Les classes latentes

Gilbert Saporta

Neuvièmes Journées d'Etude en Statistique- SFdS
Modèles statistiques pour données qualitatives, Marseille, Octobre 2000

LES CLASSES LATENTES

Gilbert Saporta

Les modèles à variables latentes ont été introduits dans les sciences humaines dès le début du siècle avec l'analyse factorielle de Spearman puis de Thurstone. Ils postulent l'existence de variables inobservables directement (appelées facteurs ou variables latentes) telles l'intelligence, l'engagement religieux etc., mais dont on peut mesurer ou observer des conséquences, ou effets, comme la fréquentation des lieux de culte, la réussite à certains tests…
L'hypothèse fondamentale est que les covariations entre variables observées (dites également « variables manifestes ») s'expliquent par la dépendance de chaque variable observée avec les variables latentes.
Connaître les variables latentes permettrait donc de diminuer les corrélations entre variables observées, d'où le principe fondamental d'indépendance conditionnelle : *les variables observées sont indépendantes conditionnellement aux variables latentes.* L'analyse en facteurs communs et spécifiques en est le cas particulier le plus connu, où variables observables et facteurs sont tous quantitatifs. Le tableau suivant présente les différentes situations :

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

Introduite principalement par Lazarsfeld vers 1950, l'analyse en classes latentes est considérée comme l'équivalent de l'analyse factorielle dans le cas entièrement qualitatif : les *p* variables observées sont qualitatives (souvent dichotomiques, mais pas nécessairement) et on postule l'existence d'une variable latente également qualitative à *k* modalités (les classes latentes).

Mais l'analyse en classes latentes peut être vue également comme une méthode de classification où les classes seraient telles qu'à l'intérieur de chacune, les variables observées seraient indépendantes, ou un modèle particulier de mélange de distributions.

On trouvera un exposé élémentaire de la théorie dans le petit livre de McCutcheon (1987), l'ouvrage de Bartholomew et Knott (1999) resituant l'analyse en classes latentes dans un contexte plus vaste. L'exposé qui suit s'inspire largement de ces deux ouvrages.

I. le modèle théorique dans le cas dichotomique

Soient *p* variables observées dichotomiques X_1, X_2,..,X_p prenant des valeurs 0 ou 1, et Y la variable latente à *k* classes, on notera p_{ij} la probabilité que X_i=1 pour un individu de la classe latente j . Si π_j est la probabilité *a priori* d'appartenir à la classe latente j, l'hypothèse d'indépendance conditionnelle donne pour le vecteur des variables observées *x*:

$$f(x) = \sum_{j=1}^{k} \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i} \quad (6.1)$$

On en déduit que la probabilité *a posteriori* qu'un individu de vecteur x appartienne à la classe latente j est :

$$h(j / \mathrm{x}) = \pi_j \prod_{i=1}^{p} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i} / f(\mathrm{x}) \quad (6.2)$$

La formule précédente permet donc d'affecter un individu à la classe latente la plus probable. Le problème statistique se ramène donc à estimer les paramètres π_j et p_{ij}, et à tester l'ajustement du modèle.
Le modèle de classes latentes peut s'étendre sans difficultés à des variables observées à plus de deux catégories, mais ne sera pas développé ici.

II L'estimation des paramètres

On dispose d'un n-échantillon d'observations et on note x_{ih} la valeur prise par X_i pour l'individu *h*. On utilise la méthode du maximum de vraisemblance : d'après l'équation (1.1), la log-vraisemblance vaut :

$$l = \sum_{h=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j \prod_{i=1}^{p} p_{ij}^{x_{ih}} (1 - p_{ij})^{1-x_{ih}} \right) \quad (6.3)$$

Depuis Goodman, la maximisation de *l* s'effectue à l'aide de la méthode itérative EM, qui semble la mieux adaptée. En voici les étapes essentielles (Bartholomew et Knott):

Comme $\sum_{j=i}^{k} \pi_j = 1$, on maximise le lagrangien $\phi = l + \lambda \sum_{j=i}^{k} \pi_j$. Les dérivées partielles sont alors :

$$\frac{\partial \phi}{\partial \pi_j} = \sum_{h=1}^{n} \left(\prod_{i=1}^{p} p_{ij}^{x_{ih}} (1-p_{ij})^{1-x_{ih}} \; / \; f(\mathrm{x_h}) \right) + \lambda = \sum_{h=1}^{n} \frac{g(\mathrm{x_h} / j)}{f(\mathrm{x_h})} + \lambda \quad (6.4)$$

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

Comme :

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

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

on a :

$$\frac{\partial \phi}{\partial p_{ij}} = \Big(\pi_j / \; p_{ij}(1-p_{ij}) \Big) \sum_{h=1}^{n} (x_{ih} - p_{ij}) g(\mathrm{x_h} / j) / \; f(\mathrm{x_h}) \qquad (6.5)$$

Ces équations se simplifient avec la formule de Bayes, en introduisant les probabilités *a posteriori* $h(j / \mathrm{x_h}) = \pi_j g(\mathrm{x_h} / j) / \; f(\mathrm{x_h})$, et en annulant les dérivées on trouve :

$\sum_{h=1}^{n} h(j / \mathrm{x_h}) = -\lambda \pi_j$ d'où en sommant sur j , et puisque la somme des π_j vaut 1, on trouve que $\lambda = -n$.

La première équation d'estimation est donc :

$$\hat{\pi}_j = \sum_{h=1}^{n} h(j / \mathrm{x_h}) / n \qquad (6.6)$$

La deuxième :

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

soit $$\hat{p}_{ij} = \sum_{h=1}^{n} x_{ih} h(j / \mathrm{x_h}) / n\hat{\pi}_j \qquad (6.7)$$

L'algorithme EM se déroule alors de la façon suivante :
on se donne arbitrairement un ensemble de probabilités *a posteriori* $h(j / \mathrm{x_h})$,
ensuite on utilise les équations 6.6 et 6.7 pour avoir des premières estimations des π_i et des p_{ij}, que l'on injecte dans 6.2 pour avoir de nouvelles valeurs des $h(j / \mathrm{x_h})$, etc.
La procédure converge, mais il peut y avoir des extrema locaux.

On peut, grâce aux équations de la vraisemblance, obtenir des erreurs standard asymptotiques, mais divers auteurs conseillent plutôt d'utiliser le bootstrap, en particulier si n est faible.

III Ajustement et choix de modèles

Une fois les paramètres estimés, on peut alors comparer les fréquences observées $n(x)$ des différents vecteurs x possibles (2^p au maximum) de variables observées (depuis 1111 jusqu'à 0000 si p=4 par exemple), avec leurs espérances données par $n\hat{f}(x)$ de l'équation 6.1.

On compare alors $G^2 = 2\sum_x n(x)\ln(\frac{n(x)}{n\hat{f}(x)})$ à un khi-deux à $\nu=2^p - k(p+1) + 1$ degrés de liberté si toutes les combinaisons de réponse ont été observées avec un effectif suffisant. Il y a en effet k – 1 probabilités π_j , et kp probabilités conditionnelles p_{ij} à estimer, soit k(p+1) – 1 paramètres.
Le modèle d'indépendance conditionnelle à k classes latentes est acceptable si G^2 est inférieur à un seuil.

En analyse exploratoire, on utilisera la même statistique pour choisir le nombre de classes : un usage courant consiste à tester des modèles emboîtés à 2, 3, 4 classes etc., et à s'arrêter dès que l'on trouve une valeur acceptable, car en général plus le nombre de classes est grand , plus le modèle s'ajuste bien.
L'utilisation de ce test est cependant problématique si p est grand, car il faut que les 2^p profils de réponse soient observés avec un effectif suffisant, sinon la distribution du khi-deux n'est plus applicable, même avec des regroupements. Certains auteurs préconisent d'utiliser des simulations bootstrap pour approcher la vraie distribution de G^2.
On préférera dans ces cas-là utiliser les critères AIC d'Akaïké, ou BIC de Schwartz :

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

pour comparer entre eux les modèles, afin d'obtenir un compromis entre modèle bien ajusté et modèle parcimonieux (avec peu de paramètres). Le meilleur modèle est celui qui minimise AIC ou BIC.

IV Un exemple (d'après Bartholomew et Knott)

Les données proviennent d'une enquête sur les Attitudes Sociales Britanniques faite en 1990, concernant 1077 répondants avec 10 questions binaires d'opinions sur les attitudes sexuelles :

1. Devrait-on rendre le divorce plus facile ?
2. Est ce que vous soutenez les lois contres la discrimination sexuelle ?
3. Opinion sur le sexe prénuptial : pas du tout opposé…… toujours opposé.
4. Opinion sur le sexe extra- marital : pas du tout opposé…… toujours opposé.
5. Opinion sur les relations sexuelles entre personnes de même sexe : pas du tout opposé…… toujours opposé.
6. Doit-on permettre aux homosexuels d'enseigner dans les écoles ?
7. Doit-on permettre aux homosexuels d'enseigner dans l'enseignement supérieur ?
8. Doit-on permettre aux homosexuels d'occuper des fonctions officielles ?
9. Doit-on permettre à un couple de lesbiennes d'adopter des enfants ?
10. Doit-on permettre à un couple d'homosexuels mâles d'adopter des enfants ?

Sur les 1024 possibilités de réponse, seules 147 ont été observées. Le tableau 6.1 donne ces combinaisons avec leurs fréquences :

Tableau 6.1

1	90	0110011100	76	4	0010011110
2	11	0110011000	77	1	1011000000
3	9	0110111000	78	1	1010011100
4	117	0110000000	79	1	1010100000
5	18	0100000100	80	1	1010011000
6	93	0100000000	81	1	1111100000
7	19	0111111100	82	2	0011011100
8	35	0010000000	83	1	1111001000
9	21	0110001100	84	1	1111101000
10	6	0111111110	85	2	1100011100
11	14	0010011100	86	1	0011000111
12	1	0111001100	87	4	0111000100
13	2	0111001110	88	3	0100111111
14	15	0111000000	89	3	0111011110
15	11	0110100000	90	4	0110100011
16	1	0010101100	91	2	0100100000
17	3	0110101000	92	1	1111111110
18	32	0100011100	93	1	1110011000
19	1	1011000100	94	4	1010000000
20	27	0110000100	95	1	0001000110
21	8	0110011111	96	1	1010111111
22	95	0110111100	97	3	1110111100
23	7	0100001100	98	1	1111001100
24	40	0110111111	99	1	0011001100
25	2	0100011110	100	1	1010001100
26	13	0110011110	101	4	0110010100
27	1	0010010100	102	2	1110111110
28	3	0110001110	103	1	0111001000
29	1	1011011100	104	9	0110000010
30	18	0111111111	105	2	0000000110
31	9	1100000000	106	2	0111011000
32	3	0001000000	107	2	0010100000
33	12	1110111111	108	2	0000001000
34	2	0110011010	109	1	0010001000
35	29	0000000000	110	1	0111100000
36	5	1000000000	111	1	0110111010
37	2	1010111110	112	1	0111000110
38	3	0100011000	113	1	0110010000
39	5	1111111111	114	1	0110100110
40	10	0000011100	115	1	0000001111
41	1	1111111000	116	1	0100111110
42	3	1110100000	117	1	0000000010
43	14	0110111110	118	1	1011000110
44	15	0111011100	119	1	1000011110
45	1	1100100100	120	2	0111101100
46	2	0110101110	121	1	0010000111
47	13	1110011100	122	2	1111000100
48	1	1110100100	123	2	0011000000
49	3	0110100100	124	1	0000000011
50	2	0100010000	125	2	1110011110
51	17	1110000000	126	1	1010000100
52	2	1011011111	127	1	0010110110
53	8	1110000100	128	2	0111100100
54	4	0100001000	129	3	0111011111
55	4	0110101100	130	2	0110000110
56	5	1110001000	131	1	0110001111
57	3	0010011111	132	1	0000100100
58	1	1111111101	133	1	0010100011
59	11	0010000100	134	1	1010010100

60	1	1100011000	135	1	0100100100
61	3	1110001100	136	3	1111111100
62	1	0111000011	137	1	1010111100
63	1	1010100011	138	1	1110000010
64	2	0111100011	139	1	0010011000
65	1	1111100100	140	1	0110101011
66	3	1111011100	141	1	0111101110
67	1	0110111101	142	1	1010011110
68	2	0010000010	143	1	1010000010
69	2	1111000000	144	1	0100000010
70	8	0010001100	145	1	0011111111
71	5	0100111100	146	1	0100011111
72	1	0011100000	147	1	0010111110
73	9	0110001000			
74	6	0000000100			
75	8	0010111100			

Les calculs qui suivent ont été obtenus avec le logiciel LATC, disponible à l'adresse http://www.arnoldpublishers.com/support/lvmfa2.htm
Le modèle le plus parcimonieux est celui à quatre classes comme le montre le tableau 6.2.

Tableau 6.2

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

Le tableau 6.3 donne les probabilités estimées des 4 classes :

Tableau 6.3

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

Le tableau 6.4 contient les estimations des probabilités $\hat{p}_{ij}$ de donner la réponse oui(1) à chacune des 10 questions , conditionnellement aux classes latentes.

Tableau 6.4

	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

On peut alors interpréter les classes comme suit : 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.
Le tableau 6.5 (sortie partielle) donne les effectifs espérés et les classes les plus probables :

Tableau 6.5

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

V Comparaison avec la classification euclidienne

En effectuant une classification euclidienne selon la technique classique suivante : analyse des correspondances multiples (pondérée), classification ascendante hierarchique avec la méthode de Ward, coupure en 4 classes, puis itération par une méthode de nuées dynamiques, on obtient une classification assez voisine, mais où la classe latente 3 n'est pas bien reconnue.

Tableau 6.6

POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU DU HAUT DE LA HIERARCHIE
314	0.02955	**********
216	0.04222	**********
453	0.04768	**********
398	0.06333	**********
565	0.08903	**********
512	0.13448	**********
1077	0.20847	**********

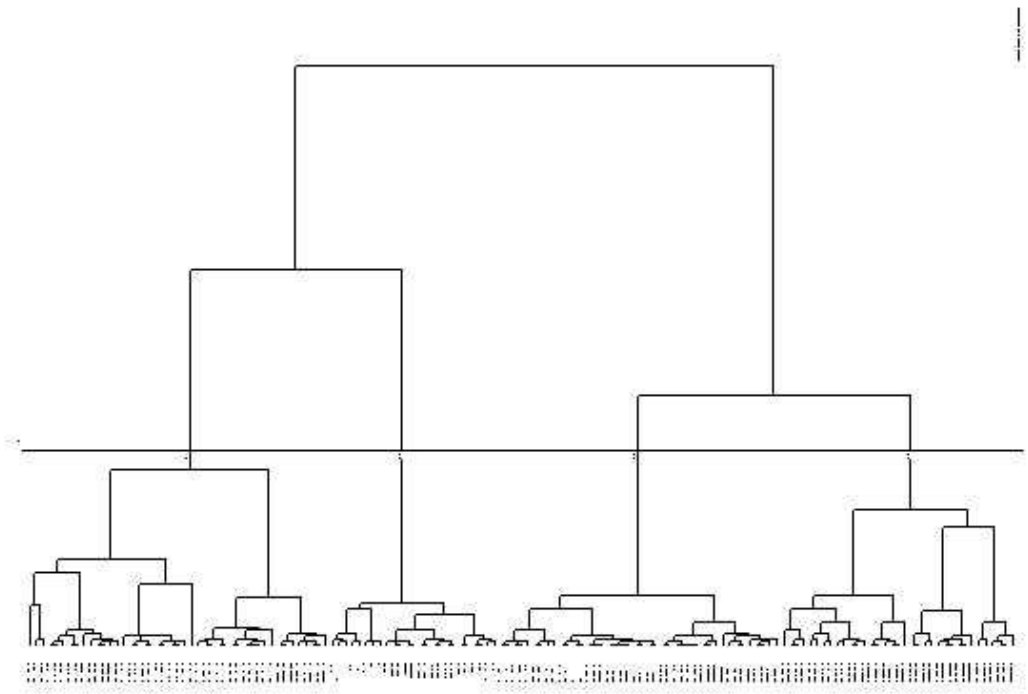

Figure 6.1

Tableau 6.7

COLONNE CLASSE HIER. / LIGNE CLASSE LAT.	CLASSE HIER. 1	CLASSE HIER. 2	CLASSE HIER. 3	CLASSE HIER. 4	ENSEMBLE
CLASSE LAT. 1	3	0	67	421	491
	0.70	0.00	58.77	99.76	45.59
	0.61	0.00	13.65	85.74	100.00
CLASSE LAT. 2	0	14	0	1	15
	0.00	12.39	0.00	0.24	1.39
	0.00	93.33	0.00	6.67	100.00
CLASSE LAT. 3	419	0	42	0	461
	97.90	0.00	36.84	0.00	42.80
	90.89	0.00	9.11	0.00	100.00
CLASSE LAT. 4	6	99	5	0	110
	1.40	87.61	4.39	0.00	10.21
	5.45	90.00	4.55	0.00	100.00
ENSEMBLE	428	113	114	422	1077
	100.00	100.00	100.00	100.00	100.00
	39.74	10.49	10.58	39.18	100.00

VI L'analyse des traits latents

La recherche de variables latentes qui ne sont plus des classes, mais des variables numériques continues, conduit à ce que l'on appelle les modèles à facteurs latents ou « latent traits models », auxquels on se réfère dans la littérature psychométrique sous le nom d'« item response theory ».
Le modèle consiste à exprimer que conditionnellement à un vecteur **z** de q variables latentes, les variables manifestes (ici dichotomiques) sont des variables de Bernoulli indépendantes avec $P(x_i = 1) = \pi_i(\mathrm{z})$, où $\pi_i(\mathrm{z}) = \pi_i(z_1,..,z_q)$ est la fonction de réponse pour l'item i. En général on prend pour cette fonction une forme logistique $\ln(\frac{\pi_i(\mathrm{z})}{1-\pi_i(\mathrm{z})}) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} z_j$.
En ajoutant que la distribution *a priori* de **z** est multinormale centrée réduite, on a le modèle dit « logit-probit ». Le modèle de Rasch est celui où q=1 et où les pentes α_{i1} sont toutes égales. Les parametres sont estimés par la méthode EM comme pour les classes latentes.
En reprenant les mêmes données qu'au paragraphe précédent, on trouve en ayant choisi q=2, les résultats suivants, grâce au programme LATV, disponible à la même adresse que LATC :

Tableau 6.8

MAXIMUM LIKELIHOOD ESTIMATES OF ITEM PARAMETERS AND STANDARD DEVIATIONS

item	α_{i0}	s.d.	α_{i1}	s.d.	α_{i2}	s.d.	P(X=1/Z=0)
1	-2.00	0.11	0.29	0.12	-0.30	0.16	0.12
2	1.71	0.10	0.32	0.12	0.47	0.13	0.85
3	1.69	0.13	1.32	0.16	0.18	0.15	0.84
4	-2.08	0.12	0.81	0.13	-0.09	0.14	0.11
5	-1.26	0.12	1.84	0.18	0.75	0.17	0.22
6	0.55	0.25	6.24	2.03	6.30	1.95	0.63
7	2.56	0.52	6.24	1.07	6.30	1.33	0.93
8	1.32	0.18	2.49	0.29	2.37	0.28	0.79
9	-3.76	1.06	4.12	1.32	-0.65	0.56	0.02
10	-9.66	12.83	6.91	10.70	-0.45	0.47	0.00

On utilise généralement pour « estimer » les facteurs latents, les espérances *a posteriori* E(Z1/X) et E(Z2/X) . Le tableau 6.9 donne est une sortie partielle fournissant ces valeurs, ainsi que les écart-types *a posteriori* et les effectifs espérés.

Tableau 6.9

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
6	3.4	-0.87	0.64	-0.11	0.67	1	0000000100
2	0.5	-0.78	0.63	0.36	0.66	2	0100010000
4	3.9	-0.78	0.63	0.36	0.66	2	0100001000
2	0.8	-0.75	0.63	0.27	0.66	1	0000001000
35	33.6	-0.73	0.66	-0.89	0.77	1	0010000000
117	101.2	-0.70	0.65	-0.70	0.74	2	0110000000
2	2.6	-0.64	0.63	-0.55	0.70	2	0100100000

Une analyse des correspondances multiples appliquée au même tableau donne pour l'essentiel un premier axe correspondant à un facteur général de permissivité, avec un effet Guttmann sur les autres axes : l'ACM ne révèle donc qu'une dimension.

Tableau 6.10

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	*

Les coordonnées sur le premier axe de l'ACM sont bien corrélées, r=0.934, avec les valeurs estimées de E(Z1/X), ce qui correspond aux observations de Aitkin et al. (1987) et à ce qu'affirment Bartholomew et Knott (1999) : *«les scores obtenus par l'ACM sont une approximation au premier ordre des* α_{ij} *»*

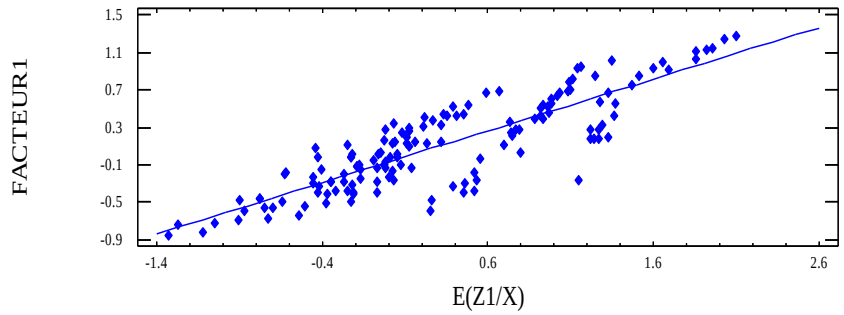

Figure 6.2

Conclusion

Peu enseignés en France, mais très populaires dans les pays anglophones et de l'Europe du Nord, les modèles à variables latentes méritent une attention particulière et peuvent être des compléments à des analyses plus classiques. Ils peuvent servir dans une optique exploratoire ou confirmatoire, mais souffrent des critiques adressées classiquement à l'analyse factorielle vis à vis des méthodes de type ACP : problèmes d'identification, d'existence des variables latentes qui ne sont jamais que des constructions, ainsi que de la non-convergence des algorithmes dans certains cas, ou de la convergence vers des extrema locaux.
Ces méthodes ne sont pas disponibles dans les grands progiciels tels SAS, SPSS, mais il existe des programmes gratuits simples d'emploi, bien que rustiques (cf. Bibliographie internet).

Bibliographie

Aitkin M., Francis B. et Raynal N. Une étude comparative d'analyses des correspondances ou de classifications et des modèles de variables latentes ou de classes latentes, ***Revue de Statistique Appliquée***, 35, 3, 53-82, 1987

Bartholomew, et D.J.& Knott, M. ***Latent Variable Models and Factor Analysis*** , 2nd edition, Arnold, 1999.

Dayton, C.M, ***Latent Class Scaling Analysis*** (Sage University Paper Series on Quantitative Applications in the Social Sciences n°126) , SAGE publications ,1998.

Everitt, B.S., ***An Introduction to Latent Variable Models*** , Chapman & Hall, London, 1984.

Goodman, L.A., Exploratory latent structure analysis using both identifiable and unidentifiable models, ***Biometrika*** , 61, 215- 231, 1974.

Goodman, L.A., The Analysis of systems of qualitative variables when some of the variables are unobservable. Part I : A modified latent structure approach , ***American Journal of Sociology***, 79, 1179- 1259, 1974.

Goodman, L.A .« Exploratory latent structure models using both identifiable and unidentifiable models, ***Biometrika***, 61, 215-231, 1974.

Hagenaars J.A. McCutcheon A.L. (eds.), ***Applied latent class analysis*** Cambridge University Press, 2002.

Heinen T., ***Latent Class and Discrete Latent Trait Models, Similarities and Differences***, (Advanced Quantitative Techniques in the Social Sciences Series), SAGE Publications, 1996 .

Lazarsfeld, P.F , The logical and mathematical foundation of latent structure analysis , In S. Stouffer (Ed.), ***Measurement and Prediction***, 362-412, Princeton, N.J :Princeton University Press,1950.

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.

McLachlan, G.J., et Krishnan, J., ***The EM algorithm and Extensions*** , Wiley, New York, 1997.

Vermunt, J.K., et Magidson, J., « Exploratory Latent Class Cluster, Factor and Regression Analysis : The Latent Gold approach ». Article présenté à la conférence de EMPS'99, Lueneburg, Germany, 1999.

Sur Internet :

Site du logiciel gratuit « lvmfa2 » (cf.Bartholomew & Knott) :
http://www.arnoldpublishers.com/support/lvmfa2.htm.

Site du logiciel « Latent Gold » distribué par Statistical Innovations:
http://latentclass.com/

« Latent Class Analysis Website », par John Uebersax :
http://ourworld.compuserve.com/homepages/jsuebersax/

« LEM » logiciel gratuit par J.K.Vermunt de l'Université de Tilburg (KUB)
http://cwis.kub.nl/~fsw_1/mto/mto_snw.htm#software