POLITECNICO
MILANO 1863

Cluster nations by educational performances

Group R9

Tutors: Alessandro Carminati, Alessandra Ragni

Students: Gaia Caringi, Ignacio Cunado, Alice Flamigni, Sara Fregnan, Francesca Gamba, Marco Ronchetti

Introduction

Educational attainment is a fundamental pillar of national development, influencing economic growth, social mobility, and overall quality of life.

GOAL: **Cluster nations** based on the percentage of low-achieving students in their schools, where a low-achieving student is defined as **having a math proficiency below Level 3.**

Dataset

From the OECD Program for International Student Assessment (PISA) we selected two datasets:

- the first one related to **students questionnaire** (administered to students)
- the second one related to **schools questionnaire** (administered to school principals)

81 countries and economies participated in the 2018 assessment, from which we extracted data on mathematics scores.

Dataset

	CNTSCHID	CNT	PRIVATESCH	STRATIO	SCHSIZE	SCH_TESTED	sum_MATH1below	mean_ESCS	mean_ESCS_std	Y_MATH1	Y_MATH1_rate	Y_BIN_MATH1
1	800006	Albania	public	18.0000	315	19	15	-1.3437474	-0.6578306	15	0.7894737	1
2	800035	Albania	public	11.6800	292	11	7	-0.8204273	0.2044703	7	0.6363636	0
3	800037	Albania	public	22.7239	1852	16	10	-0.5431063	0.6614261	10	0.6250000	0
4	800078	Albania	private	10.4516	162	14	12	-0.8737857	0.1165489	12	0.8571429	1
5	800112	Albania	public	16.1481	436	11	9	-1.2905545	-0.5701821	9	0.8181818	0
6	800116	Albania	public	18.4000	460	19	11	-0.1619158	1.2895329	11	0.5789474	1
7	800131	Albania	public	14.8214	415	10	7	-0.6617900	0.4658649	7	0.7000000	0
8	800132	Albania	public	14.4571	253	10	8	-0.4052100	0.8886447	8	0.8000000	0
9	800133	Albania	public	12.7273	210	10	10	-1.8296800	-1.4585263	10	1.0000000	0
10	800134	Albania	private	9.2593	125	13	11	0.5048385	2.3881775	11	0.8461538	1

Figure: First ten observations of preprocessed dataset

Covariates:

- **CNT** = country name
- **PRIVATESCH** = explains whether a school is public or private
- **STRATIO** = (# of students) / (# of teachers)
- **SCHSIZE** = size of a school
- **SCH-TESTED** = number of students tested in that school
- **mean-ESCS** = mean of socio-economic status
- **mean-ESCS-std** = Standardized mean socio-economic status of students

Target:

Y-MATH1-rate = rate of low-achieving students related to plausible value 1

$$g(\mu_{it}) = \mathbf{X}_{it}\beta + \mathbf{Z}_{it}\mathbf{b}_t$$

where:

- i: countries i=1,...I
- t: schools t=1,...T
- $g(\cdot)$: Link function.
- μ_{it}: Expected value of the response.
- $\mathbf{X}_{it}$: Matrix of predictors for fixed effects.
- β: Vector of coefficients for fixed effects.
- $\mathbf{Z}_{it}$: Matrix of predictors for random effects.
- $\mathbf{b}_t$: Random effects for group j.

A Dirichlet Process (DP) is the natural infinite-dimensional extension of the analogue finite dimensional Dirichlet prior.
It is a **non-parametric probabilistic model** used to define a distribution over distributions.

In a Mixture of Dirichlet Processes the clusters are not predetermined, but they are modeled by the Dirichlet process, which has the property of "creating" new clusters as new data is observed.

Methodology: Bayesian semi-parametric approach

Specializing the general model described in the paper by Kleinman and Ibrahim (1998), we decided to model our problem through a Poisson regression:

$$
\begin{aligned}
y_{it} \mid \boldsymbol{\beta}, b_i &\overset{\text{ind}}{\sim} \mathcal{P}\left(\exp\left(\boldsymbol{X}_{it}^T \boldsymbol{\beta} + b_i + \log(T_{it})\right)\right) \\
\boldsymbol{\beta} &\sim \mathcal{N}_p(0, \sigma_\beta^2 \boldsymbol{I}_p) \\
b_i \mid G &\overset{\text{iid}}{\sim} G \\
G &\sim DP(M, G_0) \\
G_0 &\sim \mathcal{N}(\mu_0, \sigma_0^2)
\end{aligned}
$$

- M: Precision parameter, controlling the variability of the Dirichlet process.
- T_{it}: Number of students in country i and school t.
- $\log(T_{it})$: Offset term to normalize the count data.
- y_{it}: Ratio of low-achieving students.
- b_i: Clustering component from the Dirichlet process, shared by subjects in the same cluster.

Software: Julia
Package: Turing

Why Julia and not Python?

- Similar languages
- Turing in Julia is an equally powerful alternative, but **more flexible**
- Julia provides superior performance in terms of speed.

Dirichlet Process Construction via Stick-Breaking

Stick-Breaking Process: A method for iteratively constructing a Dirichlet Process by breaking a "stick" into smaller pieces.

- **Inizialization**: Start with a stick of length 1, representing the total probability mass.
- **Step 1:** Sample the fraction v_1 for the first component and assign probability $p_1 = v_1$ to it:

$$v_1 \sim \text{Beta}(1, \alpha)$$

- **Step 2:** For each subsequent component, sample a fraction v_k from Beta:

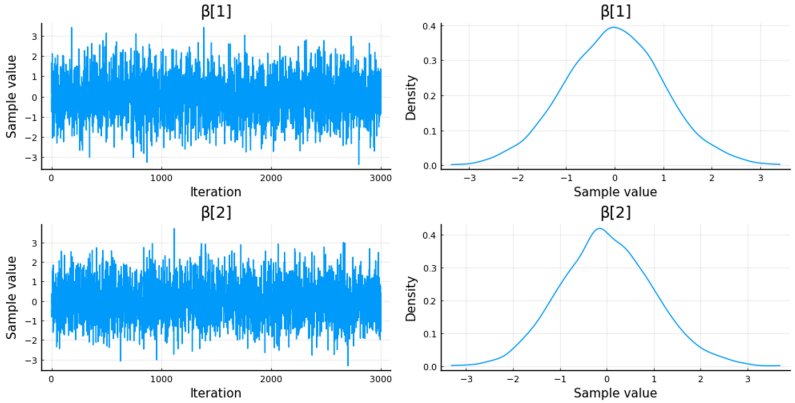
$$v_k \sim \text{Beta}(1, \alpha)$$

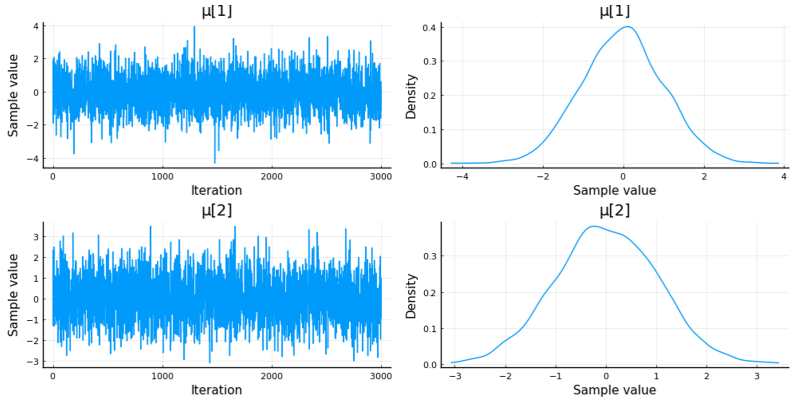
The probability of each new component is:

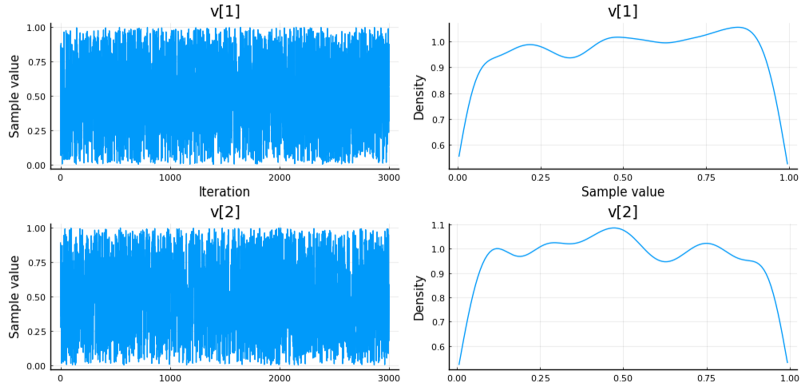
$$p_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

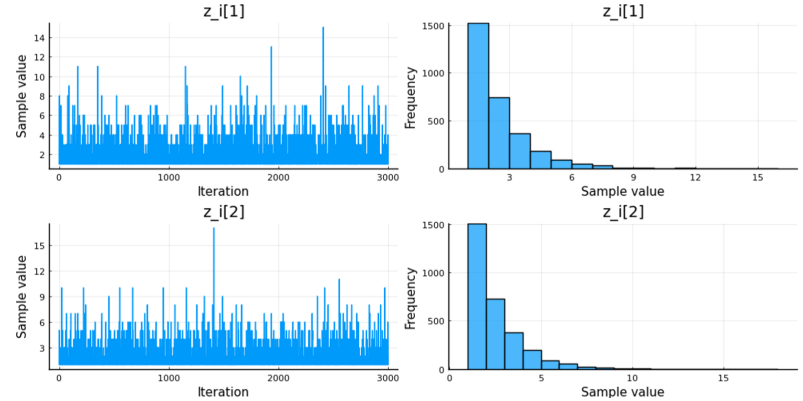
- **Continue** until the total probability mass is fully assigned.

Application: convergence of the chains



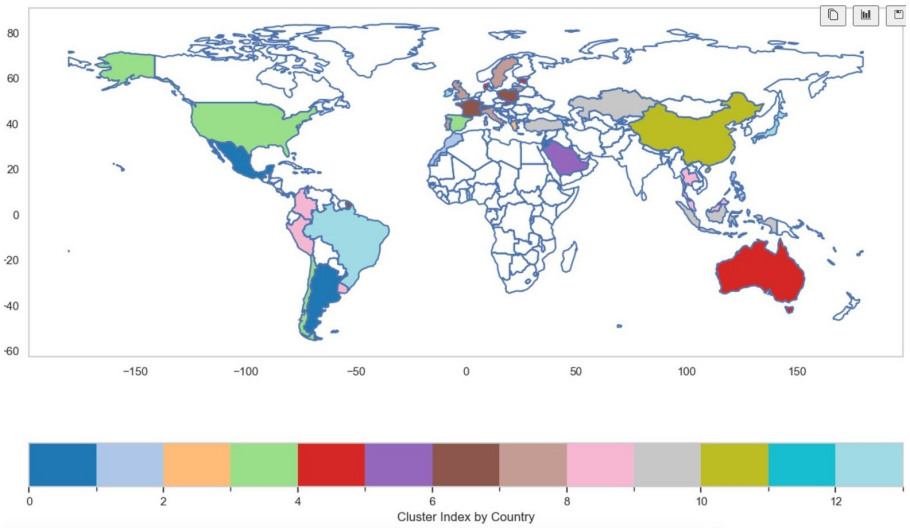




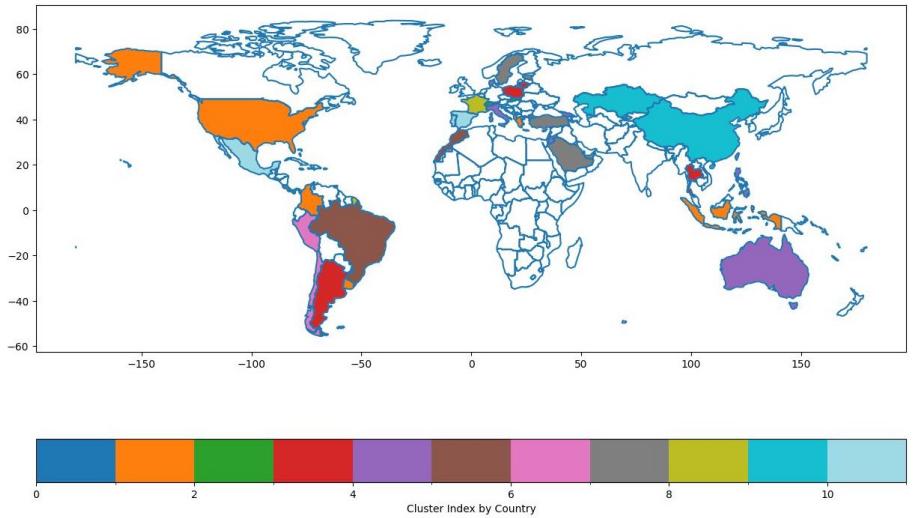


Application: clustering

Frequentist approach: 13 clusters



Bayesian approach: 11 clusters



Future developments

- Revision of the clustering
- Interpretation of the common factors within a cluster and of the differences between them
- Literature review of the paper by Rodriguez et al.
- Implementation of the nested Dirichlet Process (Rodrìguez, Dunson, and Gelfand 2008).

References

[1] Ken P. Kleinman and Joseph G. Ibrahim. A semi-parametric bayesian approach to generalized linear mixed models. *Statist. Med.*, page 2579—2596, 1998.

[2] Peter Muller, Fernando Andres Quintana, Alejandro Jara, and Tim Hanson. *Bayesian Non Parametric Data Analysis*. Springer, 2015.

[3] Abel Rodríguez, David B. Dunson, and Alan E. Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 2008.