Item selection by latent class-based methods: an application to nursing homes evaluation

Francesco Bartolucci, Giorgio E. Montanari, Silvia Pandolfi[1]

Department of Economics, Finance and Statistics
University of Perugia

AG DANK/BCS Meeting 2013

London, 8-9 November 2013

[1] pandolfi@stat.unipg.it

Outline

Motivation

The latent class model
- Basic notation
- Model assumption

Item selection procedure
- Inclusion-exclusion algorithm

Application to the ULISSE dataset
- Item selection
- Parameters estimates

Conclusions

Main references

Motivation

- The evaluation of nursing homes and the assessment of the quality of the health care provided to their patients are nowadays based on the administration of questionnaires
- These questionnaires are usually made of a large number of *polytomous items* that may lead to a lengthy and expansive administration
- Due to tiring effects, using several items may lead the respondent to provide inaccurate responses
- Methods which allow us to select *the smallest subset of items* useful for clustering are of interest
- These methods may lead to a reduction of the costs of the data collection process and a better quality of the collected data

Our contribution

- We adopt the algorithm for *item selection* proposed by Dean and Raftery (2010) which is based on the *Latent Class (LC) model* (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968; Goodman, 1974)
- The algorithm is aimed at finding the subset of items which provides the best value of the *Bayesian Information Criterion* index (BIC, Schwartz, 1978; Kass and Raftery, 1995)
- At the same time, the algorithm allows us to select the optimal number of latent classes of the LC model
- The implementation of the algorithm is based on a stepwise scheme that, starting from a reduced set of items, at each iteration performs both *inclusion and exclusion steps*
- We also include and additional step, as *random check*, aimed at initializing, with a large number of random starting values, the estimation algorithm, so as to prevent the problem of local maxima of the model log-likelihood

The ULISSE project

- The *ULISSE project* (Lattanzio *et al.*, 2010) has been carried out by a Research Group established by the Italian Ministry of Health and the Pfizer private firm
- The dataset is collected by the administration of a questionnaire concerning the *quality-of-life* of elderly patients hosted in nursing homes
- The project is based on a longitudinal survey that has been carried out since 2004, covering 17 Italian regions and 37 randomly chosen nursing homes
- We considered the data coming from the first wave of administration of the ULISSE questionnaire to a sample of 1739 patients
- The original questionnaire includes *75 polytomously-responded items*

- The items are included into different sections of the questionnaire, d, concerning:

  1. Cognitive Conditions (CC)
  2. Auditory And View Fields (AVF)
  3. Humor And Behavioral Disorders (HBD)
  4. Activities Of Daily Living (ADL)
  5. Incontinence (I)
  6. Nutritional Field (NF)
  7. Dental Disorders (DD)
  8. Skin Conditions (SC)

- The items are ordinal, with categories ordered with increasing difficulty in accomplishing tasks or severeness of the health conditions

- The LC model represents a useful tool of analysis of data collected by questionnaires made of polytomous items

- The use of the LC model is justified when the items measure one or more latent trait, such as the quality-of-life or the tendency toward a certain behavior

Basic notation

- n: sample size
- J: number of items
- Y_{ij}: response variable for subject i to item j ($i = 1, \ldots, n$, $j = 1, \ldots, J$)
- l_j: number of categories of every response variable, indexed from 0 to $l_j - 1$
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$: vector of all response variables for subject i

Model assumption

- U_i: discrete latent variable which has the same distribution for every subject i, based on k support points, labeled from 1 to k

- Each support point corresponds to a latent class in the population

- Assumption of *local independence*:

  for every subject i the random variables within $\boldsymbol{Y}_i$ are conditionally independent given U_i

- Assumption of *missing at random* (MAR; Rubin, 1976; Little and Rubin, 2002):

  the probability of the observed missingness pattern, given the observed and the unobserved data, does not depend on the unobserved data

Model parameters

- *Prior probability* (or weight) of each latent class:

$$\pi_u = p(U_i = u), \quad u = 1, \dots, k$$

- *Conditional response probabilities*:

$$\lambda_{j|u}(y) = p(Y_{ij} = y | U_i = u), \quad j = 1, \dots, J,\ u = 1, \dots, k,\ y = 0, \dots, l_j - 1$$

- The number of free parameters is:

$$g_k = (k - 1) + k \sum_j (l_j - 1)$$

- The assumption of *local independence* implies that:

$$p(\boldsymbol{y}_i|u) = p(\boldsymbol{Y}_i = \boldsymbol{y}_i|U_i = u) = \prod_j \lambda_{j|u}(y_{ij})$$

with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})$

- *Manifest probability* of response pattern $\boldsymbol{y}_i$ for subject i:

$$p(\boldsymbol{y}_i) = \sum_u p(\boldsymbol{y}_i|u)\pi_u$$

- *Posterior probability* that a subject with observed response configuration $\boldsymbol{y}_i$ belongs to latent class u:

$$q_i(u) = p(u|\boldsymbol{y}_i) = \frac{p(\boldsymbol{y}_i|u)\,\pi_u}{p(\boldsymbol{y}_i)}$$

- These posterior probabilities are used to allocate subjects in the different latent classes
- Subject i is assigned to the class with the largest posterior probability

- Given the independence of the subjects in the sample, the *log-likelihood* function of the model may be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{y}_i)$$

  - $\boldsymbol{\theta}$: vector of all free parameters affecting $p(\boldsymbol{y}_i)$

- $\ell(\boldsymbol{\theta})$ may be maximized wrt $\boldsymbol{\theta}$ by the *Expectation-Maximization (EM) algorithm* (Baum et al., 1970; Dempster et al., 1977)

- The EM algorithm alternates two steps (E-step, M-step) until convergence in $\ell(\boldsymbol{\theta})$ and is based on the *complete data log-likelihood*

Inclusion-Exclusion algorithm

- The algorithm is based on the method proposed by *Dean and Raftery (2010)* which assesses the importance of a certain item by comparing two LC models

- In the first model, the item is assumed to provide additional information about clustering allocation, beyond that contained in the already selected items

- in the second model, this item does not provide additional information useful for clustering and then it is independent of the latent variable defining the latent classes

- The two models are compared via *BIC index*, which is seen as an approximation of the Bayes Factor (Kass and Raftery, 1995)

Basic notation

- $\mathcal{I}$: full set of items
- $\mathcal{A}^{(0)}$: initial set of clustering items
- $k^{(0)}$: initial number of latent classes
- $\mathcal{A}^{(h)}$: set of items selected at the end of the hth iteration of the algorithm
- $k^{(h)}$: number of latent classes selected at the end of the hth iteration of the algorithm
- $\bar{\mathcal{A}}^{(h)}$: complement of $\mathcal{A}^{(h)}$ with respect to the full set of items
- $\hat{\ell}_k(\mathcal{A})$: maximum of the log-likelihood of the LC model applied to the data referred to the items in $\mathcal{A}$
- $g_k(\mathcal{A})$: corresponding number of free parameters

$$BIC_k(\mathcal{A}) = -2\ \hat{\ell}_k(\mathcal{A}) + g_k(\mathcal{A})\log(n)$$

$$BIC_{tot,k}(\mathcal{A}) = BIC_k(\mathcal{A}) + BIC_1(\bar{\mathcal{A}})$$

- *Inclusion step:*
  - Each item j in $\bar{\mathcal{A}}^{(h-1)}$ is singly proposed for inclusion in $\mathcal{A}^{(h)}$
  - The item with the smallest (negative) values of $BIC_{diff}(\mathcal{A}^{(h-1)}, j)$ is included in $\mathcal{A}^{(h)}$and $k^{(h)}$ is updated

$$BIC_{diff}(\mathcal{A}^{(h-1)}, j) = \min_{2 \le k \le k_{\max}} BIC_k(\mathcal{A}^{(h-1)} \cup j) - [BIC_1(j) + BIC_{k^{(h-1)}}(\mathcal{A}^{(h-1)})]$$
$$= \min_{2 \le k \le k_{\max}} BIC_{tot,k}(\mathcal{A}^{(h-1)} \cup j) - BIC_{tot,k^{(h-1)}}(\mathcal{A}^{(h-1)})$$

  - If no item yields a negative BIC_{diff}, then we set $\mathcal{A}^{(h)} = \mathcal{A}^{(h-1)}$
- *Exclusion step:*
  - Each item j in $\mathcal{A}^{(h)}$ is singly proposed for exclusion
  - The item with the highest (positive) value of $BIC_{diff}(\mathcal{A}^{(h)} \setminus j, j)$ is removed from $\mathcal{A}^{(h)}$ and $k^{(h)}$ is updated

$$BIC_{diff}(\mathcal{A}^{(h)} \setminus j, j) = BIC_{k^{(h)}}(\mathcal{A}^{(h)}) - [BIC_1(j) + \min_{2 \le k \le k_{\max}} BIC_k(\mathcal{A}^{(h)} \setminus j)]$$
$$= BIC_{tot,k^{(h)}}(\mathcal{A}^{(h)}) - \min_{2 \le k \le k_{\max}} BIC_{tot,k}(\mathcal{A}^{(h)} \setminus j)$$

  - If no item is found with a positive BIC_{diff}, then we set $\mathcal{A}^{(h)} = \mathcal{A}^{(h)}$

- The algorithm ends when no item is added to $\mathcal{A}^{(h)}$ and no item is removed from $\mathcal{A}^{(h)}$
- As in Dean and Raftery (2010), the *posterior probabilities* estimated at the end of the previous step of the algorithm are used to obtain the starting values for the EM algorithm involving the updated dataset, with one more or one less item
- At the end of both inclusion and exclusion steps we perform an additional *random check* in order to assess the convergence to the global maximum of the model log-likelihood:
  - We initialize the estimation algorithm of the current model from a large number of random starting values proportional to k
  - We take as final estimate the one corresponding to the highest log-likelihood value found at convergence of the EM algorithm

Application to the ULISSE dataset

- We estimate the LC model considering the full set of 75 item, $\mathcal{I}$, for a number of latent classes from 2 to $k_{\max}$, with $k_{\max} = 10$
- For each k, we consider $100 \times (k-1)$ *random initialization* of the EM algorithm, and we take the estimate corresponding to highest log-likelihood at convergence of the algorithm
- We select the *optimal number of latent classes* corresponding to the minimum of $BIC_k(\mathcal{I})$, that is, $k = 8$
- We order the items on the basis of the variability of their estimated conditional response probabilities across the classes, so as to select the initial set of clustering items

- In order to study the *sensitivity* of the final solution with respect to the initial set of clustering items, we consider different sizes of $\mathcal{A}^{(0)}$ equal to 3,10,20,30,75
- For each initial set of items, we select the initial number of latent classes,$k^{(0)}$, on the basis of $BIC_k(\mathcal{A}^{(0)})$, for $k = 2, \dots, k_{\max}$, and we start the item selection procedure

Comparison of the results of the inclusion-exclusion algorithm for item selection with respect to different sizes of the initial set of clustering items (in boldface are the quantities corresponding to the best solution in terms of $BIC_{tot,k}(\hat{\mathcal{A}})$)

size of $\mathcal{A}^{(0)}$	$k^{(0)}$	# items	$\hat{k}$	$BIC_k(\hat{\mathcal{A}})$	$BIC_{tot,k}(\hat{\mathcal{A}})$
3	2	53	8	129,344.90	165,353.63
10	10	50	9	124,815.80	165,320.90
20	10	50	9	124,816.50	165,321.50
30	**9**	**51**	**9**	**127,164.00**	**165,317.20**
75	8	53	8	129,351.60	165,360.00

Validation by resampling

- Given the nature of the search algorithm and the complexity of the study, the selected number of latent classes and the final set of selected items may be sensitive to the specific data used for the analysis
- We *validate the results* by sampling with replacement, from the original dataset, $B = 10$ samples of the same size, n, of the original one
- For each sample we select the optimal number of latent classes, k_b, corresponding to the minimum of $BIC_{k_b}(\mathcal{I})$, $b = 1, \dots, B$
- We then order the full set of items on the basis of the variability of their estimated conditional response probabilities across the classes
- We select the initial set of clustering items $\mathcal{A}_b^{(0)}$, $b = 1, \dots, B$, and we apply the item selection procedure
- For each sample we consider a size of the initial set of items equal to 30

Final results of the item selection strategy

j	#sel.	best	#resamp.	j	#sel.	best	#resamp.	j	#sel.	best	#resamp.
1	5	X	10	26	3	X	5	51	5	X	10
2	5	X	10	27			2	52	5	X	10
3	5	X	10	28				53	5	X	10
4	5	X	10	29			2	54	5	X	10
5	5	X	10	30				55	5	X	10
6	5	X	10	31				56	5	X	10
7	5	X	10	32	5	X	10	57	5	X	10
8	5	X	10	33	5	X	10	58	5	X	10
9	5	X	10	34	5	X	10	59	5	X	10
10	5	X	10	35	5	X	10	60	5	X	7
11	5	X	10	36				61	5	X	10
12	5	X	10	37				62	5	X	10
13	5	X	10	38	5	X	10	63			
14	5	X	10	39				64	2		2
15	5	X	10	40	5	X	10	65			
16	5	X	10	41	5	X	10	66	2		4
17	5	X	10	42	5	X	10	67			1
18	5	X	10	43	5	X	10	68	5	X	10
19				44	5	X	10	69	5	X	7
20	5	X	9	45	5	X	10	70			
21			1	46	5	X	10	71			
22				47	5	X	10	72			
23			1	48	5	X	10	73			
24			3	49	5	X	10	74			
25				50	5	X	10	75	5	X	10

j: item index - #sel.: number of times that item j has been selected with respect to the different starting sets - best: item included in the best solution -

#resamp.: number of times that item j has been selected with respect to the different samples

Comments

- 47 items are *always included* in the different final solutions, both with respect to different specifications of the initial set of items and with respect to the different samples
- 16 items are *never included* in the final solutions
- 12 items are in intermediate situations
- All items referred to sections CC, AVF, and ADL are always retained in the final solutions
- Most of the excluded items belongs to sections SC, DD, HBD and NF

Parameters estimates

- We consider the best solution, in terms of $BIC_{tot,k}(\hat{\mathcal{A}})$, provided by the inclusion-exclusion algorithm, which selects 51 items with $\hat{k} = 9$
- Estimation results are reported in terms of *item mean score*

$$\hat{\mu}_{ju} = \frac{1}{l_j - 1} \sum_y (y-1)\hat{\lambda}_{j|u}(y), \quad j \in \hat{\mathcal{A}},\ u = 1, \dots, \hat{k}$$

  - A value of $\hat{\mu}_{ju}$ close to 0 corresponds to a low probability of suffering from a certain pathology
  - A value close to 1 corresponds to a high probability of suffering from the same pathology

- We compute the *section mean score* $\hat{\bar{\mu}}_{d|u}$ as the average of $\hat{\mu}_{ju}$ for the items in $\hat{\mathcal{A}}$ composing each section d of the questionnaire
- In order to have a clearer interpretation of the results, we order the latent classes on the basis of the values of $\hat{\bar{\mu}}_{d|u}$ assumed in section ADL

Estimated section mean score, $\hat{\bar{\mu}}_{d|u}$, for each latent class u and each section d of the questionnaire, together with the estimated weights $\hat{\pi}_u$ and the difference between the largest and the smallest estimated section mean score for each section

	d								
u	1 (CC)	2 (AVF)	3 (HBD)	4 (ADL)	5 (I)	6 (NF)	7 (DD)	8 (SC)	$\hat{\pi}_u$
1	0.043	0.098	0.074	0.090	0.231	0.082	0.393	0.048	0.171
2	0.378	0.236	0.213	0.132	0.350	0.082	0.387	0.028	0.108
3	0.699	0.457	0.419	0.244	0.711	0.168	0.406	0.057	0.078
4	0.150	0.178	0.112	0.317	0.411	0.148	0.372	0.104	0.098
5	0.600	0.369	0.278	0.529	0.785	0.222	0.340	0.065	0.095
6	0.080	0.142	0.098	0.629	0.608	0.126	0.406	0.147	0.103
7	0.757	0.602	0.313	0.682	0.900	0.343	0.372	0.183	0.103
8	0.486	0.326	0.193	0.732	0.839	0.234	0.399	0.191	0.131
9	0.733	0.662	0.227	0.903	0.886	0.437	0.379	0.313	0.112
$\max_u(\hat{\bar{\mu}}_{d|u})-$ $\min_u(\hat{\bar{\mu}}_{d|u})$	0.715	0.563	0.345	0.813	0.669	0.356	0.066	0.285	

Comments

- The sections which present a high difference between the maximum and the minimum value of $\hat{\hat{\mu}}_{d|u}$ are ADL, CC and I
- The smallest among these differences is observed for section DD, which tend to discriminate less between subjects
- The first latent class, which includes around 17% of patients, corresponds to the *best health conditions* for all the pathologies measured by the sections, apart from DD and SC
- The 9th latent class, which includes around 11% of patients, corresponds to cases with the *worst health conditions* for almost all the pathologies
- Intermediated classes show a different case-mix depending on the section mean score pattern

Conclusions

- The algorithm for item selection we illustrate may lead to a *sensible reduction* of the number of items, without losing relevant information for clustering
- This implies clear advantages in terms of setting up a questionnaire which may be more easily administered, especially in a longitudinal context in which the questionnaire is periodically administered
- Reducing the dimension of the dataset also implies that it may be more easily analyzed by complex statistical models
- The clustering scheme obtained by the estimation of the LC model may be useful for *evaluating the ability of the nursing homes* to retain patients in the classes corresponding to better health conditions

Main references

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41** 164–171.

Dean, N. and Raftery, A. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, **62**, 11–35.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statitical Association*, **90**, 773–795.

Lattanzio, F., Mussi, C., Scafato, E., Ruggiero, C., Dell'Aquila, G., Pedone, C., Mammarella, F., Galluzzo, L., Salvioli, G., Senin, U., Carbonin, P. U., Bernabei, R., and Cherubini, A. (2010). Health care for older people in Italy: The U.L.I.S.S.E. project (un link informatico sui servizi sanitari esistenti per l'anziano - a computerized network on health care services for older people). *J Nutr Health Aging*, **14**, 238–42.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. S., editor, *Measurement and Prediction*, New York. Princeton University Press.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. 2nd ed. Wiley, New York.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, **63**, 581–592.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.