

Model-based classification of clustered binary data with non ignorable missing values

Francesco Lagona

Abstract A latent-class model is proposed for the classification of clustered binary data with non-ignorable missing values, when missingness is informative of class membership. The joint distribution of the complete data and the missing pattern is specified on the basis of a discrete, bivariate random effect that simultaneously allows for latent heterogeneity between subjects and a non ignorable missing data mechanism. Maximum likelihood estimates are obtained by a suitable E-M algorithm that reduces to the iterative maximization of three log-likelihood functions with case weights, alternated with weights updating. This methodology is exploited on a sample of Chinese older adults, to cluster subjects according to their cognitive impairment and ability to cope with a Mini-Mental State Examination questionnaire.

Key words: E-M algorithm, latent variable models, non-ignorable missing data, random effects

1 Introduction

In model-based classification studies, multivariate observations are clustered according to a finite number of classes in a way that the association structure between the observed variables is well approximated by this partitioning of the sample. A popular approach to model-based classification relies on the specification of the joint distribution of the variables of interest as a mixture of distributions, which represent the distribution of the variables within each class.

The focus of this paper is on model-based classification of clustered (or repeated) binary data, which arise when a battery of binary variables (or items) is observed on a sample of subjects, together with a number of subject- or item-specific covariates.

Francesco Lagona
DIPES - University Roma Tre, via G. Chiabrera 199, 00145 Rome (Italy)
e-mail: lagona@uniroma3.it

These data are typically examined by generalized linear mixed models (GLMM; [8]), specified on the basis of continuous (often normally distributed) random effects, in order to simultaneously account for the unobserved heterogeneity between subjects and the dependence between the outcomes that have been observed on the same subject. Discrete random effects, taking a finite number of values, are however more appropriate when the goal of an analysis is the classification of subjects according to a finite number of groups [1]. In this paper, the focus is on finite mixtures of logistic regressions (also known as latent-class models), that are the simplest and most popular discrete-random-effects models for clustered binary data. Under this modelling strategy, the binary outcomes on the same subject are assumed to be conditionally independent given the value of a univariate, discrete random effect.

In the case of complete data information, maximum-likelihood estimates (MLEs) of a latent-class model can be obtained by exploiting standard E-M algorithms for finite mixture models [9]. When observations are incomplete, i.e. some of the binary outcomes on a subject are missing, the process that generates the missing values should be taken in account, in order to obtain efficient MLEs [10]. If the data are ignorably missing, i.e. the probability of not observing a value does not depend on the unobserved value, then efficient MLEs can be found by maximizing the marginal likelihood, obtained by integrating the likelihood function of the complete data with respect to the missing values. If, otherwise, the data are non-ignorably missing, then a missing value is informative of the unobserved value and a more complex likelihood function must be maximized, obtained by specifying the joint distribution of the complete data and the missing pattern for each subject.

In this paper, missingness is assumed to be informative of class membership and a latent-class model that depends on a bivariate, discrete random effect is proposed for data classification. A suitable E-M algorithm is exploited for maximum-likelihood estimation, by adapting the algorithm developed by Hunt and Jorgensen [6] to handle ignorable missing values, when missingness is informative of latent-class membership. The model is exploited on a sample of Chinese older adults, obtained from the Chinese Longitudinal Health and Longevity Survey (CLHLS; [11]), to cluster subjects according to their cognitive impairment and ability to cope with a Mini-Mental State Examination questionnaire.

2 A latent class model for incomplete binary data

Clustered binary data arise when a battery of J binary variables (or items) is observed in a sample of n subjects and a binary vector $\mathbf{y}_i = (y_{i1} \dots y_{ij} \dots y_{iJ})$ is obtained for each subject (or cluster) $i = 1 \dots n$. In classification studies, these data are often examined in conjunction with a number of covariates, which are available at the item level and/or at the subject level. As a result, each subject is associated to the J rows $\mathbf{x}_{ij}^\top$ of a $J \times K$ design matrix $\mathbf{X}_i$.

To allow for missing outcomes on each subject i , we introduce a binary vector $\mathbf{r}_i = (r_{i1} \dots r_{ij} \dots r_{iJ})$, where $r_{ij} = 1$ if y_{ij} is missing and 0 otherwise. When the

data are non-ignorably missing, missingness of a value is informative of the unobserved value and, as a result, missing patterns $\mathbf{r}_i$ and complete data vectors $\mathbf{y}_i$ should be jointly modelled. Two alternative ways of specifying $p(\mathbf{y}, \mathbf{r})$ are based on a selection-model specification, say $p(\mathbf{y}, \mathbf{r}) = p(\mathbf{r}|\mathbf{y})p(\mathbf{y})$, or on a pattern-mixture specification $p(\mathbf{y}, \mathbf{r}) = p(\mathbf{y}|\mathbf{r})p(\mathbf{r})$. Aiming at clustering subjects in a number of classes, we may alternatively assume that missingness is informative of class membership. More precisely, we may introduce a discrete, bivariate random variable, say (U, V) , that takes $H \times M$ values (u_h, v_m) with probabilities $\pi_{hm}, h = 1 \dots H, m = 1 \dots M, \sum_{hm} \pi_{hm} = 1$, and assume that

$$p(\mathbf{y}_i, \mathbf{r}_i) = \sum_{h=1}^H \sum_{m=1}^M \pi_{hm} p_h(\mathbf{y}_i | \mathbf{X}_i) p_m(\mathbf{r}_i | \mathbf{X}_i), \quad (1)$$

where

$$p_h(\mathbf{y}_i | \mathbf{X}_i) = \prod_{j=1}^J p_{hij}^{y_{ij}} (1 - p_{hij})^{1-y_{ij}} \quad (2)$$

$$p_m(\mathbf{r}_i | \mathbf{X}_i) = \prod_{j=1}^J p_{mij}^{r_{ij}} (1 - p_{mij})^{1-r_{ij}} \quad (3)$$

and

$$\text{logit } p_{hij} = \text{logit } P(y_{ij} = 1 | u_h, \mathbf{x}_{ij}^T) = u_h + \mathbf{x}_{ij}^T \beta_j \quad (4)$$

$$\text{logit } p_{mij} = \text{logit } P(y_{ij} = 1 | v_m, \mathbf{x}_{ij}^T) = v_m + \mathbf{x}_{ij}^T \gamma_j, \quad (5)$$

where β_j and γ_j are parameters vectors that separately capture the effect of the covariates on the binary outcomes and the missing patterns, respectively. Latent heterogeneity between subjects is captured by the two random intercepts U and V . According to model (1), subjects are clustered into $H \times M$ classes (u_h, v_m) , allowing for flexible modelling when subjects are more homogeneous with respect to the outcomes than to the missing patterns, or viceversa. Finally, the two random intercepts are correlated, allowing for a dependence structure between the outcomes and the missing patterns, within the same subject. In summary, the model provides a clustering of subjects in classes with homogeneous responses and missingness patterns, capturing the heterogeneity of the missing mechanism within classes of subjects who are similar with respect to the outcome, after controlling for relevant covariates.

3 Estimation

For each subject i , let $M(i)$ and $O(i)$ be the subsets of the missing and the observed items, respectively, $O(i) \cup M(i) = \{1, 2, \dots, J\}$. Maximum likelihood estimates of

the parameters $\theta = (\mathbf{u}, \mathbf{v}, \pi, \beta, \gamma)$ in model (1) can be obtained by maximizing the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \sum_{\mathbf{y}_{M(i)}} \sum_{h=1}^H \sum_{m=1}^M \pi_{hm} \prod_{j=1}^J p_h(y_{ij}) p_m(r_{ij}) \\ &= \prod_{i=1}^n \sum_{h=1}^H \sum_{m=1}^M \pi_{hm} \prod_{j \in O(i)} p_h(y_{ij}) \prod_{j=1}^J p_m(r_{ij}) \end{aligned} \quad (6)$$

where $\mathbf{y}_{M(i)}$ is the vector of the unobserved outcomes of subject i . Notice that if the two random effects are independent, then $L(\theta)$ is proportional to the quantity

$$\prod_{i=1}^n \sum_{h=1}^H \pi_{h\cdot} \prod_{j \in O(i)} p_h(y_{ij}), \quad (7)$$

where $\pi_{h\cdot} = \sum_{m=1}^M \pi_{hm}$, or, in other words, missing values are ignorable. In the general case of dependent random effects, likelihood (6) can be efficiently maximized by an E-M algorithm that reduces to the iterative maximization of the following expected complete log-likelihood function (Maximization step)

$$Q(\theta | \hat{\theta}) = \sum_{i=1}^n \sum_{h=1}^H \sum_{m=1}^M w_{ihm}(\hat{\theta}) \left(\frac{\log \pi_{hm}}{\sum_{j \in O(i)} \log p_h(y_{ij})} \right), \quad (8)$$

alternated with the iterative updating of the following weights (Expectation step)

$$w_{ihm}(\hat{\theta}) = \frac{\hat{\pi}_{hm} \prod_{j \in O(i)} p_h(y_{ij} | \hat{\beta}, \hat{u}_h) \prod_{j=1}^J p_m(r_{ij} | \hat{\gamma}, \hat{v}_m)}{\sum_{h=1}^H \sum_{m=1}^M \hat{\pi}_{hm} \prod_{j \in O(i)} p_h(y_{ij} | \hat{\beta}, \hat{u}_h) \prod_{j=1}^J p_m(r_{ij} | \hat{\gamma}, \hat{v}_m)}, \quad (9)$$

which are subject-specific posterior probabilities of class membership. These weights are computed on the basis of an estimate of θ , that is obtained from a previous maximization step. Function Q can be decomposed into three functions

$$Q(\theta | \hat{\theta}) = Q_1(\pi | \hat{\theta}) + Q_2(\beta, \mathbf{u} | \hat{\theta}) + Q_3(\gamma, \mathbf{v} | \hat{\theta}),$$

that can be maximized separately. In particular, the following estimates

$$\hat{\pi}_{hm} = \frac{\sum_{i=1}^n w_{ihm}(\hat{\theta})}{n}$$

maximize $Q_1(\pi | \hat{\theta})$, while

$$Q_2(\beta, \mathbf{u} | \hat{\theta}) = \sum_{i=1}^n \sum_{h=1}^H w_{ih}(\hat{\theta}) \sum_{j \in O(i)} \log p_h(y_{ij}) \quad \left(w_{ih\cdot} = \sum_{m=1}^M w_{ihm} \right) \quad (10)$$

$$Q_3(\gamma, \mathbf{v} | \hat{\theta}) = \sum_{i=1}^n \sum_{m=1}^M w_{i\cdot m}(\hat{\theta}) \sum_{j=1}^J \log p_m(r_{ij}) \quad \left(w_{i\cdot m} = \sum_{h=1}^H w_{ihm} \right) \quad (11)$$

can be separately maximized by fitting two weighted logistic regressions on two augmented datasets that include $H \times n$ and $M \times n$ pseudo-subjects, with case weights. The algorithm is iterated up to the convergence of the estimates, whose limit is a local maximum point of the likelihood (6)[14].

If the number of classes is large, stable standard errors of the parameters of interest can be computed ([3]) on the basis of the predictive distribution of the missing values, as obtained from the estimates of the last iteration of the E-M algorithm. Precisely, missing data are imputed by sampling values from

$$p(\mathbf{y}_{M(i)} | \mathbf{y}_{O(i)}, \mathbf{r}_i, \hat{\theta}) = \sum_{h=1}^H \sum_{m=1}^M w_{ihm}(\hat{\theta}) \prod_{j \in M(i)} p_h(y_{ij} | \hat{\beta}, \hat{\mathbf{u}}),$$

to obtain naive point estimates and variance estimates of the parameters. Finally, the variance of the EM estimator is obtained as a weighted sum of the mean of the imputation variances and the empirical variance of the imputation point estimates, with weights 1 and N , where N is the number of imputation used.

4 Application

4.1 Data

The data that motivated this article are drawn from the study No. 3891 of the Inter-University Consortium for Political and Social Research ([11]). The study was carried out on 7352 Chinese subjects, aged between 80 and 106, whose cognitive impairment was examined through the Chinese Mini-Mental State Examination (MMSE) questionnaire. Questions in a MMSE questionnaire are typically compound and include a number of single items to be separately asked to the subject. Scores on each item are binary (e.g., 1 for a correct answer and 0 otherwise). With respect to the popular 30-items MMSE [2], the 23-items Chinese MMSE adopts some appropriate adjustments to make the questions more understandable and answerable among ordinary oldest old Chinese, the majority of whom are illiterate [11]. Overall, respondents were asked a 5-items orientation-related question (naming the current time, animal year, season, festival, and county), a 12-items language-related question (6 items on word recalling, 3 items on word repetition and 3 items on sentence comprehension), a 5-items calculation question (respondents are asked to subtract 3 from 20, then 3 from the previous result, and so on) and a single-item

drawing question (drawing a figure that is shown to the respondent). Only 57% of the questionnaires were complete. Partial questionnaires with a number of missing items between 1 and 10 missing items were about the 30% of the sample. The rest of the sample (13%) left more than 10 items unanswered. Of these, 7% are subjects who did not answer any questions of the questionnaire.

An outcome often reported in the literature is the observation that missing scores on tests of cognitive impairment occur more frequently among cognitively impaired patients [13, 5]. Missing values cannot therefore be ignored in the analysis, as they are informative of the level of cognitive impairment in a subject.

For each subject, relevant covariates were obtained at the time of the MMSE interview: age (in months), gender, type of residence (rural or urban), whether the subject is sedentary or active, and limits in activities of daily living (ADL; six activities including bathing, dressing, eating, indoor transferring, toileting and continence), categorized into three levels: no, one, two or more limits. In this sample, the median age at the MMSE interview is 92 years, while the lower and the upper quartiles are respectively 91 and 100 years, 59% of the subjects are males, 64% are rural residents, 45% have a sedentary lifestyle, 14% have one limit in ADL and finally 20% have two or more ADL limits.

4.2 Results

Model (1) has been estimated from the CLHLS data by varying the number H of the cognitive-impairment latent classes and the number M of the missing-mechanism latent classes, to compute the BIC criterion, $\text{BIC}(H, M) = -2 \log L_{H, M}(\hat{\theta}) + q \log n$, where n is the number of subjects in the sample and q is the number of unknown parameters (Table 1).

Table 1 BIC values

number of cognitive impairment classes	number of missing mechanism classes				
	1	2	3	4	5
1	130765.3	128773.1	127007.7	120985.8	146882.5
2	125888.2	112995.5	111963.3	121873.0	126981.7
3	124872.4	114345.2	109554.5	116441.0	159332.1
4	143221.5	131665.3	121540.5	122937.6	112994.1
5	145761.0	187872.2	134871.1	144556.9	144662.5

According to the above criterion, a model with $H \times M = 3 \times 2$ latent classes seems a good compromise between goodness of fit and parsimony. Although preliminary simulation studies have shown a reasonable performance of BIC in choosing the dimension of model (1), the general behavior of BIC (and other alternative criteria such as AIC) in the case of clustered binary data with non ignorable missing values is still an area of open research.

Table 2 Random effects distribution

cognitive impairment classes	missing mechanism $v_1 = 0.83$	classes $v_2 = -0.83$	marginal
$u_1 = -0.150$	0.371	0.123	0.494
$u_2 = -0.100$	0.005	0.056	0.061
$u_3 = 0.246$	0.198	0.247	0.445
marginal	0.574	0.426	1

Table 3 Covariate effects on cognitive functioning and missing mechanism

covariate		mental state		missing mechanism	
		est	(st. err.)	est	(st. err.)
type of question					
(ref: orientation)	language	-0.4808	(0.1504)	0.1768	(0.1644)
	calculation	-0.8117	(0.1698)	0.4902	(0.1900)
	drawing	-2.3134	(0.2562)	1.8486	(0.3000)
age (months)		-0.0035	(0.0008)	0.0056	(0.0010)
gender					
(ref: female)	male	-0.4668	(0.1320)	0.3199	(0.1603)
residence					
(ref: urban)	rural	-0.1416	(0.1163)	0.9969	(0.1441)
lifestyle					
(ref: active)	sedentary	-1.0422	(0.1288)	2.2928	(0.1856)
ADL limits					
(ref: no limits)	one	-0.3328	(0.1970)	0.8910	(0.2107)
	two or more	-0.4491	(0.1260)	1.6816	(0.1420)

Table 2 displays the estimated bivariate distribution of the random effects that are associated to the 6 latent classes, as obtained after the last iteration of the algorithm. Classes proportions confirm the high level of latent heterogeneity in the sample. Subjects are essentially clustered into two main classes $(-0.150, 0.83)$ and $(0.246, -0.83)$. MMSE questionnaires in class $(-0.150, 0.83)$ include larger numbers of unanswered items and wrong answers than those predicted by the available covariates, while questionnaires in class $(0.246, -0.83)$ include smaller numbers of missing items and wrong answers than those predicted by the covariates. However, 30% of the questionnaires is clustered into two classes, $(0.246, 0.83)$ and $(-0.150, -0.83)$, where subjects answer items less (more) often and give right answers more (less) often than expected by their covariate profiles. This indicates that the missing mechanism is heterogeneous within classes of subjects with similar levels of cognitive functioning, even after controlling for relevant covariates. As a result, the presence (absence) of unanswered questions is not necessarily an indication of poor (good) cognitive functioning, because unpredicted differences in the ability to cope with the questionnaire might reflect the absence of important covariates, such as education level, which influence the probability of answering specific

questions. Columns 1 and 2 of Table (3) show the estimated effects of the covariates on cognitive functioning (model (4)) with their standard errors, conditionally on the random effects. Impairment in orientation occurs less often than impairment in calculation, language and drawing. Wrong answers occur especially in calculation- and drawing-related items of the questionnaire. As expected, older adults are more cognitively impaired than younger adults. Overall, males are less cognitively impaired than females, confirming the results on gender differentials found by Zhang [12] on the same data used in this paper. In keeping with the outcomes reported by Gu and Qui [4], urban residents are less cognitively impaired than the rural residents, whereas physical disabilities and a sedentary life style negatively influence a subjects cognitive functioning, even after correcting for age at the interview. Columns 3 and 4 of Table (3) display parameters estimates and standard errors of model (5). Factors that negatively influence cognitive functioning increase the probability of leaving an item unanswered, in keeping with the results reported by Lagona and Zhang ([7]). Interestingly, males answer the MMSE questionnaire items less often than females.

References

1. Bouguila N: On multivariate binary data clustering and feature weighting, *Computational Statistics and Data Analysis*, **54**, 120–134 (2010).
2. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatry Research* **12**, 189–198 (1975).
3. Goetghebuer E, Ryan L: Semiparametric regression analysis of interval-censored data, *Biometrics*, **56**, 1139–1144 (2000).
4. Gu D, Qiu L. Cognitive functioning and its determinants among the oldest-old in China. *Journal of Nanjing College for Population and Management*, **2**, 3–9 (2003).
5. Herzog AR, Wallace RB. Measures of cognitive functioning in the AHEAD Study. *Journals of Gerontology: Psychological Sciences and Social Sciences* **B52**, 37–48 (1997).
6. Hunt L, Jorgensen M: Mixture model clustering for mixed data with missing information, *Computational Statistics and Data Analysis*, **41**, 429–440 (2003).
7. Lagona F, Zhang Z: A missing composite covariate in survival analysis: a case study of the Chinese Longitudinal Health and Longevity Survey, *Statistics in Medicine*, **29**, 248–261, (2010).
8. Mc Cullagh CE: *Generalized Linear Mixed Models*, Institute of Mathematical Statistics, Beachwood (Ohio) (2003).
9. McLachlan GJ, Peel D: *Finite Mixture Models*. John Wiley, New York (2000).
10. Rubin DB: Inference and Missing Data, *Biometrika*, **63**, 81–92 (1976).
11. Zeng Y, Vaupel, JW: Functional Capacity and Self-Evaluation of Health and Life of Oldest Old in China, *Journal of Social Issues*, **58**, 733–748 (2002).
12. Zhang Z. Gender differentials in cognitive impairment and decline in the oldest old in China. *Journal of Gerontology: Social Sciences* **B61**, S107–S115 (2006).
13. Zimmer Z, Martin LG, Chang M-C. Changes in functional limitation and survival among older Taiwanese, 1993, 1996, and 1999. *Population Studies*, **56**, 265–276 (2002) .
14. Wu C: On the convergence properties of the EM algorithm, *The Annals of Statistics*, **11**, 95–103 (1983).