Modello proposto (21/7/2023)

Notazione indici

$$
\begin{aligned}
&i : \text{nazione} \quad 1, \dots, n = 4 \\
&x : \text{età} \quad 1, \dots, X = 87 \\
&t : \text{tempo} \quad 1, \dots, T = 101 \\
&j : \text{coefficienti spline} \quad 1, \dots, p = 20 \\
&k : \text{cluster} \quad 1, \dots, K_j = \text{numero di cluster per coeff. j.}
\end{aligned}
$$

Modello per i tassi di morte m_ixt, definiti come il rapporto tra il numero di morti d_{ixt} e gli esposti E_{ixt}:

$$
\begin{aligned}
Y_{ixt} &= \log m_{ixt} = \log \frac{d_{ixt}}{\mathrm{E}_{ixt}} \\
Y_{ixt} &\overset{\perp\!\!\!\perp}{\sim} N\left(\sum_{j=1}^{p} \beta^{\star}_{c_{ixtjt}, jt}, \sigma_i^2 \right) \\
\beta^{\star}_{kjt} &\overset{\perp\!\!\!\perp}{\sim} N(\theta_{jt}, \tau_{jt}) \\
(\theta_{jt}, \tau_{jt}) &\overset{\perp\!\!\!\perp}{\sim} N(\phi_j, \delta_j) \times U(0, A_\tau) \\
(\phi_j, \delta_j) &\overset{\perp\!\!\!\perp}{\sim} N(\lambda, \xi) \times U(0, A_\delta) \\
(\lambda, \xi) &\overset{\perp\!\!\!\perp}{\sim} N(m_0, s_0^2) \times U(0, A_\xi) \\
\sigma_i^2 &\overset{\perp\!\!\!\perp}{\sim} U(0, A_\sigma) \\
\{\mathbf{c}_{j1}, \dots, \mathbf{c}_{jt}, \dots, \mathbf{c}_{jT}\} &\overset{\perp\!\!\!\perp}{\sim} tRPM(\alpha_j, M) \\
\alpha_j &\overset{\perp\!\!\!\perp}{\sim} Beta(a_\alpha, b_\alpha)
\end{aligned}
$$

dove $\mathbf{c}_{jt} = (c_{1jt}, \dots, c_{ijt}, \dots, c_{njt})$ sono le label delle n osservazioni dei cluster per il j-esimo coefficiente al tempo t. $A_\tau = A_\delta = A_\xi = 10, A_\sigma = 5, m_0 = 0, s_0^2 = 1, M = 2$ e $a_\alpha = b_\alpha = 1$ sono iperparametri fissati. In particolare M è il parametro di concentrazione del CRP alla base del modello tRPM.

Inferenza

Ho simulato 2000 campioni dalla posteriori tramite il Gibbs Sampler presentato nel materiale supplementare del lavoro di Page, Quintana e Dahl (https://doi.org/10.1080/10618600.2021.1987255
Ho buttato via le prime 500 di burn-in.

Le prime cose che ho indagato sono:

- il numero medio di cluster per ogni coefficiente e la sua evoluzione nel tempo,
- la probabilità di co-clustering delle nazioni per alcuni coefficienti.

L'andamento del numero medio di cluster è riassunto dalla seguente immagine.
Ci sono due cose che vorrei capire meglio:

- perché il numero medio di cluster è sempre molto prossimo all'1,
- cosa succede negli anni più recenti per avere quelle crescite (covid?).

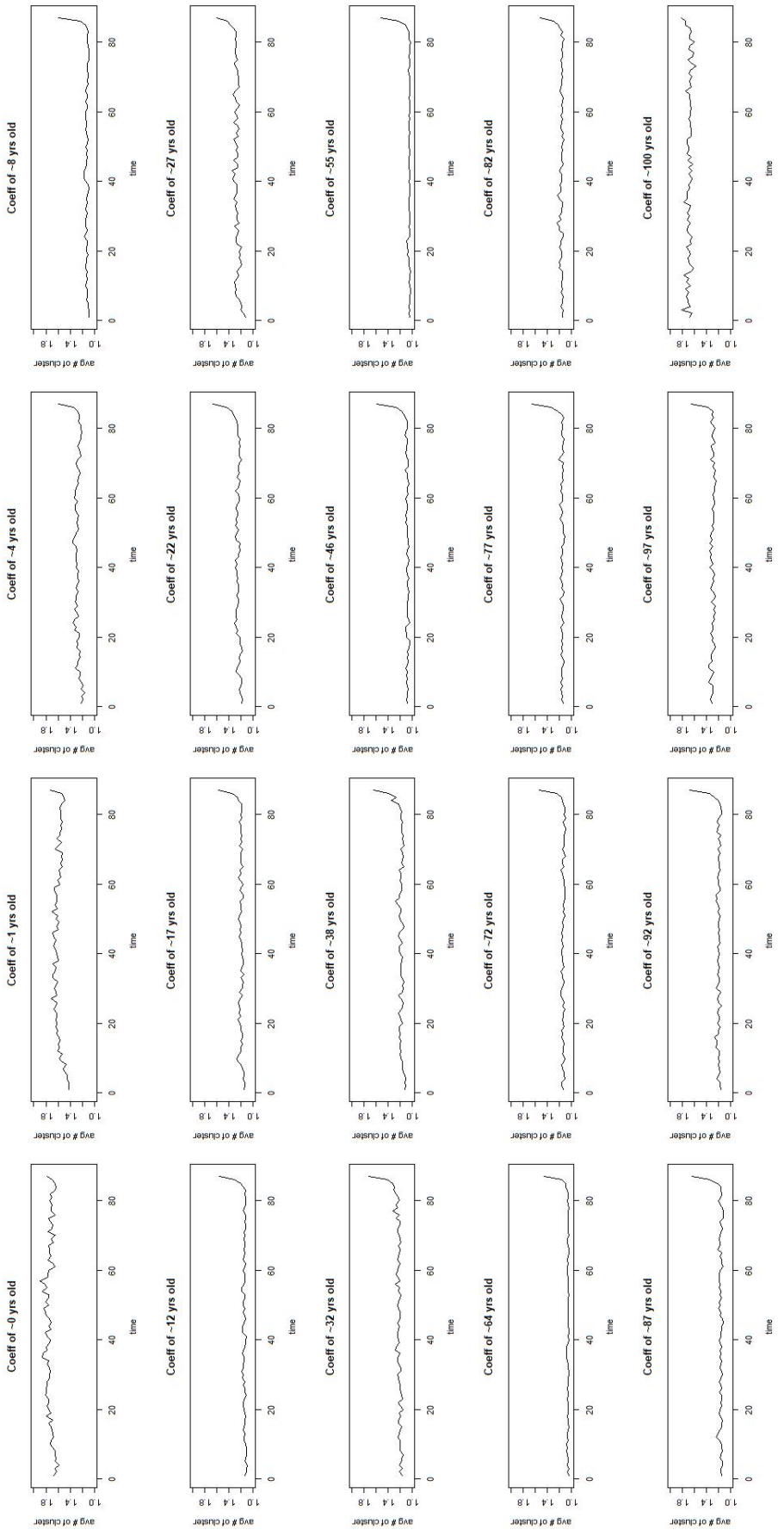

Figure 1: Evoluzione del numero medio di cluster per i 20 coefficienti.

Successivamente ho indagato le probabilità empiriche di co-clustering per vedere se riflettono ciò che si osserva nella Figura 4 del lavoro di Pavone, Legramanti e Durante (https://arxiv.org/abs/2209.12047), in particolare i coefficienti di *Infant* e *Adult* negli uomini.

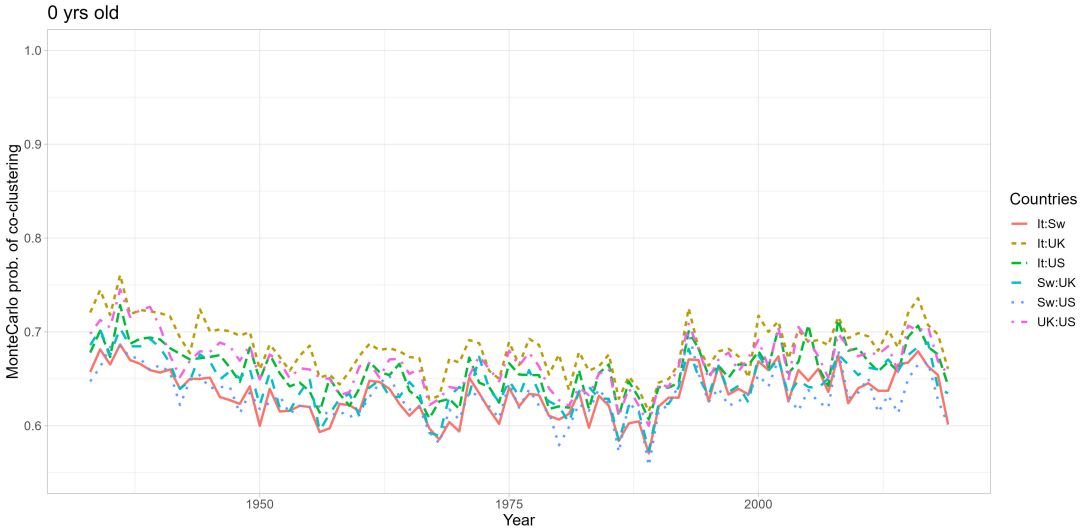

Figure 2: Probabilità di co-clustering per *Infant*.

Figure 3: Probabilità di co-clustering per *Adult*.

L'aspetto positivo che penso si possa trarre da questi due plot è che c'è una netta differenza nella "scala" delle probabilità per gli *Adult* e per gli *Infant*, che rispecchia quanto si vede nel lavoro di Pavone. Purtroppo però, confrontando le diverse curve del grafico degli *Infant*, non si notano alcune differenze nette che si vedono invece nel lavoro di Pavone, in particolare:

- fino a poco dopo il 1980 ci sono 3 cluster, con UK e US in uno e le altre due nazioni isolate;
- da poco dopo il 1980 anche UK e US si separano.

Perciò, mi aspetterei che, nel mio grafico, la curva relativa a UK:US sia nettamente più in alto delle altre per i primi 50 anni e poi che torni sul livello delle altre curve. Al contrario, mi pare che sia tutto piuttosto uguale, sia tra diverse coppie di nazioni che nel corso del tempo.