

Latent Class Factor Models for Market Segmentation: an Application to Pharmaceutical Products

*Latent Class Factor Models per la Segmentazione del Mercato:
un'Applicazione ai Prodotti Farmaceutici*

Francesca Bassi

Dipartimento di Scienze Statistiche, via Battisti 241, 35133 Padova, bassi@stat.unipd.it

Riassunto:

La segmentazione di un mercato consiste nell'individuazione di gruppi omogenei di clienti, a questo fine numerose sono le tecniche statistiche proposte in letteratura, tra queste, la cluster analysis e i modelli a classi latenti. In questo lavoro si propone, per la segmentazione del mercato, un'estensione del modello a classi latenti tradizionale, denominato latent class factor model. Tale approccio risulta più parsimonioso e fornisce risultati più facilmente interpretabili dell'approccio tradizionale. Il modello è utilizzato per segmentare il mercato dei medici di base, clienti delle case farmaceutiche.

Keywords: latent class cluster model, K-means clustering, marketing strategies.

1.Introduction

Market segmentation is an essential element of marketing. Goods and services can no longer be sold without considering heterogeneity in customer needs. Segmentation is a grouping task for which a large variety of methods are available, including loglinear models, clustering methods and mixture models (Wedel and Kamakura, 2000).

Latent class (LC) analysis attempts to explain the observed association between variables in a multiway contingency table by introducing unobservable underlying classes (clusters). The LC approach to clustering is model-based, the fundamental assumption is that of local independence: units in the same latent class share a common joint probability distribution among the observed variables (Vermunt, 1997). LC clustering may be viewed as a probabilistic variant of K-means clustering and it provides a way not only to formalise the K-means approach in terms of a statistical model, but also to extend it in several directions. Magidson and Vermunt (2001) propose an extension of the traditional LC approach, called LC factor model, which increases the number of latent variables rather than that of latent classes. The basic LC factor model, which contains R mutually independent, dichotomous latent variables, has exactly the same number of parameters as a LC cluster model with $R+1$ classes.

The LC factor model is applied to identify segments in the pharmaceutical market. To propose the appropriate drug to the appropriate doctor, pharmaceutical industries cannot rely solely on sensations expressed by their agents, specific analyses are necessary. The pharmaceutical sector is characterised by a high level of competitiveness, more limited economic budgets than years ago and, at the same time, expensive sales and promotion activities. In this context, it is very important to understand which factors influence doctors in prescribing medicines, so to design appropriate marketing strategies.

2. Latent class factor models

An LC model for four nominal variables A, B, C and D , and one latent variable X , is:

$$\pi_{ijklt}^{ABCDX} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X} \quad (1)$$

where π_{ijklt}^{ABCDX} is the proportion of units in the five-way contingency table, π_t^X is the probability of being in latent class $t=1,2,\dots,T$; $\pi_{it}^{A|X}$ is the probability of obtaining response i to item A , from members of class t , $i=1,2,\dots,I$; and $\pi_{jt}^{B|X}$, $\pi_{kt}^{C|X}$, $\pi_{lt}^{D|X}$, $j=1,2,\dots,J$, $k=1,2,\dots,K$, $l=1,2,\dots,L$, are the conditional probabilities of items B, C and D . One goal of traditional LC analysis is to determine the smallest number of latent classes T which is sufficient to explain the associations observed among the items. The final step of LC analysis is to classify units into the appropriate latent class. For any response pattern $(ijkl)$, posterior membership probabilities may be estimated through the Bayes theorem. Cases are then assigned to the class for which the posterior probability is highest. Magidson and Vermunt (2001) refer to this as an LC cluster model because the goal of classification into T homogeneous groups is identical to that of cluster analysis. Rejection of a T -class LC model due to lack of fit means that the local independence assumption does not hold. The traditional model-fitting strategy is to fit a $T+1$ -class model to the data, but alternative strategies may be considered, to see if they lead to more parsimonious models, as well as models more congruent with initial hypotheses. Magidson and Vermunt (2001 and 2004) show that, by increasing dimensionality by adding latent variables rather than latent classes, the resulting LC factor model often fits data better than the LC cluster model with the same number of parameters. In addition, LC factor models are identified in some situations when the traditional LC model is not. Traditional LC models with four or more classes may be interpreted in terms of two or more component latent variables. For example, a latent variable X consisting of four classes can be represented in terms of two dichotomous latent variables V and W , using the following correspondences: $X=1$ corresponds with $V=1$ and $W=1$; $X=2$ with $V=1$ and $W=2$; $X=3$ with $V=2$ and $W=1$; $X=4$ with $V=2$ and $W=2$. For four nominal variables, the four-class LC model may be reparameterised as an LC factor model with two dichotomous latent variables, as follows:

$$\pi_{ijklrs}^{ABCDVW} = \pi_{rs}^{VW} \pi_{ijklrs}^{ABCD|VW} = \pi_{rs}^{VW} \pi_{irs}^{A|VW} \pi_{jrs}^{B|VW} \pi_{krs}^{C|VW} \pi_{lrs}^{D|VW} \quad (2)$$

Any LC model is equivalent to a loglinear one with unobserved variables:

$$\begin{aligned} \ln F_{ijklrs}^{ABCDVW} = & \lambda + \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ir}^{AV} + \lambda_{jr}^{BW} + \lambda_{kr}^{CV} + \\ & + \lambda_{lr}^{DV} + \lambda_{is}^{AW} + \lambda_{js}^{BW} + \lambda_{ks}^{CW} + \lambda_{ls}^{DW} + \lambda_{irs}^{AVW} + \lambda_{jrs}^{BVW} + \lambda_{krs}^{CVW} + \lambda_{lrs}^{DVW} \end{aligned} \quad (3)$$

where F_{ijklrs}^{ABCDVW} is the absolute frequency in the generic cell of a six-way contingency table and λ are the parameters defining a loglinear model (Haberman, 1979).

The basic LC factor model contains two or more dichotomous latent variables which are mutually independent of each other and which exclude higher-order interactions from the conditional response probabilities, from equation (3):

$$\ln F_{ijklrs}^{ABCDVW} = \lambda + \lambda_r^V + \lambda_s^W + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ir}^{AV} + \lambda_{jr}^{BV} + \lambda_{kr}^{CV} + \lambda_{lr}^{DV} + \lambda_{is}^{AW} + \lambda_{js}^{BW} + \lambda_{ks}^{CW} + \lambda_{ls}^{DW}$$

The basic LC factor model with R independent factors has the same number of distinct parameters as a traditional LC cluster model with $R+1$ classes. This offers a great advantage in parsimony and results are often easier to interpret.

3. Market segmentation for pharmaceutical products

The data was collected from 487 Italian general practitioners. On a seven-point scale, doctors expressed how important the following items were in inducing them to prescribe a drug proposed by a pharmaceutical industry: (1) attention of the industry for doctors' updating (ATT), (2) frequency and regularity of visits by pharmaceutical representatives (FRE), (3) assistance on diagnostic and therapeutic problems (ASS), (4) consideration for doctors' experience and suggestions (EXP), (5) quality of training of pharmaceutical representatives (QUA), (6) information on industry activities (INF), and (7) global quality of information and promotion activities (PRO).

LC cluster and factor models were estimated in order to identify market segments¹. Table 1 lists (left) the results obtained by estimating the LC cluster model which revealed the best fit to the data ($L^2 = 884,1580$, $p=0,99$, $BIC=-1.894,3726$). On the right, it lists the results obtained by estimating a basic LC 2-factor model, which shows a better fit ($BIC=-1.932,3203$) and is more parsimonious than the LC cluster with four classes (30 parameters to be estimated instead of 38).

Both models identify four segments but with different dimensions and characteristics. Looking at conditional probabilities on the left side of Table 1, names can be assigned to the clusters. In cluster 1 (33%) we find loyal practitioners who are very concerned about the frequency and regularity of visits and the quality of training of representatives, the attention of the pharmaceutical industry for doctors' updating, and the global quality of information and promotion activities. In cluster 2 (30% of doctors) contains practitioners uninterested in promotion and information by industries. In cluster 3 (23%) doctors can be defined as loyal and demanding at the same time; all items are important for them in order to choose among drugs. Lastly, in cluster 4 (14%) doctors are not very sensitive to representatives' activity, and sometimes even annoyed by it.

Referring to the results obtained with the LC 2-factor model, level 1 of both factors identifies practitioners sensitive to representatives' activities. Level 2 represents doctors who pay little attention to pharmaceutical representatives and their information activity. All practitioners, in any case, evaluate as important the frequency and regularity of visits and the global quality of training of representatives. Looking now more deeply into the four segments identified by the two binary factors, (1,1)² (31% of respondents) contains

¹ The software Latent Gold 4.0 was used. The author wishes to thank prof. Paolo Mariani for supplying the data.

² Digits refer to factor and level respectively.

loyal and demanding practitioners who are highly interested in all activities performed by the industries. (2,1) identifies loyal doctors (30%), an interesting segment for pharmaceutical industries, although not so sensitive to their activities as the first. In (2,2) (19%) we find practitioners uninterested in information and promotion. (1,2) (19%) identifies doctors sometimes annoyed by information and promotion activities.

Table 1 LC cluster and basic 2-factor models – estimation results³

	LC cluster				LC factor			
Size	0,3287	0,3028	0,2316	0,1369	Factor 1		Factor 2	
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Level 1	Level 2	Level 1	Level 2
ATT=1	0,0497	0,3540	0,0035	0,7719	0,0189	0,4459	0,1612	0,3834
ATT=2	0,2458	0,3931	0,0748	0,1985	0,1407	0,3456	0,2401	0,2474
ATT=3	0,7045	0,2529	0,9216	0,0296	0,8404	0,2086	0,5987	0,4073
FRE=1	0,0360	0,2376	0,0036	0,4770	0,0125	0,2907	0,1026	0,2294
FRE=2	0,1234	0,2481	0,0419	0,2541	0,0723	0,2481	0,1469	0,1810
FRE=3	0,8405	0,5144	0,9545	0,2689	0,9152	0,4612	0,7505	0,5896
ASS=1	0,1959	0,4925	0,0089	0,7951	0,1016	0,5508	0,2152	0,5031
ASS=2	0,3295	0,3245	0,0965	0,1727	0,2247	0,2765	0,2487	0,2535
ASS=3	0,4745	0,1830	0,8946	0,0321	0,6737	0,1727	0,5361	0,2434
EXP=1	0,2849	0,5437	0,0323	0,9087	0,1699	0,6123	0,2586	0,6028
EXP=2	0,2941	0,2701	0,1390	0,0812	0,2157	0,2244	0,2254	0,2114
EXP=3	0,4210	0,1862	0,8287	0,0101	0,6144	0,1633	0,5160	0,1859
QUA=1	0,0168	0,0671	0,0000	0,5613	0,0087	0,2022	0,0166	0,2477
QUA=2	0,1265	0,2292	0,0014	0,2972	0,0702	0,2345	0,0986	0,2382
QUA=3	0,8568	0,7037	0,9986	0,1415	0,9210	0,5633	0,8848	0,5141
INF=1	0,4761	0,5832	0,0611	0,9136	0,3064	0,6375	0,2800	0,7794
INF=2	0,2713	0,2466	0,1697	0,0760	0,2177	0,2067	0,2467	0,1567
INF=3	0,2526	0,1703	0,7692	0,0103	0,4759	0,1558	0,4733	0,0638
PRO=1	0,0161	0,3244	0,0000	0,9241	0,0056	0,4628	0,1143	0,4254
PRO=2	0,2244	0,4935	0,0031	0,0745	0,0951	0,3672	0,2605	0,1832
PRO=3	0,7595	0,1821	0,9969	0,0015	0,8992	0,1700	0,6252	0,3914

In our case, the LC 2-factor model performs better than the traditional LC cluster model: it fits the data better (the percentage of classification errors decreases from 16,54% to 0,09%) and it is more parsimonious. Moreover, identification of latent factors facilitates the interpretation of results, clusters are more neatly described, and marketing strategies can therefore be more efficiently designed.

References

- Haberman S. J. (1979) *Analysis of Qualitative Data*, Vol.2, Academic Press, New York.
- Magidson J, Vermunt J. K. (2001) Latent class factor and cluster models, bi-plots and related graphical displays, *Sociological Methodology*, 31, 223-264.
- Magidson J, Vermunt J. K. (2004) Latent class models in Kaplan D (ed.) *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Ch. 10, 175-198, Sage, Thousand Oaks.
- Vermunt J. K. (1997) *Loglinear Models for Event Histories*, Sage, Thousand Oaks.
- Wedel M., Kamakura W. A. (2000) *Market Segmentation: Concepts and Methodological Foundations*, Kluwer Academic, Boston.

³ Categories are rescaled to three to keep the dimension of the contingency table within reasonable limits.