L'ANALYSE EN VARIABLES LATENTES

- Une autre manière d'analyser des relations entre variables observées (variables **manifestes**)
- Les modèles à variables **latentes** (ou facteurs) postulent l'existence de variables inobservables directement telles l'intelligence, l'engagement religieux, etc. mais dont on peut mesurer ou observer les effets, comme la fréquentation des lieux de culte, la réussite à certains tests…

- Hypothèse fondamentale: les covariations entre variables observées s'expliquent par la dépendance de chaque variable observée avec les variables latentes.
- principe d'indépendance conditionnelle :

  les variables observées sont indépendantes conditionnellement aux variables latentes.

	Variables latentes	
Variables observées	**qualitatives**	**quantitatives**
qualitatives	Analyse des classes latentes	Analyse des traits latents
quantitatives	Analyse des profils latents	Analyse factorielle

Bibliographie

- Bartholomew, et D.J.& Knott, M. ***Latent Variable Models and Factor Analysis*** , 2nd edition, Arnold, 1999.
- Everitt, B.S., ***An Introduction to Latent Variable Models*** , Chapman & Hall, London, 1984.
- Hagenaars J.A., McCutcheon A.L. (eds.), ***Applied latent class analysis,*** Cambridge University Press, 2002.
- Lazarsfeld, P.F. et Henri, N.W., ***Latent Structure Analysis*** , Houghton Mifflin, Boston, 1968.
- McCutcheon, A.L, ***Latent Class Analysis*** , (Sage University Paper Series on Quantitative Applications in the Social Sciences n°64) , SAGE publications, 1987.

- Logiciels et sites Web:
  - logiciel gratuit " lvmfa2 " (cf.Bartholomew & Knott) : http://www.arnoldpublishers.com/support/lvmfa2.htm
  - logiciel gratuit “LEM” par J.K.Vermunt de l'Université de Tilburg http://cwis.kub.nl/~fsw_1/mto/mto_snw.htm#software
  - logiciel " Latent Gold " distribué par Statistical Innovations: http://www.statisticalinnovations.com/products/latentgold_v4.html
  - " Latent Class Analysis Website ", par John Uebersax : http://ourworld.compuserve.com/homepages/jsuebersax/

A. L'analyse factorielle

- « Factor analysis » ou analyse en facteurs communs et spécifiques
- Charles Spearman (1901) : **modèle unifactoriel**

$$x_i = \lambda_i f + u_i \qquad i = 1, \ldots p$$

- f facteur commun, u_i facteur spécifique, λ_i saturation ou « factor loading »

Exemple

- Corrélation entre résultats scolaires

  Classics, english, french

$$R = \begin{pmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{pmatrix}$$

$$\begin{cases} x_1 = 0.983\, f + u_1 \\ x_2 = 0.844\, f + u_2 \\ x_3 = 0.794\, f + u_3 \end{cases}$$

- Thurstone: **modèle plurifactoriel ,** k facteurs

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + u_i \qquad i = 1,\ldots p$$

- Modèle de régression multiple multilinéaire sur variables inobservables
- Hypothèses :
  - x_i centrés (réduits)
  - f_j centrés réduits indépendants entre eux
  - u_i centrés indépendants entre eux $V(u_i)=\psi_i$
  - u_i indépendants des f_j

Formulation matricielle

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u}$$

- Décomposition de la variance

$$V(x_i) = \sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i = h_i^2 + \psi_i$$

Communauté + spécificité

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$$

Indépendance conditionnelle

- Matrice de variance des manifestes et des facteurs

$$\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \mathbf{I} \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \qquad \boldsymbol{\Sigma}_{11/2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \left(\boldsymbol{\Sigma}_{22} \right)^{-1} \boldsymbol{\Sigma}_{21}$$

$$\boldsymbol{\Sigma}_{\mathbf{x/f}} = \boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \boldsymbol{\Psi}$$

Matrice diagonale: l'AF explique les corrélations

Indétermination des facteurs

- Modèle surparamétré
- Invariance par transformation orthogonale

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u} = \mathbf{\Lambda GG'f} + \mathbf{u}$$

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi} = \mathbf{\Lambda GG'\Lambda}' + \mathbf{\Psi}$$

- Contrainte mathématique:

$$\mathbf{\Lambda'\Psi}^{-1}\mathbf{\Lambda} \quad \text{diagonale}$$

Degré de liberté

- Nombre d'équations p(p+1)/2 $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
- Nombre de contraintes k(k-1)/2 $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$
- Nombre d'inconnues pk+p λ_{ij} et ψ_i

$$s = \frac{p(p+1)}{2} + \frac{k(k-1)}{2} - p(k+1) = \frac{1}{2}\left[(p-k)^2 - (p+k)\right]$$

- S=0 solution unique mais..
- S<0 indétermination
- S>0 plus d'équations que d'inconnues: le cas utile! Représentation parcimonieuse

- S=0 solution artificielle , Ψ_i peut être négatif
- Cas exacts:

  P=3 k=1

  P=6 k=3

  P=10 k=6

  P=15 k=10

  P=21 k=15

  …

Estimation

- **Méthode des facteurs principaux**
  - 1ère estimation des $h_i^2 = 1 - \psi_i = \sum_j \lambda_{ij}^2$

    par $R^2\left(x_i; x_1, .., x_{i-1}, x_{i+1}, ..\right) \sim R^{2'}\left(x_i; f_1, .., f_k\right)$

    D'où estimation de Ψ
  - Diagonalisation de R- Ψ Matrice de corrélation réduite $\mathbf{R - \Psi \simeq \Lambda \Lambda'}$
  - Rééstimation de Ψ etc.

Estimation (2)

- Maximum de vraisemblance
  - $-2\ln(L) \sim \chi^2_s$

Rotation des facteurs

- Indétermination: GG'=I $\Delta=\Lambda G$
- Recherche de « structures simples »
  - Chaque variable corrélée avec peu de facteurs
  - Chaque facteur avec peu de variables
- Rotation « varimax »
  - Maximise la somme des variances des carrés des saturations intra-colonnes, normalisée par la communauté

$$\sum_{i=1}^{p} V_j \left(\frac{\delta_{ij}^2}{h_i^2} \right)$$

- Nombreuses autres méthodes
  - Orthogonales: quartimax, orthomax, parsimax
  - Obliques: biquartimin, oblimin, quartimin …

  fournit des facteurs corrélés après rotation.

  « *Oblique rotations often produce more useful patterns than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable* »

Estimation des scores

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{u}$$

f coefficients d'un modèle linéaire!

- **Scores de Bartlett**

Moindres carrés généralisés

$$\hat{\mathbf{f}} = \left(\Lambda' \Psi^{-1} \Lambda\right)^{-1} \Lambda' \Psi^{-1} \mathbf{x}$$

- **Scores de Thompson** (approche bayesienne)
  - Loi conditionnelle de x/f: $N_p(\Lambda f, \psi)$
  - Loi *a priori* de f: $N_k(0;I)$
  - Espérance de f *a posteriori*

$$\hat{\mathbf{f}} = \left(\mathbf{I} + \Lambda' \Psi^{-1} \Lambda\right)^{-1} \Lambda' \Psi^{-1} \mathbf{x}$$

```
proc factor data=bagnole method=principal
nfactors=2 priors=smc preplot
rotate=varimax reorder plot ;

var CYL PUIS LON LAR POIDS VITESSE;

run;
```

The FACTOR Procedure
Initial Factor Method: Principal Factors

Prior Communality Estimates: SMC

CYL	PUIS	LON	LAR	POIDS	VITESSE
0.73488960	0.91006240	0.86119199	0.76176890	0.89957099	0.84314001

Eigenvalues of the Reduced Correlation Matrix: Total = 5.0106239 Average = 0.83510398

	Valeur propre	Différence	Proportion	Cumulée
1	4.25891481	3.55620941	0.8500	0.8500
2	0.70270540	0.50308529	0.1402	0.9902
3	0.19962012	0.20503633	0.0398	1.0301
4	-.00541621	0.06408051	-0.0011	1.0290
5	-.06949672	0.00620677	-0.0139	1.0151
6	-.07570349		-0.0151	1.0000

2 factors will be retained by the NFACTOR criterion.

Factor Pattern

	Factor1	Factor2
POIDS	0.90145	-0.23116
PUIS	0.88599	0.36856
LON	0.87407	-0.35609
CYL	0.85610	0.07966
LAR	0.78351	-0.33246
VITESSE	0.74198	0.51939

Variance Explained by Each Factor

Factor1	Factor2
4.2589148	0.7027054

Final Communality Estimates: Total = 4.961620

CYL	PUIS	LON	LAR	POIDS	VITESSE
0.73925315	0.92081180	0.89079542	0.72441494	0.86604445	0.82030045

analyse en facteurs communs et spécifiques 17:20 Wednesday, November 30,

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.75773	0.65257
2	-0.65257	0.75773

Rotated Factor Pattern

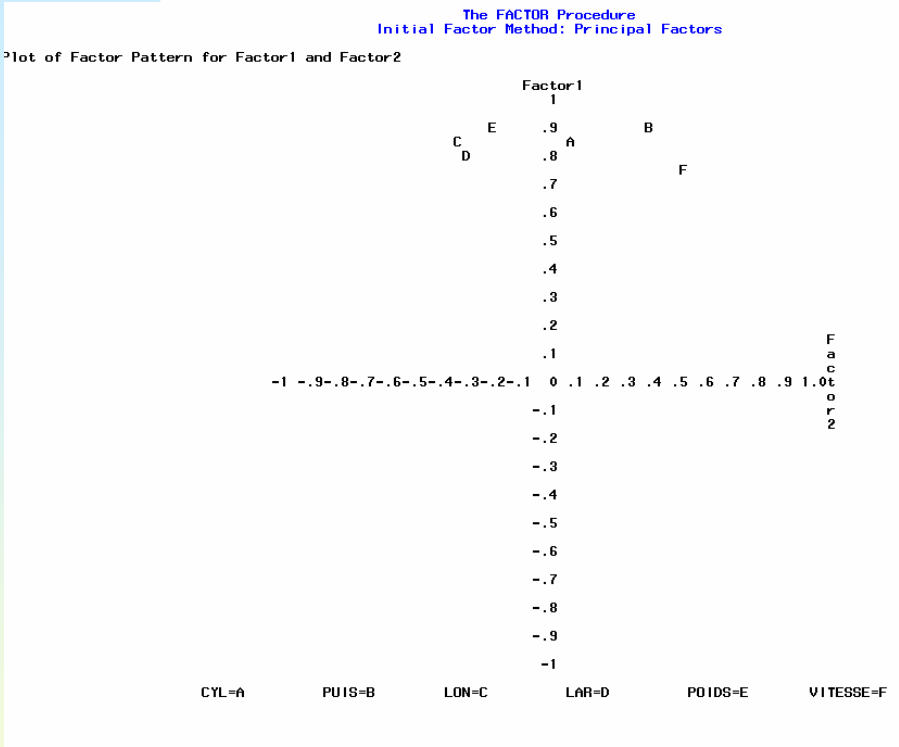
	Factor1	Factor2
LON	0.89468	0.30057
POIDS	0.83390	0.41311
LAR	0.81064	0.25939
VITESSE	0.22328	0.87775
PUIS	0.43083	0.85744
CYL	0.59671	0.61903

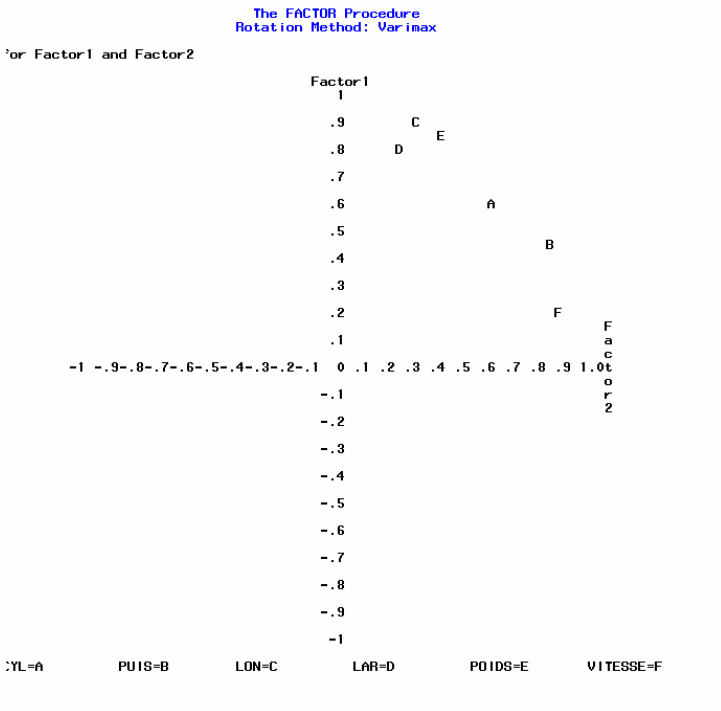
Variance Explained by Each Factor

Factor1	Factor2
2.7444961	2.2171241

Final Communality Estimates: Total = 4.961620

CYL	PUIS	LON	LAR	POIDS	VITESSE
0.73925315	0.92081180	0.89079542	0.72441494	0.86604445	0.82030045


The FACTOR Procedure
Initial Factor Method: Principal Factors
Plot of Factor Pattern for Factor1 and Factor2
Factor1
Factor2
CYL=A PUIS=B LON=C LAR=D POIDS=E VITESSE=F


The FACTOR Procedure
Rotation Method: Varimax
or Factor1 and Factor2
Factor1
Factor2
YL=A PUIS=B LON=C LAR=D POIDS=E VITESSE=F

B. Les classes latentes

- Lazarsfeld (1950): équivalent de l'analyse factorielle dans le cas entièrement qualitatif : les *p* variables observées sont qualitatives (souvent dichotomiques) et on postule l'existence d'une variable latente également qualitative à *k* modalités (les classes latentes).
- Une méthode de classification où les classes sont telles que les variables observées sont indépendantes dans chaque classe
- Un modèle particulier de mélange de distributions.

I. le modèle théorique dans le cas dichotomique

- *p* variables observées dichotomiques X_1, X_2,..,X_p prenant des valeurs 0 ou 1
- Y variable latente à *k* classes,
- p_{ij} la probabilité que X_i=1 pour un individu de la classe latente j .
- π_j probabilité *a priori* d'appartenir à la classe latente j

- Hypothèse d'indépendance conditionnelle:

$$f(x) = \sum_{j=1}^{k} \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i}$$

- probabilité *a posteriori* qu'un individu de vecteur x appartienne à la classe latente j

$$h(j / \mathrm{x}) = \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i} \; / \; f(\mathrm{x})$$

II L'estimation des paramètres

maximum de vraisemblance avec EM:

$$l = \sum_{h=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j \prod_{i=1}^{p} p_{ij}^{x_{ih}} (1 - p_{ij})^{1 - x_{ih}} \right)$$

$$\sum_{j=i}^{k} \pi_j = 1 \qquad \phi = l + \lambda \sum_{j=i}^{k} \pi_j$$

$$\frac{\partial \phi}{\partial \pi_j} = \sum_{h=1}^{n} \left(\prod_{i=1}^{p} p_{ij}^{x_{ih}} (1 - p_{ij})^{1 - x_{ih}} / f(\mathrm{x_h}) \right) + \lambda = \sum_{h=1}^{n} \frac{g(\mathrm{x_h} / j)}{f(\mathrm{x_h})} + \lambda$$

$$\frac{\partial \phi}{\partial p_{ij}} = \sum_{h=1}^{n} \pi_j \frac{\partial g(\mathrm{x_h} / j)}{\partial p_{ij}} / f(\mathrm{x_h})$$

$$\frac{\partial g(\mathrm{x_h} / j)}{\partial p_{ij}} = \frac{\partial}{\partial p_{ij}} \exp\left(\sum_{i=1}^{p} \left(x_{ih} \ln(p_{ij}) + (1 - x_{ih}) \ln(1 - p_{ij}) \right) \right) =$$

$$g(\mathrm{x_h} / j) \left\{ \frac{x_{ih}}{p_{ij}} - \frac{1 - x_{ih}}{1 - p_{ij}} \right\} = (x_{ih} - p_{ij}) g(\mathrm{x_h} / j) / p_{ij}(1 - p_{ij})$$

- D'où:

$$\frac{\partial \phi}{\partial p_{ij}} = \left(\pi_j / p_{ij}(1-p_{ij})\right) \sum_{h=1}^{n} (x_{ih} - p_{ij}) g(\mathrm{x_h} / j) / f(\mathrm{x_h})$$

- en introduisant les probabilités *a posteriori* :

$$h(j / \mathrm{x_h}) = \pi_j g(\mathrm{x_h} / j) / f(\mathrm{x_h})$$

- en annulant les dérivées :

$$\sum_{h=1}^{n} h(j / \mathrm{x_h}) = -\lambda \pi_j$$

$$\sum_{h=1}^{n} (x_{ih} - p_{ij}) h(\mathrm{x_h} / j) / p_{ij}(1-p_{ij}) = 0$$

$$\hat{\pi}_j = \sum_{h=1}^{n} h(j / \mathrm{x}_h) / n$$

$$\hat{p}_{ij} = \sum_{h=1}^{n} x_{ih} h(j / \mathrm{x}_h) / n\hat{\pi}_j$$

$$h(j / \mathrm{x}) = \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i} \; / \; f(\mathrm{x})$$

III Ajustement et choix de modèles

- Le test du G^2

$$G^2 = 2\sum_x n(\mathrm{x}) \ln\left(\frac{n(\mathrm{x})}{n\hat{f}(\mathrm{x})}\right)$$

- nombre de paramètres à estimer:

  k – 1 probabilités π_j et kp probabilités conditionnelles p_{ij}

  = k(p+1) – 1 paramètres.

- G^2 est à comparer à un khi-deux à

  $\nu = 2^p - k(p+1) + 1$ degrés de liberté
- Tests d'hypothèses emboîtées sur k
- En général si p est grand, les 2^p réponses possibles ont des effectifs insuffisants voire nuls: loi du khi-deux inappliquable.
- Simulations, ou bootstrap pour la loi de G^2

- Critères AIC d'Akaïké, ou BIC de Schwartz :

  $AIC = -2\ln(L) + 2(k(p+1) - 1)$

  $BIC = -2\ln(L) + \ln(n).(k(p+1) - 1)$

  vraisemblance pénalisée

- recherche de modèles parcimonieux, minimisant AIC ou BIC

IV Exemple (Bartholomew et Knott)

enquête sur les Attitudes Sociales Britanniques faite en 1990, concernant 1077 répondants avec 10 questions binaires d'opinions sur les attitudes sexuelles.

Sur les 1024 possibilités de réponse, seules 147 ont été observées:

1	90	0110011100
2	11	0110011000
3	9	0110111000
4	117	0110000000
5	18	0100000100
6	93	0100000000
7	19	0111111100
8	35	0010000000
9	21	0110001100
10	6	0111111110

- X1 Devrait-on rendre le divorce plus facile ?
- X2 Est ce que vous soutenez les lois contre la discrimination sexuelle ?
- X3 Opinion sur le sexe prénuptial : pas du tout opposé…… toujours opposé.
- X4 Opinion sur le sexe extra- marital
- X5 Opinion sur les relations sexuelles entre personnes de même sexe .
- X6 Doit-on permettre aux homosexuels d'enseigner dans les écoles ?
- X7 Doit-on permettre aux homosexuels d'enseigner dans l'enseignement supérieur ?
- X8 Doit -on permettre aux homosexuels d'occuper des fonctions officielles ?
- X9 Un couple de lesbiennes devrait–il avoir le droit d'adopter des enfants ?
- X10 Un couple d'homosexuels mâles devrait–il avoir le droit d'adopter des enfants ?

Nb. de classes	AIC	BIC
2	9328	9432
3	8946	9105
4	**8850**	**9064**
5	8852	9121

$\hat{\pi}_1 = 0.4611 \quad \hat{\pi}_2 = 0.0139 \quad \hat{\pi}_3 = 0.4169 \quad \hat{\pi}_4 = 0.1081$

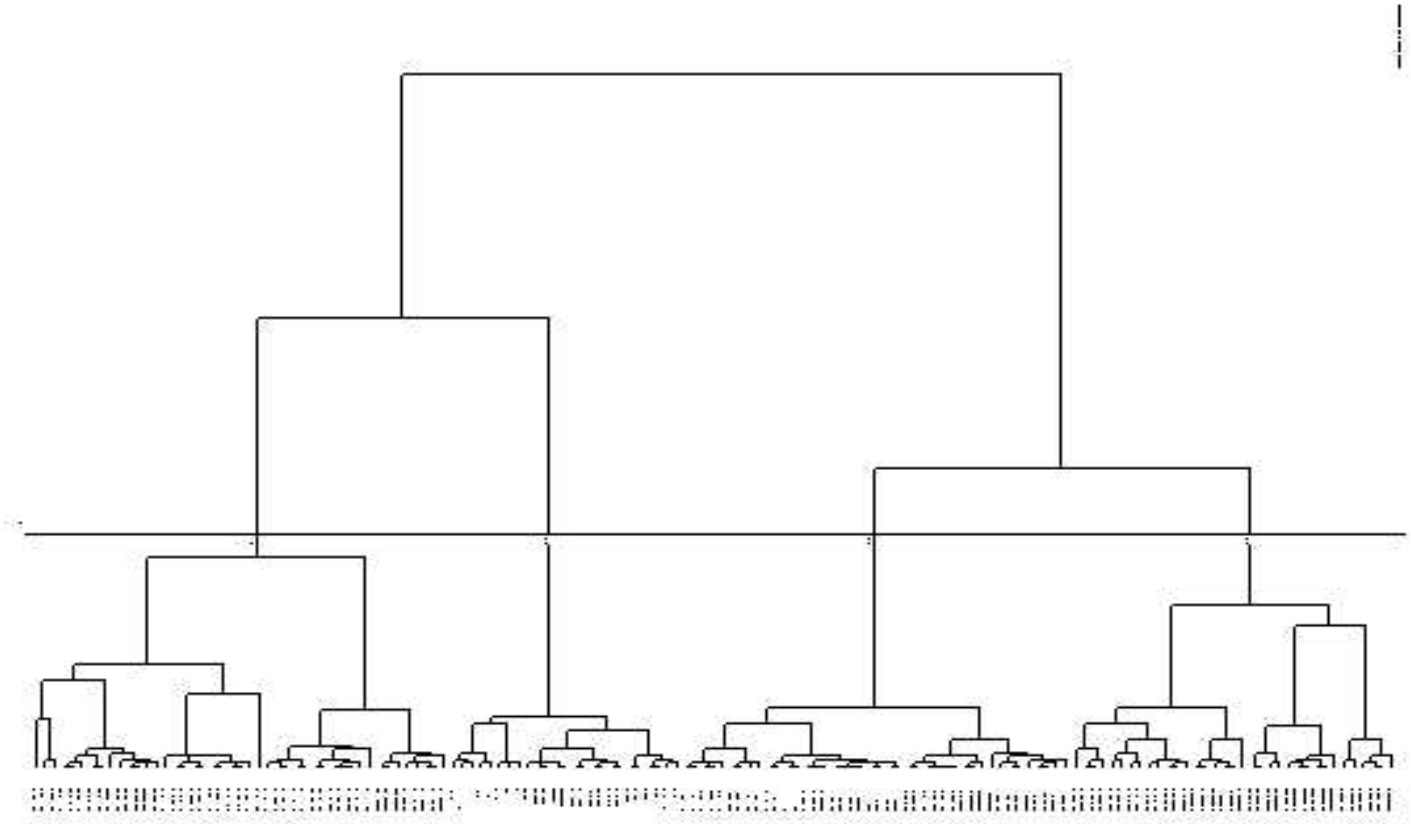
estimations des probabilités de donner la réponse oui(1) à chacune des 10 questions , conditionnellement aux classes latentes.

	classe 1	classe 2	classe 3	classe 4
X1	0.1360	0.0667	0.0947	0.2144
X2	0.7656	0.6000	0.8712	0.9246
X3	0.6319	0.8667	0.8620	0.9635
X4	0.0822	0.2667	0.1319	0.3089
X5	0.0681	0.6000	0.3822	0.8299
X6	0.0081	0.0000	0.8721	1.0000
X7	0.0589	0.2000	0.9829	1.0000
X8	0.2077	0.2667	0.9141	1.0000
X9	0.0463	1.0000	0.1071	0.9790
X10	0.0000	1.0000	0.0000	0.8505

la classe 1 est non-permissive, la classe 2 (de très faible effectif) se distingue par une grande permissivité pour l'adoption par des homosexuels mais est très négative sur les items 6, 7 et 8. La classe 3 est permissive sur tout sauf l'adoption, la classe 4 est permissive à peu près sur tous les items.

OBS.FREQ.	E(FREQ)	LAT. CLASS	RESPONSE VECTOR
90	114.608	3	0110011100
11	10.825	3	0110011000
9	6.661	3	0110111000
117	125.255	1	0110000000
18	19.173	1	0100000100
93	72.966	1	0100000000
19	10.831	3	0111111100
35	38.337	1	0010000000
21	18.862	3	0110001100
6	4.348	4	0111111110
14	16.949	3	0010011100
1	2.737	3	0111001100
2	0.315	3	0111001110
15	11.223	1	0111000000
11	9.161	1	0110100000
1	1.583	3	0010101100
3	1.549	3	0110101000
32	18.344	3	0100011100
1	0.142	1	1011000100
27	33.119	1	0110000100
8	7.973	4	0110011111
95	71.011	3	0110111100
7	3.887	3	0100001100
40	38.911	4	0110111111
2	2.252	3	0100011110
13	15.139	3	0110011110

Comparaison avec une classification classique



Comparaison avec une classification classique

- ACM, classification hiérarchique (Ward), coupure en 4 classes et nuées dynamiques

	CLASSE HIER. 1	CLASSE HIER. 2	CLASSE HIER. 3	CLASSE HIER. 4	ENSEMBLE
CLASSE LAT. 1	3	0	67	421	491
CLASSE LAT. 2	0	14	0	1	15
CLASSE LAT. 3	419	0	42	0	461
CLASSE LAT. 4	6	99	5	0	110
ENSEMBLE	428	113	114	422	1077

C. Traits latents

- “Latent trait models” , “ item response theory ”
- Recherche de variables latentes continues
- conditionnellement à un vecteur **y** de q variables latentes, les p variables manifestes (en général dichotomiques) sont des Bernoulli indépendantes.

$$P(x_i = 1) = \pi_i(\mathrm{y})$$

$$\ln\left(\frac{\pi_i(\mathrm{y})}{1-\pi_i(\mathrm{y})}\right) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j$$

- modèle “ logit-probit ” si la distribution *a priori* de **y** est multinormale centrée réduite
- Le modèle de Rasch est celui où q=1 (une seule variable latente) avec des pentes α_{ij} toutes égales.
- Les α_{ij} s ’interpretent comme des scores pour les catégories des variables manifestes.
- Bartholomew a montré que les scores obtenus par l’ACM sont une approximation au premier ordre des α_{ij} . Voir Aitkin et al. (RSA 1987,3,53-82)

Sortie partielle du modèle à deux facteurs

POSTERIOR ANALYSIS:

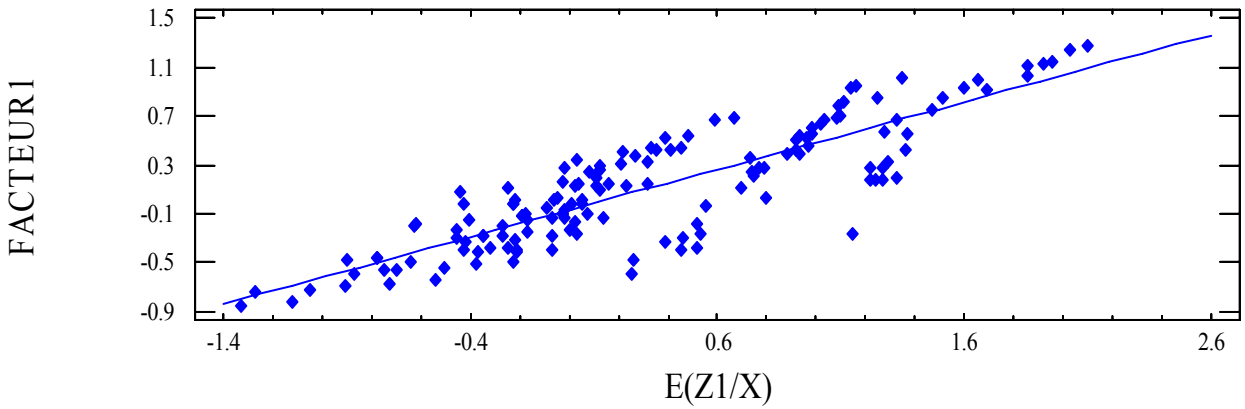
OBS	EXPECT	E(Z1/X)	SD1	E(Z2/X)	SD2	+RESP	RESPONSE PATTERN
29	27.8	-1.33	0.74	-0.74	0.85	0	0000000000
93	75.6	-1.27	0.71	-0.48	0.81	1	0100000000
5	3.4	-1.12	0.72	-1.00	0.85	1	1000000000
9	8.7	-1.05	0.70	-0.75	0.81	2	1100000000
3	1.5	-0.91	0.69	-0.96	0.81	1	0001000000
18	13.7	-0.90	0.63	-0.01	0.66	2	0100000100
3	2.2	1.66	0.44	-0.38	0.77	8	0111011111
2	0.1	1.69	0.44	-0.79	0.69	8	1011011111
1	0.7	1.86	0.45	-0.36	0.81	8	1010111111
40	27.9	1.86	0.44	0.19	0.89	8	0110111111
12	6.4	1.92	0.45	-0.03	0.87	9	1110111111
1	1.6	1.96	0.46	-0.23	0.85	8	0011111111
18	16.5	2.03	0.48	0.13	0.90	9	0111111111
5	4.1	2.10	0.50	-0.11	0.88	10	1111111111

ACM

HISTOGRAMME DES 10 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	0.3384	33.84	33.84	**********
2	0.1425	14.25	48.09	**********
3	0.1085	10.85	58.94	**********
4	0.1005	10.05	68.99	**********
5	0.0874	8.74	77.73	**********
6	0.0787	7.87	85.61	**********
7	0.0617	6.17	91.77	**********
8	0.0366	3.66	95.44	**********
9	0.0287	2.87	98.31	**********
10	0.0169	1.69	100.00	**********

- Corrélation de 0.934 entre E(Z1/X) et le premier facteur de l'ACM.



Conclusion

- Peu enseignés en France, les modèles à variables latentes méritent une attention particulière et peuvent être des compléments à des analyses plus classiques.
- Optique exploratoire ou confirmatoire.
- Les résultats d ’ACM ou de classification peuvent servir d ’initialisation aux modèles à variables latentes,
- Mêmes critiques que celles adressées à l’analyse factorielle vis à vis des méthodes de type ACP : problèmes d’identification, d’existence des variables latentes, de convergence des algorithmes...