# Optimal model-based clustering with multilevel data

*Fulvia Pennoni*

*Department of Statistics and Quantitative Methods*

*University of Milano-Bicocca*

http://www.statistica.unimib.it/utenti/pennoni/

Email: fulvia.pennoni@unimib.it

*joint work with S. Bacci and F. Bartolucci*

# Outline

- Introduction
- Multilevel Latent Class (MLC) model with covariates
- Proposed predictive clustering method and cluster allocation
- Simulation results to evaluate the prediction accuracy
- Application to the Italian data from the TIMSS & PIRLS 2011 international survey
- References

# Introduction

- We deal with observed binary or polytomous item responses related to individuals when they are nested in groups
- We consider a finite-mixture latent variable model to account for a flexible way to find homogeneous latent classes of individuals and groups
- We address the problem of predicting the allocations of the latent variables at cluster and individual level on the basis of the observed data
- Among the proposed approaches the Maximum A-Posteriori (MAP) probability is commonly employed
- We propose an alternative rule for the posterior classification which allocates jointly individuals and groups

# General MLC model formulation and assumption

- $\boldsymbol{Y}_h = (\boldsymbol{Y}_{h1}, \dots, \boldsymbol{Y}_{hn_h})'$ vector of dichotomous or ordinal polytomous responses related to individual $i$, $i = 1, \dots, n_h$, and group $j$, $j = 1, \dots, r$ for $h = 1, \dots, H$ groups (or clusters)

- We consider a discrete latent variable $Z_{hi}$ to represent the bidimensional latent variables which identify homogenous clusters $U_h$ and classes $V_{hi}$

- We define two probabilities associated to $Z_{hi}$: the probability of each possible discrete state $z = 1, \dots, k_Z$ as

$$q_{h,z} = p(Z_{h1} = z) = \lambda_{h,u(z)}\lambda_{h1,v(z)|u(z)}$$

and the probabilities of changes across states as

$$q_{hi,z|\bar{z}} = p(Z_{hi} = z|Z_{hi-1} = \bar{z}) = \lambda_{hi,v(z)|u(z)}, \quad \text{if } u(z) = u(\bar{z})$$

where $\lambda$ denotes the conditional probability of belonging to the latent classes according to the cluster

# General MLC model formulation and assumption

- We assume that $\lambda_{hi,v}$ and $\lambda_{hi,v|u} = p(V_{hi} = v|U_h = u)$ are constant for all groups and individuals

- We also consider covariates collected at group $\boldsymbol{W}_h$ level as $\lambda_{h,u} = p(U_h = u|\boldsymbol{W}_h = \boldsymbol{w}_h)$ and at individual level $\boldsymbol{X}_{hi}$ as $\lambda_{hi,v|u} = p(V_{hi} = v|U_h = u, \boldsymbol{X}_{hi} = \boldsymbol{x}_{hi})$

- We employ a multinomial logit parameterization:

$$\log \frac{\lambda_{hi,v|u}}{\lambda_{hi,1|u}} = \boldsymbol{x}'_{hi}\psi_v^{(V)}, \quad v = 2, \dots, k_V,$$
$$\log \frac{\lambda_{h,u}}{\lambda_{h,1}} = \boldsymbol{w}'_h\psi_u^{(U)}, \quad u = 2, \dots, k_U,$$

$\psi_v^{(V)}$ and $\psi_u^{(U)}$ are vectors of regression coefficients on the logit of individual and group covariates ($v = 1$ and $u = 1$ are arbitrarily chosen as reference categories)

# Prediction methods: MAP

- The model is estimated by maximizing the complete log-likelihood function by the Expectation-Maximization algorithm
- The estimated posterior probabilities of belonging to a certain latent class for each individual and to certain latent cluster for each group are provided by suitable forward recursions
- MAP consists in selecting the classes having the highest posterior probability, which corresponds to the conditional distribution of this variable given the observed data and it has to be done for each latent variable (at each level)

# Prediction methods: MAP

- MAP is applied by either considering the marginal (denoted by suffix (1) in the following) or the conditional probabilities (denoted as (2))

- In (1) the maximization is performed separately for the posterior probabilities related to the latent variables at cluster and individual level

- In (2) instead we first predict the latent variable for each cluster and second we predict each individual-specific latent variable conditional on the value predicted for the corresponding cluster-level latent variable

# Prediction method: Viterbi

- We propose to jointly consider the allocation of individuals and groups by adapting a suitable version of the Viterbi algorithm (Viterbi, 1967, Juan and Rabiner, 1991)

- To decode the optimal sequence of clusters and classes at the same time the algorithm involves the following quantities:
  - $\phi_{i\boldsymbol{y}|z} = p(\boldsymbol{Y}_{hi} = \boldsymbol{y}_{hi} | U_h = u, V_{hi} = v)$: the conditional probability of the response vector of individual $i$ given the latent variables
  - $\hat{p}_1(z, \boldsymbol{y}_h)$: the estimated values of the conditional posterior probabilities of belonging to a certain latent class and to the group $h$ given the observed values
  - $\hat{p}_i(z, \boldsymbol{y}_h)$: the estimated conditional posterior probabilities as above for each individual $i = 2, \ldots, n_h$

# The proposed clustering algorithm

- The required steps are the following:
  - I Compute the joint posterior probabilities for 1

$$\hat{p}_1(z, \boldsymbol{y}_h) = \hat{\phi}_{1\boldsymbol{y}|z}\hat{q}_h(z);$$

for $z = 1, \ldots, k_Z$ and for every group $h = 1, \ldots, H$

  - II Compute the joint posterior probabilities for each $i$

$$\hat{p}_i(z, \boldsymbol{y}_h) = \hat{\phi}_{i\boldsymbol{y}|z} \max_{\bar{z}=1,\ldots,k_Z} [\hat{p}_{i-1}(\bar{z}, \boldsymbol{y}_h)\hat{q}_{hi}(z|\bar{z})]$$

for $i = 2, \ldots, n_h$ and $z = 1, \ldots, k_Z$

  - III Find the optimal state

$$\hat{\boldsymbol{z}}^*_{n_h}(\boldsymbol{y}_h) = argmax_{(z=1,\ldots,k_z)}\ \hat{p}_{n_h}(z, \boldsymbol{y}_h)$$

for $z = 1, \ldots, k_Z$ and $i = n_h$

  - IV Predict jointly clusters and classes allocation by considering

$$\hat{\boldsymbol{z}}^*_i(\boldsymbol{y}_h) = argmax_{(z=1,\ldots,k_z)}\ [\hat{q}_i(z, \boldsymbol{y}_h)\hat{q}_{hi+1}(\hat{\boldsymbol{z}}^*_{hi+1}, \boldsymbol{y}_h|z)],$$

for $z = 1, \ldots, k_Z$ and $i = n_h - 1, \ldots, 1$

# Entropy for LMC

- The entropy helps to assess the separability between the latent components (Lukočienė and Vermunt, 2009)
- The following three entropy measures account for the degree of separation between clusters and classes
  - For the clusters

$$\mathrm{EN}_U = -\sum_{h=1}^{H}\sum_{u=1}^{k_U} \hat{p}_{uh}(u|\boldsymbol{y}) \log \hat{p}_{uh}(u|\boldsymbol{y})/H$$

  - For the classes of individuals when the marginal (1) or the conditional (2) approach is used

$$\mathrm{EN}_{V_1} = -\sum_{i=1}^{n}\sum_{v=1}^{k_V} \hat{p}_{ih}(v|\boldsymbol{y}) \log \hat{p}_{ih}(v|\boldsymbol{y})/n_h$$

$$\mathrm{EN}_{V_2} = -\sum_{i=1}^{n_h}\sum_{v=1}^{k_V} \hat{p}_{ih}(v|u,\boldsymbol{y}) \log \hat{p}_{ih}(v|u,\boldsymbol{y})/n_h.$$

# Simulation design 1

- In order to establish a comparison between the proposed algorithm and the current one we planned two different simulation designs
- The first aims at evaluating the prediction accuracy when the groups don't exist in the model generating the data and the second when the data are generated by a precise hierarchal structure with discrete latent variables for groups and individuals
- The first simulation evaluates the accuracy of MAP and Viterbi when the model is misspecified
- The simulation design is the following:
  - $U \sim N(0, 1)$ (clusters) and $V \sim N(\mu_u, 1)$ (classes)
  - $H =$ (50,100) with $n_h = (10, 25, 50)$
  - $Y_{hij}$ binary variable with $r =$ (8,10) items

# Simulation design 1 (cont.)

- In the first simulation design we consider the following 72 situations with $k_V$ and $k_U$ ranging from 3 to 5

| # Scenario | $H$ | $n_h$ | $r$ | $k_U$ | $k_V$ | # Scenario | $H$ | $n_h$ | $r$ | $k_U$ | $k_V$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 10 | 8 | 3 | 3 | 37 | 100 | 10 | 8 | 3 | 3 |
| 2 | 50 | 10 | 8 | 3 | 4 | 38 | 100 | 10 | 8 | 3 | 4 |
| 3 | 50 | 10 | 8 | 3 | 5 | 39 | 100 | 10 | 8 | 3 | 5 |
| 4 | 50 | 10 | 8 | 4 | 4 | 40 | 100 | 10 | 8 | 4 | 4 |
| 5 | 50 | 10 | 8 | 4 | 5 | 41 | 100 | 10 | 8 | 4 | 5 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 32 | 50 | 50 | 12 | 3 | 4 | 68 | 100 | 50 | 12 | 3 | 4 |
| 33 | 50 | 50 | 12 | 3 | 5 | 69 | 100 | 50 | 12 | 3 | 5 |
| 34 | 50 | 50 | 12 | 4 | 4 | 70 | 100 | 50 | 12 | 4 | 4 |
| 35 | 50 | 50 | 12 | 4 | 5 | 71 | 100 | 50 | 12 | 4 | 5 |
| 36 | 50 | 50 | 12 | 5 | 5 | 72 | 100 | 50 | 12 | 5 | 5 |

# Simulation design 1 (cont.)

- For each scenario by considering the estimated entropy, we compare the rates of different allocation (DISagreement) within the MAP and the Viterbi for: groups and individuals when the marginal ($DIS_U$) and conditional MAP ($DIS_{V_1}$ and $DIS_{V_2}$) are evaluated
- We find that:
  - $DIS_U$ (latent clusters) rate shows a median of 0.245
  - $DIS_{V_1}$ (latent classes, marginal) rate shows a median of 0.130
  - $DIS_{V_2}$ (latent classes, conditional) rate shows a median of 0.055

| Scenario | $EN_U$ | $EN_{V_1}$ | $EN_{V_2}$ | $DIS_U$ | $DIS_{V_1}$ | $DIS_{V_2}$ |
|---|---|---|---|---|---|---|
| 1 | 0.707 | 0.425 | 0.384 | 0.086 | 0.056 | 0.019 |
| 2 | 0.495 | 0.495 | 0.399 | 0.106 | 0.122 | 0.034 |
| 3 | 0.390 | 0.488 | 0.383 | 0.085 | 0.133 | 0.025 |
| 4 | 0.729 | 0.488 | 0.395 | 0.138 | 0.139 | 0.036 |
| 5 | 0.598 | 0.472 | 0.365 | 0.139 | 0.131 | 0.027 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 68 | 0.490 | 0.576 | 0.568 | 0.302 | 0.064 | 0.052 |
| 69 | 0.353 | 0.654 | 0.621 | 0.422 | 0.152 | 0.135 |
| 70 | 0.788 | 0.577 | 0.566 | 0.330 | 0.072 | 0.057 |
| 71 | 0.499 | 0.652 | 0.618 | 0.520 | 0.156 | 0.138 |
| 72 | 0.708 | 0.642 | 0.603 | 0.550 | 0.156 | 0.135 |

# Simulation design 2

- The second simulation study is carried out to evaluate the prediction accuracy according to the correct allocation rates

- The simulation design is as follows:
  - $U$ and $V$ are discrete latent variables with $k_U = 3$ (clusters) and $k_V = 3$ (classes)
  - Sampling weights of $U$ correspond to the weights of a the Gaussian quadrature nodes and those of $V$ are related to the components of $U$ through an inverse logit transformation
  - To create more/less distinctive clusters and classes, as proposed in Yu and Park (2014) we consider the Gaussian quadrature nodes with zero mean and with the following choices for $\sigma_U$ and $\sigma_V$:

| | | |
|---|---|---|
| low | $\sigma_U = 0.2$ | $\sigma_V = 0.2$ |
| intermediate | $\sigma_U = 1.0$ | $\sigma_V = 1.0$ |
| high | $\sigma_U = 2.0$ | $\sigma_V = 2.0$ |
| mixed 1 | $\sigma_U = 0.2$ | $\sigma_V = 1.0$ |
| mixed 2 | $\sigma_U = 1.0$ | $\sigma_V = 0.2$ |

# Simulation design 2 (cont.)

- The second simulation design is made of $r = 8$ binary items so that we evaluate the proposal with 20 scenarios and 50 different datasets
- Where the number of groups is $H = 50{,}100$ and number of individuals in each group is $n_h = 10, 50$

| # Scenario | $H$ | $n_h$ | $\sigma_U$ | $\sigma_V$ |
|---|---|---|---|---|
| 1 | 50 | 10 | 0.2 | 0.2 |
| 2 | 50 | 10 | 0.2 | 1.0 |
| 3 | 50 | 10 | 1.0 | 0.2 |
| 4 | 50 | 10 | 1.0 | 1.0 |
| **5** | 50 | 50 | 0.2 | 0.2 |
| 6 | 50 | 50 | 0.2 | 1.0 |
| **7** | 50 | 50 | 1.0 | 0.2 |
| 8 | 50 | 50 | 1.0 | 1.0 |
| 9 | 100 | 10 | 0.2 | 0.2 |
| 10 | 100 | 10 | 0.2 | 1.0 |
| 11 | 100 | 10 | 1.0 | 0.2 |
| 12 | 100 | 10 | 1.0 | 1.0 |
| **13** | 100 | 50 | 0.2 | 0.2 |
| 14 | 100 | 50 | 0.2 | 1.0 |
| **15** | 100 | 50 | 1.0 | 0.2 |
| 16 | 100 | 50 | 1.0 | 1.0 |
| 17 | 50 | 10 | 2.0 | 2.0 |
| 18 | 50 | 50 | 2.0 | 2.0 |
| 19 | 100 | 10 | 2.0 | 2.0 |
| 20 | 100 | 50 | 2.0 | 2.0 |

# Simulation design 2 (cont.) results

- The values of the index of DISagreement are lower than those obtained within the first simulation study conducted with the misspecified model

- Related to the group allocations the median rate of disagreement $DIS_U$ is equal to 0.089

- Related to the cluster allocations $DIS_{V_1}$ has a median value of 0.076 and $DIS_{V_2}$ of 0.010

- The worst cases are related to scenarios (5,7,13,15) - see bold values in the table, where $DIS_U$ may reach 0.58 and $DIS_{V_1}$ and $DIS_{V_2}$ may reach 0.40

- The higher rate is reached when there are many individuals in each group and the latent groups are less distinctive

# Application to TIMM&PIRLS Italian data

- We use the large-scale surveys TIMMS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) conducted in 2011
- We consider the achievement scores at the fourth grade when the Italian pupils are 9 to 10 years old
- We consider 5 ordinal items according to the 5 plausible values for each subject
- The ordered categories are: 0: student performed below the low international benchmark (IB), 1: student performed at or above the low IB, but below the intermediate IB, 2: student performed at or above the intermediate IB, but below the high IB, 3: student performed at or above the intermediate IB, but below the high international benchmark, 4: student performed at or above the advanced IB

# Application to TIMM&PIRLS Italian data

- The covariates are collected from the background parents' questionnaires, the principals' questionnaire of the schools and from external data archives

- The covariates (Grilli et al., 2016) are:
  - gender
  - home resources for learning (scores)
  - early literacy/numeracy tasks
  - Italian language spoken at home (1 if yes)
  - school adequate environment and resources
  - school is safe and orderly
  - socio-economic condition of the area where the school is located (gross value added at province level, from an external data archive)
  - dummy variables for five Italian geographical areas (North-West, North-East, Centre, South, South-Islands)

# Application (cont.) results

- We evaluate the model and the proposed prediction algorithm for four different data structures and models: *i*) unidimensional responses (Maths) without covariates; *ii*) as *i*) with covariates; *iii*) multidimensional ordinal responses (Maths, Reading and Science) without covariates; *iv*) as *iii*) with covariates

- We estimated each model and we applied our proposal to classify students and schools for increasing values of $k_U$ and of $k_V$ ranging from 4 to 6 and from 3 to 6, respectively

- We evaluate the corresponding allocation of students (latent clusters) to the estimated latent classes (of schools) according with the estimated entropy values

- We report the absolute frequencies and the relative rates of schools and students differently classified between the MAP and the proposed approach (compared to the marginal (1) and conditional (2) MAP)

# Application (cont.) results

- The entropy measures suggest that some latent classes and, less, latent clusters are very well separated
- For the univariate response the disagreement rate related to the school's allocations, between MAP and Viterbi lies in the interval from 0.005 to 0.050
- This corresponds to a number of schools that are differently classified from 2 to 14 with an average equal to 5 (2.5%)
- The above rates are lower when multivariate responses are considered and a multidimensional model is estimated: the number of schools that are differently classified ranges from 2 to 8
- The above rates are not particularly influenced by the presence/absence of covariates

# Application (cont.) results 1 (cont.)

- We report the results for the highest and lowest value of $k_U$ and $k_V$ for each case *i), ii), iii)* and *iv)*
- $\#U, \#V_1, \#V_2$ are the number of schools and students which are differently classified with (1) and (2) respectively

| $k_U$ | $k_V$ | # U | # $V_1$ | # $V_2$ | $DIS_U$ | $DIS_{V_1}$ | $DIS_{V_2}$ |
|---|---|---|---|---|---|---|---|
| ***Maths**; no covariates* | | | | | | | |
| 4 | 3 | 3 | 90 | 5 | 0.015 | 0.024 | 0.001 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 5 | 6 | 14 | 107 | 9 | 0.070 | 0.029 | 0.002 |
| ***Maths**; with covariates* | | | | | | | |
| 4 | 3 | 4 | 80 | 6 | 0.020 | 0.021 | 0.002 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 5 | 6 | 7 | 149 | 11 | 0.035 | 0.040 | 0.003 |
| ***Maths, Reading, Grammar**; no covariates* | | | | | | | |
| 3 | 2 | 3 | 23 | 1 | 0.015 | 0.006 | 0.000 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 6 | 5 | 8 | 50 | 2 | 0.040 | 0.013 | 0.001 |
| ***Maths, Reading, Grammar**; with covariates* | | | | | | | |
| 3 | 2 | 1 | 22 | 0 | 0.005 | 0.006 | 0.000 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 6 | 5 | 5 | 32 | 5 | 0.025 | 0.009 | 0.001 |

# Application (cont.) results

- For the student's allocations the marginal MAP (1) performs strongly worse than conditional the MAP (2), when compared with the new proposal
- The disagreement rates under the marginal MAP (1) range from a minimum of 35 students to a maximum of 149 students when a univariate response is considered
- The above values are lower for the multidimensional responses: the number of students that are allocated differently ranges from 10 to 77 when compared with the marginal MAP (1)
- Under the conditional MAP (2) the differences are substantially reduced, ranging from 0 to 11 students for the univariate model and from 0 to 3 students for the multidimensional model

# Main References


- Bacci, S. and Gnaldi, M. (2015). A classification of university courses based on students satisfaction: an application of a two-level mixture item response model. Advances in Data Analysis and Classification, **49**, 927–940.
- Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press, Boca Raton.
- Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, **36** 491–522.
- Bray, B. C., Lanza, S. T., and Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural equation modeling: a multidisciplinary journal*, **22**, 1–11.
- Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.
- Dias, J. G. and Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, **23**, 643–659.
- Foy, P., Arora, A., and Stanco, G. M. (2013). *TIMSS 2011 User Guide for the International Database*. ERIC.
- Goodman, L. A. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, **37**,1–22.

# Main References


- Gnaldi, M., Bacci, S., and Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, **10**, 53–70.
- Grilli, L., Pennoni, F., Rampichini, C., and Romeo, I. (2016). Exploiting TIMMS&PIRLS combined data: multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*, **10**, 2405–2426.
- Juang, B. and Rabiner, L. (1991). Hidden Markov models for speach recognition *Technometrics*, **33**, 251–272.
- Lukočienė, O. and Vermunt, J. K. (2009) Determining the number of components in mixture models for hierarchical data. In: Fink, A., Lausen, B., Seidel, W. and Ultsch, A. (eds.), *Advances in data analysis, data handling and business intelligence*, 241-249. Springer: Berlin-Heidelberg.
- Pennoni, F. (2014). *Issues on the Estimation of Latent Variable and Latent Class Models*. Scholars' Press, Saarbücken.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum de- coding algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.
- Yu, H.-T. and Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate Behavioral Research*, **49**, 232–244.
- Welch, L. R. (2003). Hidden Markov Models and the Baum-Welch Algorithm, *IEEE Information Theory Society Newsletter*, **53**, 1-13.