Bayesian statistics project

Louise LIGONNIERE, Amina MANSEUR, Lila MEKKI

January 2025

1 Introduction

This project explores a time series clustering method introduced by Fröhwirth-Schnatter and Kaufmann [3] which aims to group multiple time series using finite-mixture models. We consider a panel of N time series $y_{i,t}$, with timesteps $t = 1, \ldots, T$ and units $i = 1, \ldots, N$. The central hypothesis is that these N time series belong to K latent groups, with all time series within a given group characterized by the same econometric model.

For instance, to model stationary time series, we may use finite mixtures of $AR(p)$ models, where the auto-regressive parameters differ across the K groups: for $k = 1, \ldots, K$, and for i belonging to group k:

$$y_{i,t} = c_k + \delta_{1,k} y_{i,t-1} + ... + \delta_{p,k} y_{i,t-p} + \epsilon_{i,t}, \text{ where } \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_k^2) \tag{1}$$

This base case can be extended to model several interesting problems:

- The clustering of time series based on finite mixtures of dynamic regression models, where dependence on exogenous variables is added ;
- Random-effects models that account for unobserved heterogeneity within each group ;
- Markov switching auto-regressive models that incorporate structural breaks at unknown dates.

In the method we consider, group membership of a certain time series is unknown a priori and is estimated along with the group-specific parameters (unlike methods where grouping is performed a priori). Estimation is performed using Bayesian Markov Chain Monte Carlo simulation methods.

2 The chosen Bayesian methodology

2.1 The model

We first seek to formulate a time series model for each univariate time series $\mathbb{Y}_i = \{y_{i,1}, ..., y_{i,T}\}$, $i = 1, \ldots, N$.

Sampling density : The sampling density for $\mathbb{Y}_i$ is given by (depending on the model we formulate):

$$p(\mathbb{Y}_i|\theta) = \prod_{t=t_0}^{T} p(y_{i,t}|\mathbb{Y}_i^{t-1}, \theta), \text{where } \theta \in \Theta, \mathbb{Y}_i^{t-1} = \{y_{i,1}, ..., y_{i,t-1}\} \tag{2}$$

For instance, if $y_{i,t}$ follows a $AR(p)$ model, we have the following:

$$y_{i,t}|\mathbb{Y}_i^{t-1}, \theta \sim \mathcal{N}(c + \delta_1 y_{i,t-1} + ... + \delta_p y_{i,t-p}, \sigma^2) \tag{3}$$

Clustering : The N time series are assumed to originate from K distinct groups. We define the latent group indicator $S_i = k$ if $\mathbb{Y}_i$ belongs to group k:

$$p(\mathbb{Y}_i|S_i, \theta_1, ..., \theta_K) = p(\mathbb{Y}_i|\theta_{S_i}) = \begin{cases} p(\mathbb{Y}_i|\theta_1) & \text{if } S_i = 1 \\ \vdots & \\ p(\mathbb{Y}_i|\theta_K) & \text{if } S_i = K \end{cases} \tag{4}$$

The same model is applied to all clusters, but with different parameters θ_k for each group. Furthermore, within each cluster, the $\mathbb{Y}_i$'s are assumed to be independent. From this specification, we can derive the following joint distribution:

$$p(\mathbb{Y}_1, ..., \mathbb{Y}_N|S_1, ..., S_N, \theta_1, ..., \theta_K) = \prod_{k=1}^{K} \prod_{i:S_i=k} p(\mathbb{Y}_i|\theta_k) \tag{5}$$

Probabilistic structure for the groups : We also need to define a probabilistic model for the group indicators $\mathbb{S} = (S_1, ..., S_N)$ and a corresponding prior probability. We assume that $S_1, ..., S_N$ are a priori independent. We consider two possible models:

1. In the first case, we assume **complete prior ignorance** regarding the group membership of a certain unit. In that case, S_i has a uniform prior, given by:

$$\Pr(S_i = k|\eta_1, ..., \eta_{K-1}) = \eta_k \tag{6}$$

   where η_k is the relative size of group k and $\eta_K = 1 - \sum_{k=1}^{K-1} \eta_k$. $(\eta_1, ..., \eta_K)$ are unknown model parameters estimated along with the data.

2. In the second case, a unit-specific factor might contain information on how to group the time series. This factor could be economic, geographic, or sociopolitical. We can then define a **logit-type model** for $\Pr(S_i = k)$. For instance, when $K = 2$ and using a unit-specific exogenous variable z_i, we have the following prior for S_i:

$$\Pr(S_i = 2|\gamma_1, \gamma_2, z_i) = \frac{\exp(\gamma_1 + z_i\gamma_2)}{1 + \exp(\gamma_1 + z_i\gamma_2)} \tag{7}$$

   where (γ_1, γ_2) are unknown parameters to be estimated from the data.

The logit-type structure can be extended for more than two groups and more exogenous variables. Additionally, note that if $\gamma_2 = 0$, equality (7) reduces to (6) with a different parametrization for the group sizes. Conversely, if $\gamma_2 \neq 0$, it indicates that z_i helps predict group membership.

2.2 Estimation

In this section, we explain how the model parameters and group indicators are estimated. The **unknown model parameters** are:

- Parameters of the different groups : $(\theta_1, ..., \theta_K)$
- Parameters in the probabilistic distribution of group indicators $(S_i)_{i=1}^N$:

$$\phi = \begin{cases} (\eta_1, ..., \eta_K) & \text{for the ignorance structure (6)} \\ (\gamma_1, \gamma_2) & \text{for the logit-type structure (7)} \end{cases}$$

To estimate these parameters, we first need to define a **prior density** for $(\theta_1, ..., \theta_K, \phi)$. The first assumption is the independence of $\theta_1, ...\theta_K$ and ϕ. We then use the following prior specification:

- Conditionally conjugate priors $p(\theta_k)$ for $\theta_1, ...\theta_K$, that are conjugate to $p(\mathbb{Y}_i|\mathbb{S}, \theta_1, ..., \theta_K)$;
- A Dirichlet prior $(\eta_1, ..., \eta_K) \sim \mathcal{D}(e_0, ..., e_0)$ for the ignorance structure (6) ;
- A normal prior for each γ_j (as there does not exist a conditionally conjugate prior) for the logit-type structure (7).

The estimation is carried out using **Markov Chain Monte Carlo (MCMC)** through data augmentation, alternating between 2 steps:

1. Classification for fixed parameters: for $i = 1, ..., N$, we sample S_i from the posterior distribution given by: for $k = 1, ..., K$:
$$\Pr(S_i = k|\mathbb{Y}, \theta_1, ..., \theta_K, \phi) \propto p(\mathbb{Y}_i|\theta_k) \cdot \Pr(S_i = k|\phi) \tag{8}$$

2. Estimation for a fixed classification:
   - We sample $(\theta_1, ..., \theta_K)$ from the posterior $p(\theta_1, ..., \theta_K|\mathbb{S}, \mathbb{Y})$. If the groups are disjoint, we simply estimate each θ_k within group k using MCMC for time series models.
   - We sample ϕ from the posterior $p(\phi|\mathbb{S}, \mathbb{Y})$:
     - For the ignorance structure (6) with a Dirichlet prior, the posterior of $\phi = (\eta_1, ..., \eta_K)$ follows a Dirichlet distribution[1] ;
     - Under the logit-type structure (7), the posterior $p(\phi|S_1, ..., S_N)$ does not have a closed form, so we use a Metropolis-Hastings algorithm to sample ϕ.

2.3 Choosing the optimal number of clusters

Determining the optimal number of clusters K in model-based clustering is crucial. The selection is based on maximizing the posterior probability using Bayes' theorem :

$$p(M_K \mid y) \propto p(y \mid M_K)p(M_K).$$

where $p(y|M_K)$ is the marginal likelihood and $p(M_K)$ the prior probability of the model M_K.

Two common prior approaches can be used: a **uniform prior** where $p(M_K) = 1/K_{\max}$, where $K_{\max}$ is the maximum number of clusters considered, and a **prior penalizing large clusters**, such as a truncated Poisson distribution $K \sim P(\lambda)$, which reduces the likelihood of too many clusters. Maximizing the posterior probability often reduces to maximizing the marginal likelihood $p(y|M_K)$ [2]:

$$p(y \mid M_K) = \int p(y \mid \psi, K)p(\psi)\, d\psi,$$

where $p(y \mid \psi, K)$ is the likelihood of the data given the parameters ψ and $\psi = (\theta_1, \ldots, \theta_K, \phi, S)$ includes the model parameters [3].

The calculation of $p(y \mid M_K)$ is complex, especially with multiple clusters and complex parametric distributions. To address this, the **bridge sampling** approach proves particularly effective.

Bridge sampling is based on the following identity:

$$p(y) = \frac{\mathbb{E}_{g(\theta)}[p(y \mid \theta)p(\theta)h(\theta)]}{\mathbb{E}_{p(\theta|y)}[h(\theta)g(\theta)]},$$

where $g(\theta)$ is a **proposal distribution** close to the posteriori $p(\theta \mid y)$, and $h(\theta)$ is a **bridge function** that minimizes the relative mean square error. The method combines i.i.d. samples from $g(\theta)$ and $p(\theta \mid y)$ to estimate $p(y)$. The optimal bridge function is explained in Gronau et al. [4][4]. An iterative scheme updates the approximation at each step t:

$$\hat{p}(y)^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{p(y|\tilde{\theta}_i)p(\tilde{\theta}_i)}{s_1 p(y|\tilde{\theta}_i)p(\tilde{\theta}_i)+s_2\hat{p}(y)^{(t)}g(\tilde{\theta}_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{g(\theta_j)}{s_1 p(y|\theta_j)p(\theta_j)+s_2\hat{p}(y)^{(t)}g(\theta_j)}},$$

where $\tilde{\theta}_i \sim g(\theta)$ and $\theta_j \sim p(\theta \mid y)$.

[1] If the prior had parameters $e = (e_1, ..., e_K)$, the updated parameters are given by $e' = (e_1 + x_1, ..., e_K + x_K)$ where $x_1, ..., x_K$ correspond to the observed number of units in each cluster k.

[2] It is true in the case of a uniform prior for $p(M_K)$.

[3] ϕ corresponds to the probabilistic structure parameters for the groups (either $(\eta_1, \ldots, \eta_K)$ or $(\gamma_1, \ldots, \gamma_K)$).

[4] It is given by $h(\theta) = C \cdot \frac{1}{s_1 p(y|\theta)p(\theta)+s_2 p(y)g(\theta)}$, where $s_1 = \frac{N_1}{N_1+N_2}$ and $s_2 = \frac{N_2}{N_1+N_2}$, with N_1 and N_2 representing respectively the number of samples from $p(\theta \mid y)$ and $g(\theta)$.

3 Application

In this study, we apply the time series clustering method previously presented to economic data, specifically focusing on the evolution of the Gross Domestic Product (GDP) of various countries worldwide.

3.1 Data

The data was obtained from the "Global Economic Monitor" database of the World Bank, which provides GDP estimates in constant 2010 dollars for multiple countries. Initial data preprocessing was carried out to remove missing values:

- The first years of observation of the time series were excluded due to a high number of missing values ;
- Countries with remaining missing values were also removed.

After preprocessing, we retained complete annual time series for 71 countries, spanning the period from 2000 to 2024. To ensure stationarity and normalization of the series, we calculated the annual GDP growth rate using the following formula:

$$\tau_t = \frac{y_t - y_{t-1}}{y_{t-1}},$$

where y_t represents the GDP at year t, and y_{t-1} the GDP of the previous year. This transformation enables us to work with data that are better suited for time series analysis, while enhancing their robustness and interpretability.

3.2 Model selection

To select the optimal model, we tested several configurations, including $AR(1)$ and $AR(2)$ models, with and without constant term, by pooling all time series. After comparing model performance in terms of likelihood, we selected the $AR(1)$ model with no constant. This choice ensures a simpler solution while maximizing likelihood and reducing the risk of overfitting. We then have for i belonging to group k:

$$y_{i,t} = \delta_{1,k} y_{i,t-1} + \epsilon_{i,t}, \text{ where } \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_k^2) \tag{9}$$

We thus have 2 parameters to estimate for each group, and we specify the following priors: a Gaussian distribution for θ_k, and a Half-normal distribution for σ_k.

3.3 Update of parameters in an auto-regressive model

Since we estimate a $AR(1)$ model in each cluster (in step 2 of the estimation procedure), we need to update parameters using a Bayesian approach. To do so, we first have to compute the conditional sampling density:

$$\begin{aligned} p(\mathbb{Y}_k|\theta_k, \sigma_k) &= \prod_{i \in k} p(y_{i,1}, ..., y_{i,T}|\theta_k, \sigma_k) \\ &= \prod_{i \in k} p(y_{i,1}|\theta_k, \sigma_k) \prod_{t=2}^{T} p(y_{i,t}|y_{i,t-1}, ..., y_{i,1}, \theta_k, \sigma_k) \\ &= \prod_{i \in k} p(y_{i,1}|\theta_k, \sigma_k) \prod_{t=2}^{T} \exp\left(\frac{-1}{2\sigma_k^2}(y_{i,t} - \alpha_k - \beta_k y_{i,t-1})^2\right) \end{aligned}$$

We can then update the parameters using Gibbs sampler by iteratively sampling (using the priors we previously specified for $p(\theta_k)$ and $p(\sigma_k)$) [5]:

$$\begin{cases} \theta_k^{(t+1)} \sim p(\theta_k|\theta_k^{(t)}, \sigma_k^{(t)}, \mathbb{Y}_k) \propto p(\mathbb{Y}_k|\theta_k, \sigma_k) \cdot p(\theta_k) \\ \sigma_k^{(t+1)} \sim p(\sigma_k|\theta_k^{(t)}, \sigma_k^{(t)}, \mathbb{Y}_k) \propto p(\mathbb{Y}_k|\theta_k, \sigma_k) \cdot p(\sigma_k) \end{cases}$$

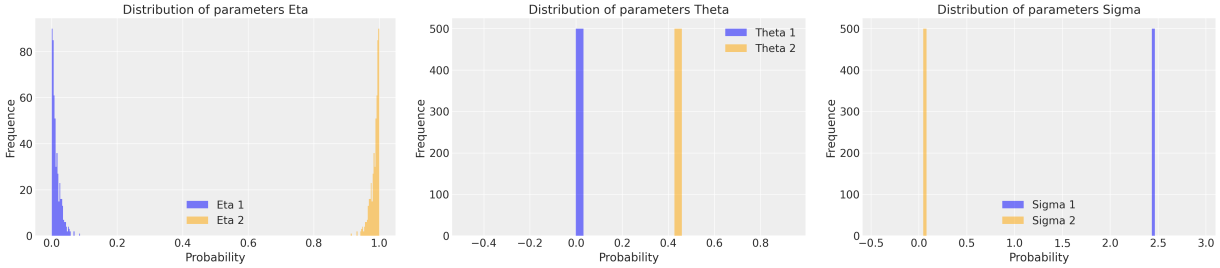
[5] See Chib [2], Barnett G. [1].

3.4 Difficulties encountered

During the implementation, several challenges were identified:

- **Difficulty in selecting an appropriate model a priori:** Since the model is tested across all time series together, identifying the most relevant model beforehand was not straightforward.
- **Difficulty in selecting an exogenous variable for determining cluster membership:** Incorporating an external factor (such as an economic or geographic variable) into the clustering process posed challenges in both choice and validation.
- **Computational intensity of the model:** The estimation of our Bayesian model proved to be computationally expensive, requiring substantial resources. Due to these constraints, we limited the number of iterations to 500, balancing computational feasibility and convergence.
- **Implementation:** The implementation of the bridge sampling function proved to be quite complex.

4 Results



(a) Distribution of the Eta's estimators

(b) Distribution of the Theta's estimators

(c) Distribution of the Sigma's estimators

Figure 1: Distribution of the parameter estimators: Eta, Theta, and Sigma.

After running the algorithm on the ignorance structure for $K = 2$ clusters and a number $n = 500$ iterations, we get the following results. All countries appear in the same cluster. We believe that the issue is not obviously based on the model (since we tried both the ignorance and logit structure for simulated data and it featured several clusters) but rather on the nature of the time series considered. Indeed, the GDP dynamics can be quite correlated on the whole even if the GDP differ from one country to another. We tried to apply the logit model too, but the input exogeneous variable we choosed seemed to be inappropriate for the execution of the model.

5 Conclusion

Tests run on simulated data proved efficient in detecting clusters of time series which was enthusiasming at first. However, shifting to real data, we have not been able to recover clusters. We reckon that the nature of the data did not allow us to get the clusters. It would have been interesting to run our algorithms on other data, though quite computationally costly.

Code

You can find the GitHub repository for this project at the following link: `https://github.com/AminaManseur29/Bayesian_statistics_project.git`.

References

[1] Barnett G., Kohn R., S. S. W. J. (1995). Markov chain monte carlo estimation of autoregressive models with application to metal pollutant concentration in sludge. *Mathl. Comput. Modelling Vol. 22, No. 10-12, pp. 7-13.*

[2] Chib, S. (2001). Markov chain monte carlo methods: Computation and inference. *Handbook of Econometrics.*

[3] Fröhwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89.

[4] Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2020). A tutorial on bridge sampling. *Journal of Mathematical Psychology.* Available online at `www.elsevier.com/locate/jmp`.