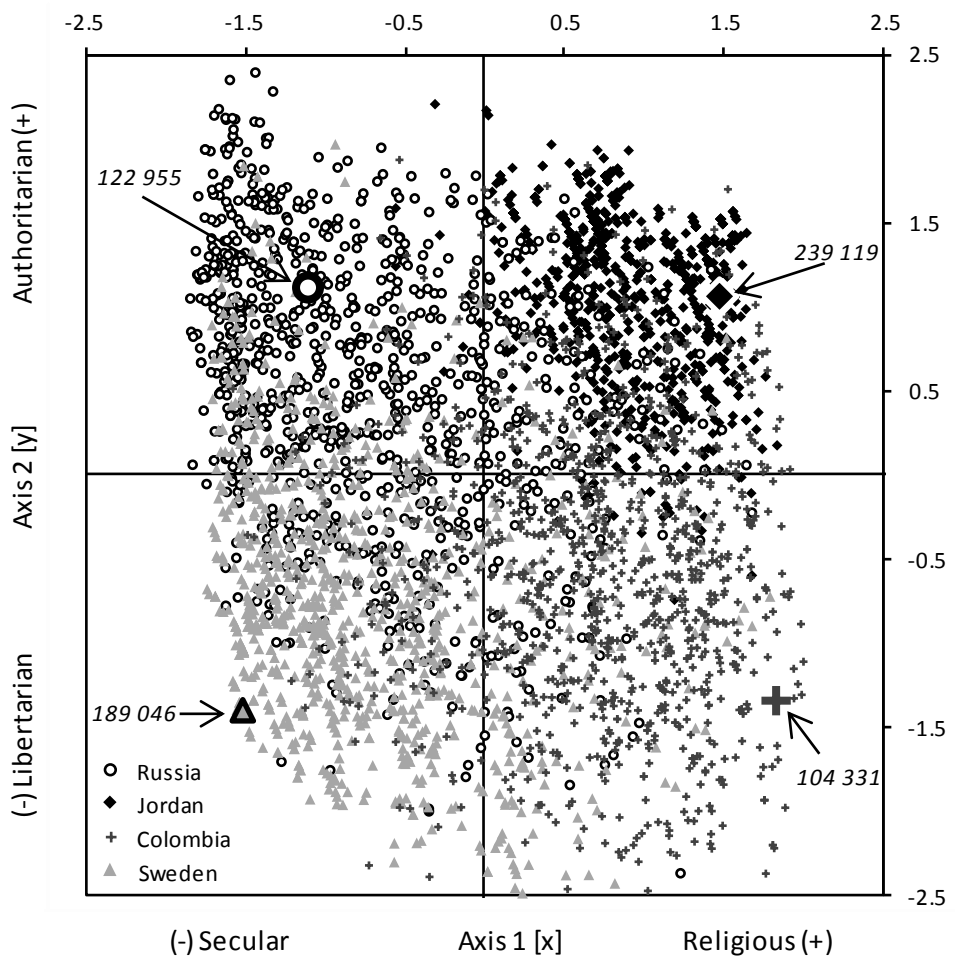
Modalités et variables après rotation ($k = 4$)





Apport de la rotation (2/2)

Scores des individus après rotation ($k = 4$)



Conclusion

- Ecriture de la solution analytique pour la rotation planaire en ACP de données mixtes
- Proposition d'une procédure itérative de rotation pour $k > 2$
- Illustration du bon comportement numérique de l'approche
- Développement du package R PCAmixdata
- Intérêt possible de la rotation sur une vraie étude de cas

Quelques références

- Chavent, M., Kuentz, V., Liquet B., Saracco, J. (2012), *The PCAmixdata R package*, The CRAN R Project.
- Chavent, M., Kuentz, V., Saracco, J. (2012), Orthogonal rotation in PCAMIX, *ADAC*, 6(2), 131-146.
- Kiers, H.A.L., (1991), Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197-212.
- Lakatos, Z. (2012), The Cultural Values-Economic Growth Nexus: A Critical Reassessment, *D*octoral thesis, Faculty of Social Sciences, Eotvos Lorand University of Sciences (ELTE TaTK), Budapest.