ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Dynamic clustering and modeling of temporal data subject to common regressive effects

Louise Bonfils, Allou Samé and Latifa Oukhellou

Université Gustave Eiffel, COSYS-GRETTIA, Champs-sur-Marne, France

Abstract.

Clustering is used in many applicative fields to summarize information into a small number of groups. Motivated by behavioral extraction issues from urban data, the interest of this paper is to propose a classification method that allows modeling the evolution of cluster profiles over time while considering common regressive effects. The parameters of the proposed model are estimated using variational approximation because maximum likelihood estimation is not suitable in this case. The ability of the model to estimate parameters is evaluated using various simulated data and compared with two other models.

1 Introduction and Motivation

In many application domains, clustering observations into a reduced set of classes is meaningful to highlight common aspects within the clusters. Considering urban data collected in the energy or mobility domains, clustering gives insights on the typical user behavior patterns ([1], [2]). Customers' habits and preferences can also be classified to build recommendation systems, for example ([3]).

Usually, the clustering of user behaviors or customer preferences does not consider potential changes or evolutions. Incentive policies, price changes or innovations can lead to changes in these behaviors and habits. Thus, it may be interesting to consider the dynamic and evolving aspect of behavior in the classification task. The evolution of behaviors in clustering problems is often taken into account by using segmentation methods to identify periods where behaviors are static and constant, then perform clustering on these specific periods. The segmentation phase can be performed manually based on solid assumptions or using stochastic methods such as Hidden Markov Models ([4]).

This paper presents a model that attempts to group similar observations into a reduced set of clusters while estimating class profiles through a dynamic approach using auto-regressive processes. The proposed model summarizes information in a small group of clusters while dynamics and evolutions are simultaneously considered without going through a first segmentation step. Moreover, the hypothesis that known factors have a global effect on observations is made. The idea is to consider the impact of those exogenous factors, which are common to all observations, and unknown endogenous effects, specific to the clusters, which the model aims to estimate.

As said before, the proposed model aims at building classes in an unsupervised way while modeling class centers dynamically and estimating the effects of common exogenous factors. We place ourselves in the framework of latent variable model using mixture models. These models allow a certain flexibility to build more or less complex models. This is why they are appreciated and often used in the case of classification problems with latent variables ([5]) using in particular the Expectation-maximization algorithm to solve the optimization problem.

However, in [6], the authors point out that the EM algorithm is not suitable for some complex generative models involving multiple dynamics latent processes. In [7] the authors estimate time dependent effects via *Variable Neighborhood Search* algorithms. But, the dynamic aspect of class centers is not taken into account by this algorithm. This topic has been explored in [8] using variational inference methods. These techniques seem to be better suited than the adapted versions of the EM algorithm to solve dynamic latent variable estimation problems. These previous results motivate the choice to use variational approximation to estimate the proposed mixture model.

The second section of this paper is devoted to the construction of the model and its presentation. Section 3 presents the inference methods and the algorithm used for parameters estimation. Finally, the last section is devoted to the testing and evaluation of the proposed model by comparing it with other models using different datasets and three criteria.

2 Model definition

To formalize the model, let's consider the following notations:

- $(\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)$ a set of n observations, where $\mathbf{x}_i = (x_{it})_t$ is a sequence of T observed data, with $\forall t$, $x_{it} \in \mathbb{R}$,
- $\mathbf{u}_t$ $(t \in [\![1, T]\!])$ a $(p+1)$-dimensional vector representing p exogenous and observable factors. We include the constant value 1 in the vector to take into account a level parameter (bias).

The model proposed in this article assumes that the series $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ can be grouped into K clusters. It is characterized by a regressive common component reflecting the effect of known observed factors, and by cluster-specific profiles reflecting the effect of latent dynamic factors. According to this assumption, we consider that x_{it} can be explained by the following model:

$$\forall i \in [\![1, n]\!], \quad \forall t \in [\![1, T]\!]; \qquad x_{it} = \mathbf{u}_t' \mathbf{a} + \sum_{k=1}^{K} z_{ik} b_{kt} + e_{it}, \tag{1}$$

where z_{ik} is a binary variable equal to 1 if the observation i belongs to the class k and 0 otherwise. We assume that z_i, satisfying $z_i = k$ if $z_{ik} = 1$, follows a Multinomial distribution with parameters $\boldsymbol{\pi} = (\pi_k)_{k=1,\dots,K}$. Also, the profile $(b_{kt})_{t=1,\dots,T}$ corresponds to the unobservable group-specific profiles, e_{it} is a centered and normally distributed noise with variance v_k and $\mathbf{a} = (a_0, \dots, a_p) \in \mathbb{R}^{(p+1)}$ refers to the regression coefficients associated to exogenous factors and a_0 denotes the level coefficient.

The latent profiles $(b_{kt})_{t=1,..,T}$ are modeled as first-order auto-regressive processes as follows:

$$\forall t \in [\![1, T]\!], \forall k \in [\![1, K]\!], \quad b_{kt} = \Phi_k b_{kt-1} + \nu_{kt}, \tag{2}$$

where, ν_{kt} is a centered Gaussian noise with variance w_k, and b_{k0} is normally distributed with μ_{k0} and σ_{k0} as mean and variance parameters. The coefficient Φ_k satisfies the stationnarity constraint $|\Phi_k| < 1$. Then, using previous elements, the vector of parameters of the model is as follows:
$\mathbf{\Theta} = \{(v_k, w_k, \pi_k, \Phi_k, \mu_{k0}, \sigma_{k0})_{k=1,...K}, \mathbf{a}\}$.

The model defined by Equation (1) is not identifiable. In fact, the coefficient a_0 can be confused with class profiles $(b_{kt})_{(k,t)}$. In this case, it is necessary to add a constraint to the model. In the present case, by setting $\tilde{\mathbf{a}} = (a_1, \dots, a_p)$, and noting $\tilde{\mathbf{u}}_t$ the corresponding p-dimensional exogenous variables, we have:

$$\mathbf{u}_t'\mathbf{a} + \sum_k z_{ik} b_{kt} = a_0 + \tilde{\mathbf{u}}_t'\tilde{\mathbf{a}} + \sum_k z_{ik} b_{kt} = (a_0 - \alpha) + \tilde{\mathbf{u}}_t'\tilde{\mathbf{a}} + \sum_k z_{ik}(b_{kt} + \alpha).$$

Thus, depending on the value of α, there is an infinite number of choices for a_0 and b_{kt}. To ensure the identifiability, we add the following constraint to the model: $\sum_{k=1}^{K} \sum_{t}^{T} \mathbf{b}_{kt} = 0$.

After presenting the model parameters and assumptions, the next session is dedicated to the theory related to variational inference methods and algorithm used for estimation.

3 Variational inference for parameter estimation

In our case, the complex structure of the model makes parameter estimation via the maximum likelihood method and the EM algorithm intractable. It is therefore necessary to get around this problem by using variational inference methods. To do so, a function $F(q, \mathbf{\Theta})$ is introduced and built from a distribution q over the latent variables ($\mathbf{b}$,$\mathbf{z}$) called *variational distribution*. The function F is defined such as: $F(q, \mathbf{\Theta}) = \mathbf{E}_q(\mathcal{L}_c(\mathbf{\Theta})) + H(q)$, where $H(q)$ is the entropy of the distribution q, and $\mathcal{L}_c$ refers to the complete log-likelihood of the model. The function F is called "The Evidence Lower Bound" because it satisfies the following equation ([9]):

$$\mathcal{L}(\mathbf{x}; \mathbf{\Theta}) \geq F(q(\mathbf{z}, \mathbf{b}), \mathbf{\Theta}).$$

The main goal is to estimate the variational distribution and estimate model parameters by maximizing the Evidence Lower Bound. To simplify this maximization problem while ensuring an accurate parameter estimation, some assumptions are usually made. In that case, the function q has the following form: $q(\mathbf{z}, \mathbf{b}) = \prod_{i=1}^{n} q_z(z_i) \prod_{t=0}^{T} \prod_{k=1}^{K} q_b(b_{kt})$, where q_z is the distribution of latent variable z_i and q_b is the distribution of the processes (b_{kt}). In this model, variables b_{kt} were naturally supposed to be Gaussian with mean parameters m_{kt}

and variance λ_k. The variables z_i are distributed according to a Multinomial distribution with parameters $(\tau_{ik})_{k=1,\dots,K}$. This variational distribution leads us to introduce *variational parameters* that will be estimated by maximizing the Evidence Lower Bound. The variational parameters of the models are:
$\{\boldsymbol{\tau} = \{(\tau_{ik})_{k=1,..,K;i=1,..,n}\}, \mathbf{m} = \{(m_{kt})_{k=1,..,K;t=0,..,T}\}, \boldsymbol{\lambda} = \{(\lambda_k)_{k=1,..,K}\}\}$.

Using previous elements made in Section 2, the Evidence Lower Bound can be explicitly written. The algorithm used for parameter estimation iteratively maximizes the Evidence Lower Bound according to each parameter one by one while considering the others as fixed. Updating class centers variational parameters $(m_{kt}^{(q+1)})_{(k,t)}$ requires an adapted version of Kalman filter ([8]).

The initialization consists in setting a starting point for parameters. Initial values are chosen for variance parameters $(v_k^{(0)}, w_k^{(0)}, \sigma_0^{(0)})$, proportion parameters $(\pi_k^{(0)})$, variational variances $(\lambda_k^{(0)})$. Then, initial values are computed for $(m_{kt}^{(0)})$, $(\tau_{ik}^{(0)})$ and coefficients $\mathbf{a}$ using the K-means algorithm.

It is assumed that the algorithm has converged to a solution when the updated class centers are sufficiently closed to those obtained in the previous iteration. In other words, the stopping criterion for this algorithm is, with $\epsilon \to 0$, $\frac{1}{KT}\sum_{t,k}(m_{kt}^{(q+1)} - m_{kt}^{(q)})^2 < \epsilon$. Once this condition is reached, the algorithm stops.

4 Evaluation of the model

The algorithm has been implemented and tested using simulated data. First, it is important to define criteria to evaluate the performances of the proposed model and to compare them for different data sets, to the performances obtained with models used as references.

4.1 Criteria for model performance evaluation

As a reminder, the model is supposed to be able to identify the global exogenous effects, classify observations, and estimate class centers as dynamic processes. The objective is to evaluate the model on these three aspects using three criteria. Notice that cluster labels have been reorganized to maximize the classification rate. First, the mean square error is used to evaluated the ability of the model to estimate class profiles: $\text{CRIT}_1 = \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}(\hat{m}_{kt} - b_{kt})^2$. Then, the ability of the model to identify and estimate exogenous effects is evaluated using the mean square error computed on exogenous factors such as: $\text{CRIT}_2 = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{u}_t'\hat{\mathbf{a}} - \mathbf{u}_t'\mathbf{a})^2$. Finally, the correct classification rate is used to evaluate the model: $\text{CRIT}_3 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{z_i=\hat{z}_i\}}$.

We will compare the performances of the proposed model to those obtained with two reference models:

- **Constant class center model:** This model consists of estimating class centers as constant values over time.

- **Two-step regression model:** This model estimates the exogenous effects and then the dynamic class centers using an adapted version of the algorithm in which regression coefficients are no longer updated. The comparison between the proposed model and this allows us to show the utility to include exogenous effect estimation inside iterations.

4.2 Results

The model is evaluated by generating various data sets to cover a large sample of different cases. Data sets with two and four classes and different numbers of observations and sequences have been generated. For each configuration, models have been tested on two hundred various datasets. First, we consider the fixed time window $T = 100$ and vary the number of observations ($n = 20$ and $n = 150$). Then, we fix the number of observations to $n = 100$ and set the time window $T = 80$ and $T = 300$. The following results are obtained with four clusters.

	CRIT$_1$		CRIT$_2$		CRIT$_3$	
T=100	**n=20**	**n=150**	**n=20**	**n=150**	**n=20**	**n=150**
Proposed Model	**0,460**	**0,133**	**0,071**	**0,059**	**0,988**	**1,000**
Constant Model	6,994	6,981	0,151	0,151	0,792	0,806
Two-step Regression	0,519	0,333	0,148	0,136	0,987	0,992
n=100	**T=80**	**T=300**	**T=80**	**T=300**	**T=80**	**T=300**
Proposed Model	**0.847**	**0.761**	**0,072**	**0,049**	0,977	**0.999**
Constant Model	3.921	3.669	0.085	0.066	0.610	0.882
Two-step Regression	0.857	0.777	0.085	0.069	**0.978**	0.997

Table 1: Average criteria calculated for data sets of different sizes

According to the results shown in Table 1, we can note that the criteria are decreasing with the size of the temporal window (T) and the number of observations (n). It means that the more data there is, the more accurate the model is. The previous table also shows that, according to all of the three criteria, the proposed model outperforms the other approaches. The performances of the proposed model compared to the model with constant class centers highlight the interest of estimating the class profiles dynamically. Although, the proposed model performances are close to the "two-step regression" ones, we observed that it requires fewer iterations to converge.

5 Conclusion

This paper presents a dynamic latent variable model to solve a classification problem by considering class centers' evolutions over time. Indeed, the main objective of the model is to estimate class profiles as random walks. Moreover, the model is able to estimate the effect of known exogenous factors on the observations.

In this article, the number of clusters K is assumed to be known. Further investigations can be made on the choice of this hyper-parameter thanks to selection criteria such as the BIC criterion for example. The presented model represents a first step towards a more general model where exogenous effects are not global but specific to each cluster, which highlights structural effects within clusters.

Future works should concern the use of the proposed model to characterize occupant behavior in buildings, for a better estimation of energy performances.

References

[1] Juan David Rodriguez Cote and Marco Diana. Exploring the benefits of a traveller clustering approach based on multimodality attitudes and behaviours. *Transportation Research Procedia*, 25:2556–2569, 2017.

[2] Emilie Devijer, Yannig Goude, and Jean-Michel Poggi. Clustering electricity consumers using high-dimensional regression mixture models. *Appl Stochastic Model Bus Ind.*, pages 1–19, 2019.

[3] Fátima Rodrigues and Bruno Ferreira. Product recommendation based on shared customer's behaviour. *Procedia Computer Science*, 100:136–146, 12 2016.

[4] J Liisberg, J.K Møller, H Bloem, J Cipriano, and Henrik Madsen. Hidden markov models for indirect classification of occupant behaviour. *Sustainable Cities and Society*, 27:83–98, 2016.

[5] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley, New-York, 2007.

[6] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[7] Stéphane Bonhomme and Manresa Elena. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83:1147–1184, 2015.

[8] Hani El Assaad, Allou Samé, Gérard Govaert, and Patrice Aknin. A variational expectation-maximisation algorithm for temporal data clustering. *Computational Statistics and Data Analysis*, 103:206–228, 2016.

[9] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2018.