MATRIX-VARIATE HIDDEN MARKOV REGRESSIONS

Salvatore D. Tomarchio [1], Antonio Punzo [1] and Antonello Maruotti [2]

[1] Department of Economics and Business, University of Catania, (e-mail: daniele.tomarchio@unict.it, antonio.punzo@unict.it)

[2] Department of Law, Economics, Political Sciences, and Modern Languages, LUMSA University, (e-mail: a.maruotti@lumsa.it)

ABSTRACT: We present two families of matrix-variate hidden Markov regression models, which differ in how they handle covariates (i.e., as fixed or random). The models achieve parsimony by using the eigen-decomposition of the components' covariance matrices. A two-step fitting strategy is implemented due to the high number of parsimonious models. These models are then investigated on a real dataset.

KEYWORDS: hidden Markov, matrix-variate, model-based clustering.

1 Introduction

Hidden Markov models (HMMs) are widely used for analyzing longitudinal data due to their mathematical flexibility. HMMs can also be modified to incorporate covariates, resulting in hidden Markov regression models (HMRMs), which are useful in regression settings (Bartolucci *et al*, 2012).

Broadly speaking, HMRMs can be divided into two main groups based on whether the covariates contribute to assigning observations to hidden states. The first group involves observed covariates that act as fixed effects shared by all units in the same hidden state, resulting in hidden Markov regression models with fixed covariates (HMRMFCs). Examples of this category can be found in studies by Bartolucci and Farcomeni (2015), and Maruotti and Punzo (2017). The second group, on the other hand, treats observed covariates as random and includes information about their distribution in the model to facilitate clustering. This approach leads to hidden Markov models with random covariates (HMRMRCs) as demonstrated in studies by Punzo *et al* (2018, 2021).

The focus of our study is to present and examine HMRMFCs and HMRMRCs as potential tools for analyzing matrix-variate longitudinal data. These models will be referred to as MV-HMRMFCs and MV-HMRMRCs, respectively. This type of data is typically obtained by observing $P \times R$ matrices of variables for I units over T periods. In essence, the data can be organized into a four-dimensional array with dimensions of $P \times R \times I \times T$.

To achieve parsimony, the two covariance matrices of each hidden state are subjected to eigen-decomposition. Because of the different formulations, the overall number of models is different between the two families. In the case of MV-HMRMFCs, only the covariance matrices of the response variables are available for each state, producing 98 MV-MRMFCs. On the other hand, for MV-HMRMRCs, both the response and covariate covariance matrices are available in each state, leading to 9604 MV-HMRMRCs. Therefore, a convenient approach for fitting the MV-HMRMRCs is employed to reduce the required computational effort.

We examine a dataset obtained from the Italian National Institute of Statistics to explore the relationship between unemployment and labor force participation in the Italian labor market. The data is structured in a two-factor design based on gender and age groups, and it covers four years at the provincial level.

2 Methodology

Let $\{\mathcal{Y}_{it}; i=1,\dots,I, t=1,\dots,T\}$ be a sequence of response variables, where each $\mathcal{Y}_{it}$ is a matrix of dimension $P \times R$ referring to the ith observation for the tth time point. The main assumption of an MV-HMM is that the random matrices in the above sequence are conditionally independent given a hidden process $\{S_{it}; i=1,\dots,I, t=1,\dots,T\}$ that follows a first-order Markov chain with state-space $\{1,\dots,k,\dots,K\}$. This process is governed by the initial probabilities π_{ik} =Pr$(S_{i1}=k)$, $k=1,\dots,K$, and the transition probabilities $\pi_{ik|j}$ =Pr$(S_{it}=k|S_{it-1}=j)$, $t=2,\dots,T$ and $j,k=1,\dots,K$, where j refers to the state previously visited. We assume a matrix-variate normal distribution for the observations at every time occasion, that is, $f\left(\mathcal{Y}_{it}=\mathbf{Y}_{it}|S_{it}=s_{it}\right) \sim MVN_{P\times R}(\mathbf{M}_k, \mathbf{\Sigma}_k, \mathbf{\Psi}_k)$, where $\mathbf{M}_k$ is the $P \times R$ mean matrix, and $\mathbf{\Sigma}_k$ and $\mathbf{\Psi}_k$ are the $P \times P$ and $R \times R$ covariance matrices related to the P rows and R columns, respectively, for latent state k.

In numerous longitudinal studies, apart from the series of responses, there exists a series of covariates $\{\mathcal{X}_{it}; i=1,\dots,I, t=1,\dots,T\}$, being each $\mathcal{X}_{it}$ a matrix of dimension $Q \times R$, that we would like to functionally relate to the former. Thus, we have to extend MV-HMMs to the two regression-based categories introduced in Section 1. By starting with the fixed covariates approach (MV-HMRMFCs), in each latent state k, we are interested in modeling the conditional distribution

$$f\left(\mathcal{Y}_{it}=\mathbf{Y}_{it}|\mathcal{X}_{it}=\mathbf{X}_{it}, S_{it}=k\right), \tag{1}$$

by assuming a linear functional form for its expectation

$$\mathbb{E}(\mathcal{Y}_{it} = \mathbf{Y}_{it} | \mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k; \mathbf{B}_k) = \mathbf{B}_k \mathbf{X}^*_{it}, \tag{2}$$

where $\mathbf{B}_k$ is a $P \times (1+Q)$ matrix of regression coefficients and $\mathbf{X}^*_{it}$ is a $(1+Q) \times R$ matrix having a vector of ones in the first row (to incorporate the intercept in the model) and the Q covariates from the second row onwards.

When the random covariates approach (MV-HMRMRCs) is considered, in each latent state k, we model the joint distribution

$$\begin{aligned} & f(\mathcal{Y}_{it} = \mathbf{Y}_{it}, \mathcal{X}_{it} = \mathbf{X}_{it} | S_{it} = k) = \\ & f(\mathcal{Y}_{it} = \mathbf{Y}_{it} | \mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k) \, f(\mathcal{X}_{it} = \mathbf{X}_{it} | S_{it} = k), \end{aligned} \tag{3}$$

by also assuming (2).

To introduce parsimony in (1) and (3), we apply the eigen-decomposition to the covariance matrices, as commonly done in the model-based clustering literature (see, e.g. Tomarchio *et al*, 2022). This creates two families of models: 98 parsimonious MV-HMRMFCs and 9604 parsimonious MV-HMRMRCs.

Parameter estimation is implemented via a maximum likelihood approach based on the expectation conditional-maximization (ECM) algorithm (Meng and Rubin, 1993) and recursions widely used in the HMM literature (Baum *et al*, 1970). To make computationally affordable the fitting of 9604 parsimonious MV-HMRMRCs, a two-step fitting strategy (not discussed here for the sake of space) is implemented.

From a classification perspective, by using a maximum *a posteriori* probabilities approach (Punzo *et al*, 2021), each unit is classified to one of the K hidden states, at each time point. This information can be useful to track how the observations move between the hidden states as well as to identify which state is mainly sojourned by each observation.

3 Real data example

We examine the relationship between unemployment and the Labor Force Participation (LFP) of 106 Italian provinces, utilizing data from the Italian National Institute of Statistics (ISTAT). Our analysis focuses on the four years from 2018 to 2021. The unemployment and LFP for each province are recorded in a two-factor percentage format, categorized by gender (male and female) and age (15-24, 25-34, 35-49, 50-74). Therefore, both variables are presented in a four-way array format, with dimensions of $2 \times 4 \times 106 \times 4$.

By limiting here our discussion to the results obtained after the fitting of parsimonious MV-HMRMRCs, we found that the best solution according to the Bayesian information criterion (BIC) has $K = 5$ hidden states. The estimated regression coefficients (omitted here due to space constraints) indicate a negative sign in 80% of the cases. This suggests that the so-called discouraged worker effect is widespread across the provinces of Italy. The estimated mean matrices (omitted here due to space constraints) illustrate that the states can be sorted according to the levels of unemployment, both in gender and age factors. Specifically, the unemployment levels consistently decrease from the first state to the fifth state. Looking at the classification obtained by assigning each state to the province it mainly sojourns, it appears that there is a geographical pattern. The first two states seem to be predominantly composed of provinces located in the southern part of Italy, while the other three states appear to contain provinces located in the central and northern parts of the country.

References

BARTOLUCCI, F., & FARCOMENI, A. 2015. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*, **71**, 80–89.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2012. *Latent Markov models for longitudinal data*. Boca Raton: CRC Press.

BAUM, L. E., PETRIE, T., SOULES, G., & 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.

MARUOTTI, A., & PUNZO, A. 2017. Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis*, **113**, 475–496.

MENG, X., & RUBIN, D. B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

PUNZO, A., INGRASSIA, S., & MARUOTTI, A. 2018. Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Statistics in Medicine*, **37**, 2797–2808.

PUNZO, A., INGRASSIA, S., & MARUOTTI, A. 2021. Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. *Statistical Papers*, **62**, 1519–1555.

TOMARCHIO, S. D., PUNZO, A., & MARUOTTI, A. 2022. Parsimonious hidden Markov models for matrix-variate longitudinal data. *Statistics and Computing*, **32**, 1–18.