Dynamic Classification through Three-Step Estimation: Evidence from a Multinational Longitudinal Study of Myasthenia Gravis Patients

Luca Brusa[1], Fulvia Pennoni[1], Francesco Bartolucci[2], and Lorenzo Maggi[3]

[1] University of Milano-Bicocca, Milan 20126, Italy
[2] University of Perugia, Perugia 06123, Italy
[3] Foundation IRCCS Neurological Institute Carlo Besta, Milan 20133, Italy

Abstract. We present a three-step likelihood inference approach for the analysis of multidimensional longitudinal data when a large number of response variables are collected at each time occasion. We show the results obtained using this method to identify patients with different levels of Myasthenia Gravis severity and assess symptom progression over time.

Keywords: Hidden Markov models, Likelihood inference, Multidimensional latent class item response models, Observational study

This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this contribution is published in *Statistics for Innovation II, SIS 2025, Short Papers, Contributed Sessions 1*, and is available online at https://doi.org/10.1007/978-3-031-96303-2.

1 Introduction

Stepwise approaches in the context of discrete latent variable models have been proposed to overcome the challenges of full maximum likelihood estimation, which may arise due to the high dimensionality of the data and the need to avoid confounding [4]. Considering observations repeated over time, coming from the administration of a wide range of items, and following the three-step approach proposed in [2], we first estimate a multidimensional item response theory model [3]. This step is followed by the assignment of individuals to the predicted latent classes and, as a third step, by the weighted maximum likelihood estimation of the structural component of a hidden Markov (HM) model [1] with covariates, based on weights accounting for the uncertainty of classification.

The proposed approach is illustrated through the analysis of responses to items repeatedly administered over time to a group of adult patients with Myasthenia Gravis (MG), a rare chronic autoimmune disease affecting the neuromuscular junction [6]. The data were collected as a part of an international longitudinal study conducted from 2020 to 2024, using a battery of 84 items designed to map eight different symptom dimensions. Through the proposed approach, we aim to identify groups of MG patients with different symptoms and varying levels of disease severity, and to assess whether clinical and demographic characteristics are associated with symptom burden at baseline and follow-up.

Section 2 presents the approach, while Section 3 illustrates the application and results. Section 4 provides concluding remarks.

2 Three-step approach

For a sample of n individuals observed at T time occasions, let $Y_{ij}^{(t)}$ denote response variable resulting from the administration of item j, where $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $j = 1, \ldots, r$. In particular, each variable $Y_{ij}^{(t)}$ is categorical with c_j categories, labeled from 0 to $c_j - 1$. These responses are collected in the vectors $\boldsymbol{Y}_i^{(t)} = (Y_{i1}^{(t)}, \ldots, Y_{ir}^{(t)})'$ for every individual i at occasion t, while $\boldsymbol{X}_i^{(t)}$ denotes the corresponding vector of covariates.

In this section, we present each of the three steps of the proposed approach for the analysis of data consisting of observations of the variables described above. For this aim we initially treat the responses provided by the same subject at different time occasions as if they belong to distinct subjects. We then denote by M the total number of observations across all subjects and time points and then the response vectors are alternatively denoted as $\boldsymbol{Y}_h = (Y_{h1}, \ldots, Y_{hr})'$, $h = 1, \ldots, M$, where h is defined on the basis of a one-to-one mapping of (i, t).

2.1 Multidimensional latent class graded response model

In the presence of multiple dimensions, let $\boldsymbol{U}_h = (U_{h1}, \ldots, U_{hD})'$ denote an observation-specific latent vector, with D representing the number of dimensions.

Each random vector $\boldsymbol{U}_h$ follows a discrete distribution with k support points denoted as $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$.

Under the latent class graded response (LCGR) model [7], the conditional response probabilities $\phi_{jy|u} = p(Y_{hj} = y|\boldsymbol{U}_h = \boldsymbol{\xi}_u)$ are parameterized as

$$\log \frac{p(Y_{hj} \geq y|\boldsymbol{U}_h = \boldsymbol{\xi}_u)}{p(Y_{hj} < y|\boldsymbol{U}_h = \boldsymbol{\xi}_u)} = \log \frac{\phi_{jy|u} + \ldots + \phi_{j,c_j-1|u}}{\phi_{j0|u} + \ldots + \phi_{j,y-1|u}} = \sum_{d=1}^{D} \delta_{jd}\xi_{ud} - \beta_{jy}, \quad (1)$$

where ξ_{ud} is the support point (or ability level) for latent class $u = 1, \ldots, k$ and dimension $d = 1, \ldots, D$, δ_{jd} is an indicator variable equal to 1 if item j measures latent trait d, and β_{jy} is the difficulty parameter for category y of item j.

The log-likelihood function of the model can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{M} \log p(\boldsymbol{Y}_h = \boldsymbol{y}_h) = \sum_{h=1}^{M} \log \left[\sum_{u=1}^{k} p(\boldsymbol{Y}_h = \boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{\xi}_u)\pi_u \right],$$

where $\boldsymbol{\theta}$ is the vector of all model parameters, $\boldsymbol{y}_h$ is the observed value of $\boldsymbol{Y}_h$, and $\pi_u = p(\boldsymbol{U}_h = \boldsymbol{\xi}_u)$ is the weight of class u. Maximum likelihood estimation is performed by the expectation-maximization algorithm [5], which iteratively maximizes $\ell(\boldsymbol{\theta})$ based on the complete-data log-likelihood function defined as

$$\ell^*(\boldsymbol{\theta}) = \sum_{h=1}^{M} \sum_{u=1}^{k} z_{hu} \log \left[p(\boldsymbol{Y}_h = \boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{\xi}_u)\pi_u \right],$$

where z_{hu} is an indicator variable equal to 1 if observation h comes from latent class u and 0 otherwise.

2.2 Model assignment of respondents to predicted latent classes

Based on the model estimates, the posterior distribution of the latent variables given the responses is computed, for each observation $h = 1, \ldots, M$, as follows:

$$\hat{q}_{hu} = p(\boldsymbol{U}_h = \boldsymbol{\xi}_u|\boldsymbol{Y}_h = \boldsymbol{y}_h) = \frac{p(\boldsymbol{Y}_h = \boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{\xi}_u)\pi_u}{p(\boldsymbol{Y}_h = \boldsymbol{y}_h)}. \quad (2)$$

Due to the longitudinal structure of the data, we consider now, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, the posterior probabilities $\hat{q}_{iu}^{(t)}$, computed on the basis of Equation (2), and $\hat{q}_{iuv}^{(t)} = \hat{q}_{iu}^{(t-1)}\hat{q}_{iv}^{(t)}$. Individuals are assigned to a certain latent class at each time t using the maximum-a-posteriori rule:

$$\hat{u}_i^{(t)} = \underset{u=1,\ldots,k}{\operatorname{argmax}}\ \hat{q}_{iu}^{(t)}, \qquad i = 1, \ldots, n,\ t = 1, \ldots, T.$$

2.3 Estimation of the structural model

We are interested in estimating the effects of the covariates on the initial and transition probabilities of $\boldsymbol{V}_i = (V_i^{(1)}, \ldots, V_i^{(T)})'$, where $V_i^{(t)}$ denotes the categorical response variable with k categories referred to the latent class of individual i at time occasion t. We then assume the multinomial logit models

$$\log \frac{p(V_i^{(1)} = u|\boldsymbol{x}_i^{(1)})}{p(V_i^{(1)} = 1|\boldsymbol{x}_i^{(1)})} = \beta_{0u} + (\boldsymbol{x}_i^{(1)})'\boldsymbol{\beta}_{1u}, \quad u = 2, \ldots, k,$$

$$\log \frac{p(V_i^{(t)} = v|V_i^{(t-1)} = u, \boldsymbol{x}_i^{(t)})}{p(V_i^{(t)} = u|V_i^{(t-1)} = u, \boldsymbol{x}_i^{(t)})} = \gamma_{0uv} + (\boldsymbol{x}_i^{(t)})'\boldsymbol{\gamma}_{1uv}, \quad u, v = 1, \ldots, k, \ u \neq v,$$

with $t = 2, \ldots, T$. Here $\boldsymbol{\beta} = (\beta_{0u}, \boldsymbol{\beta}'_{1u})'$ and $\boldsymbol{\gamma} = (\gamma_{0uv}, \boldsymbol{\gamma}'_{1uv})'$ are vectors of regression coefficients for the initial and transition probabilities, respectively. Maximum likelihood estimation is performed through weighted complete-data log-likelihood functions, with weights corresponding to the estimated posterior probabilities $\hat{q}_{iu}^{(t)}$ and $\hat{q}_{iuv}^{(t)}$ computed in the second step, thus accounting for the uncertainty of classification:

$$\ell_1^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{u=1}^{k} \hat{q}_{iu}^{(1)} \log p(V_i^{(1)} = u|\boldsymbol{x}_i^{(t)}),$$

$$\ell_2^*(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{u=1}^{k} \sum_{v=1}^{k} \hat{q}_{iuv}^{(t)} \log p(V_i^{(t)} = v|V_i^{(t-1)} = u, \boldsymbol{x}_i^{(t)}).$$

3 Application

In the following, we present the analysis of data collected from 1,702 patients affected by MG, a condition characterized by weakness in the voluntary muscles, with variable involvement of ocular, bulbar, respiratory, axial, and limb muscles. The eight dimensions of symptoms are listed in Table 1. Individual covariates are also collected at the baseline, including age, gender, housing situation, comorbidities and medical treatments.

In the first step of the analysis, to facilitate interpretability, the LCGR model is estimated fixing the number of latent classes at three. Considering the multidimensional structure of the data, models with different numbers of latent traits, obtained by iteratively aggregating items in different dimensions, are compared using the Bayesian information criterion. The best model identifies four dimensions corresponding to: (*i*) HADS, SLEEP, FACIT, HUI; (*ii*) MG-QOL, DYSPNEA; (*iii*) MG-ADL; (*iv*) EQ-5D-5L. Due to the large number of estimated parameters, the following summary measure for the estimated conditional probabilities in Equation (1) is considered: $\hat{\mu}_{j|u} = \frac{1}{c_j - 1} \sum_{y=0}^{c_j - 1} y \hat{\phi}_{jy|u}$; it ranges from 0 to 1 and measures the probability for every symptom in each latent group.

Figure 1 depicts the estimated average values for each dimension. We notice

Table 1. Description of the eight dimensions of the questionnaire and the corresponding number of items administered to individuals suffering of MG

Label	Description of the categories of symptoms
MG-ADL (8)	Symptoms severity of MG in activities of daily livings
HADS (14)	Levels of hospital anxiety and depression
EQ-5D-5L (11)	Generic quality of life
MG-QOL (15)	MG-related quality of life
SLEEP (6)	Overall sleep quality and disturbance
DYSPNEA (5)	Frequency, severity and duration of shortness of breath
FACIT (13)	Fatigue levels and impact days
HUI (12)	Health status and health-related quality of life

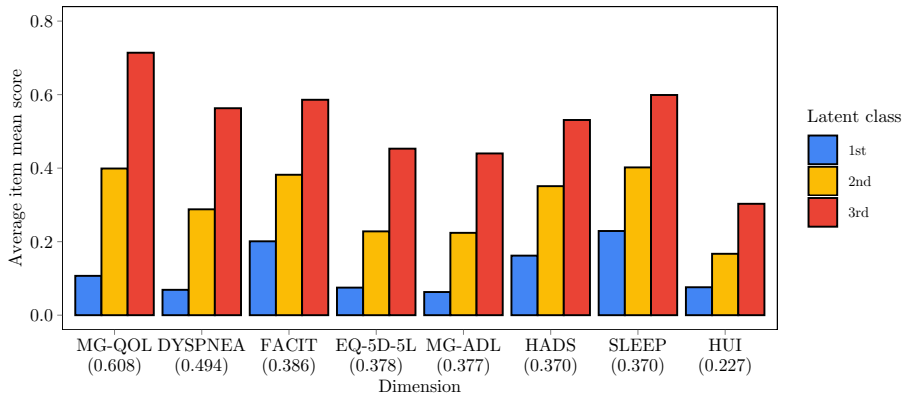

Fig. 1. Average value of the item mean scores $\hat{\mu}_{j|u}$ for each dimension; values in parentheses represent the difference between the averages in the third and in the first latent classes

that the first latent class includes patients with the mildest symptoms; this group is mainly characterized by items from dimension SLEEP, and comprises around 31.4% of patients across all time occasions. Conversely, the third latent class, which represents patients with the highest symptom burden, is primarily characterized by items in the MG-QOL dimension, and includes 27.8% of patients. Furthermore, the estimated mean differences between the first and third latent classes, shown in the figure, indicate that the typologies MG-QOL and DYSPNEA are those discriminating at most between latent classes.

In the second step, patients are classified into groups based on the model estimated posterior probabilities of class membership for each time occasion, as illustrated in Section 2.2. It is important to note that each patient may be assigned to different latent classes across the observed years.

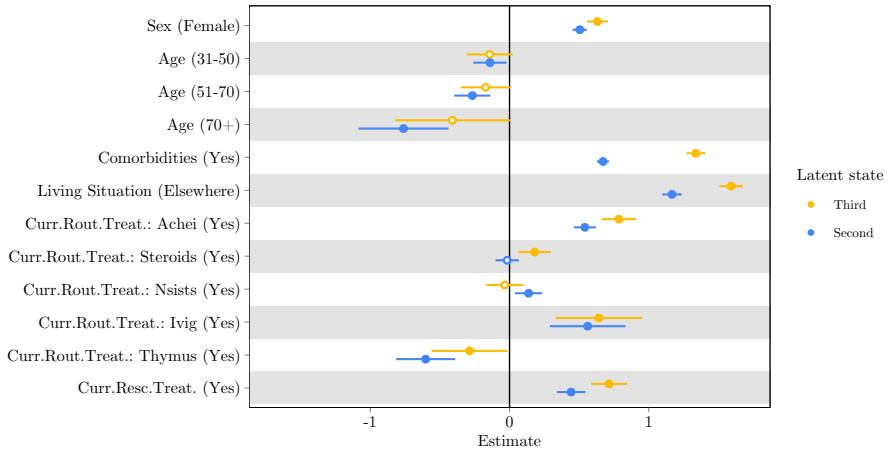

Fig. 2. Point estimates and related confidence intervals (95% level) for the regression parameters affecting the distribution of the initial probabilities of the second (blue) and third (yellow) latent states of the HM model with respect to the first; baseline categories for the following covariates are: Age (18-30), Living situation (Own home); abbreviations: Current Routine Treatments (Curr.Rout.Treat), Current Rescue Treatment (Curr.Resc.Treat)

In the third step, the initial and transition probabilities of the HM for class membership are estimated, accounting for the effects of covariates through the multinomial logit parameterizations presented in Section 2.3, with a weighted log-likelihood to account for uncertainty of classification. Results of the effects of the covariates on the initial probabilities are shown in the coefficient plot in Figure 2. In particular, we observe that women exhibit significantly more pronounced symptoms than men, with an odds ratio of 1.879 for being in the third latent class versus the first. Additionally, patients living in their own home show milder symptoms compared to those living elsewhere, with an odds ratio for the third class versus the first equal to 4.911. Finally, patients who have undergone treatment with an acetylcholinesterase inhibitor (Achei) are more likely to present a higher symptom burden, with an odds ratio of 2.194 for the third state relative to the first.

The estimated transition probabilities indicate that disease progression is generally limited, with stable conditions being more common for patients. For example, only around 28% of them show a reduction in symptoms during follow-up, as they transition from the third to the second latent state. Regarding the estimated effects of covariates on transition probabilities, it is particularly notable that steroid-based treatments lead to greater persistence in the same latent state compared to no treatment.

4 Conclusions

The proposed three-step procedure is suitable for analyzing large longitudinal multidimensional data as: (*i*) it reduces dimensionality using information criteria to aggregate redundant item dimensions; (*ii*) it dynamically classifies patients into latent subgroups characterized by distinctive disease burden; and (*iii*) it evaluates the covariates that influence the switch between groups over time.

Future work can be developed to propose methods for jointly selecting both the number of latent classes and item aggregation, as well as evaluating the model for its ability to predict events following those observed, so that it can serve as an early warning system.

Acknowledgments. Brusa, Pennoni, and Bartolucci acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by the European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006. The authors also acknowledge Vitaccess and the Scientific Advisory Board for providing access to the data.

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC, Boca Raton (2013)
2. Bartolucci, F., Montanari, G.E., Pandolfi, S.: Three-step estimation of latent Markov models with covariates. Comput. Stat. Data Anal. 83, 287–301 (2015)
3. Bartolucci, F., Bacci, S., Gnaldi, M.: Statistical Analysis of Questionnaire. Chapman & Hall/CRC, Boca Raton (2016)
4. Bolck, A., Croon, M., Hagenaars, J.: Estimating latent structure models with categorical variables: one-step versus three-step estimators. Political Anal. 12, 3–27 (2004)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm (With Discussion). J. R. Stat. Soc. B 39, 1–38 (1977)
6. Dresser, L., Wlodarski, R., Rezania, K., Soliven, B.: Myasthenia Gravis, epidemiology, pathophysiology and clinical manifestations. J. Clin. Med. 10, 22–35 (2021)
7. Samejima, F.: Estimation of ability using a response pattern of graded scores. Psychometric Society, Richmond, VA, 34, 1–97 (1969)