Estimation of mortality curves using a Dirichlet process prior

Davide Agnoletto, Tommaso Rigon & Bruno Scarpa

Statistical methods and models for complex data — Poster Session, 21st September 2022

Introduction

Modeling human mortality has been a challenge involving many statisticians over the years. The approach that we adopted models a population of age-at-death mortality curves as a mixture of Multinomial random variables. Prior knowledge about the phenomenon is expressed assuming the ideal exact ages-at-death to follow a Dirichlet process.

Motivating application

- Consider a real data problem: mortality curve of each Italian municipality for year 2020 (male population).
- The curves referring to small municipalities are affected by a large amount of noise due to their small population and to the consequent small number of deaths.

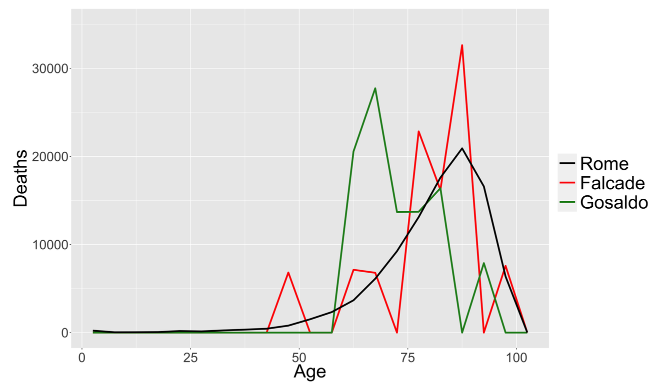

Figure 1. The mortality curve for a heavily populated municipality as Rome is smooth, while those of mountain municipalities as Falcade and Gosaldo are very irregular.

- Then we need a model which produces a smooth estimated curve for each municipality, representing the true signal (i.e. the true behavior of mortality phenomenon) hidden in each curve.

Model formulation

Exact ages at death

- Denote with $\mathbf{y} = (y_1, \dots, y_n)$ the vector of exact ages at death of each subject for a population of size n (usually $n = 10^5$).
- A flexible nonparametric Bayesian density estimation model for $\mathbf{y}$ is

$$y_i \mid \tilde{p} \sim \tilde{p}$$
$$\tilde{p} \sim \mathrm{DP}(\alpha, P_0)$$

$P_0 \rightarrow$ base probability measure providing the initial information on $\tilde{p}$;

$\alpha \rightarrow$ precision parameter controlling the degree of shrinkage of $\tilde{p}$ towards P_0.

- Since a Dirichlet process induces a finite-dimensional Dirichlet distribution when support is partitioned, then $\mathbb{P}\{y_i \in [0,5)\}, \dots, \mathbb{P}\{y_i \in [100, +\infty)\}$ is distributed as a $\mathrm{Dir}(\alpha P_0[0,5), \dots, \alpha P_0[100,+\infty))$, where $P_0[x, x+5)$ represents the probability mass assigned to each age class by the base measure.

Mixture model

- Unfortunately, the exact age at death of each subject is an **ideal** and **unknown** information, however the observed 5-years-age-classes age-at-death distribution is a simple aggregation of $\mathbf{y}$

$$d_x = \sum_{i=1}^{n} \mathbb{1}\{y_i \in [x, x+5)\}.$$

→ We can think at each curve as the outcome of n realizations from a 21-classes multinomial random variable and the population of J raw curves to come from at most H latent groups

$$d_0^j, \dots, d_{100}^j \mid G_j = h \overset{\text{i.i.d.}}{\sim} \mathrm{Multinomial}(n, \pi_{0h}, \dots, \pi_{100h})$$
$$G_j \sim \mathrm{Cat}(1, w_1, \dots, w_H).$$

Prior specification

- Prior distribution for mixture weights is chosen to favor automatic adaption of the model dimension

$$w_1, \dots, w_H \sim \mathrm{Dir}\left(\frac{1}{H}, \dots, \frac{1}{H}\right).$$

- The induced prior distribution for each group h is

$$\pi_{0h}, \dots, \pi_{100h} \sim \mathrm{Dir}(\alpha P_0[0,5), \dots, \alpha P_0[100, +\infty)).$$

Some remarks

✓ Each estimated curve is based on the information coming from the raw curves in the group and from the base measure, hence the model provides a kind of **borrowing of information**.

✓ The model automatically learns the number of clusters.

✓ Different shapes and trends of the curves are well detected.

! Choice of α is critical.

Gibbs-sampling

1. Update group composition: $\mathbb{P}(G_j = h \mid -) \propto w_h \cdot \pi_{0h}^{d_0^j} \cdot \dots \cdot \pi_{100h}^{d_{100}^j}$;
2. Update mixture weights

$$w_1, \dots, w_H \mid - \sim \mathrm{Dir}\left(\frac{1}{H} + s_1, \dots, \frac{1}{H} + s_H\right)$$

where $s_h = \sum_j \mathbb{1}_{(G_j = h)}$ indicates the size of h-th group;

3. Update deaths probabilities for each group

$$\pi_{0h}, \dots, \pi_{100h} \mid - \sim \mathrm{Dir}\left(a^*_{0,h}, \dots, a^*_{100,h}\right)$$

where $a^*_{x,h} = \alpha P_0[x, x+5) + \sum_{j: G_j = h} d_x^j$.

2020 Italian municipalities data analysis

- 7763 raw curves are considered.
- 2020 Italian male population curve is chosen as base measure.

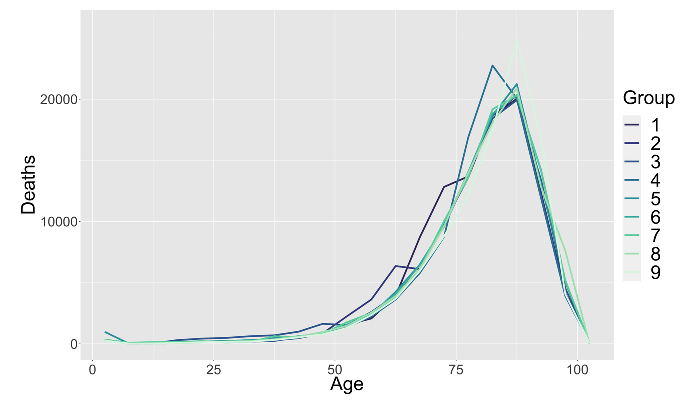

Figure 2. Estimated curves for each group.

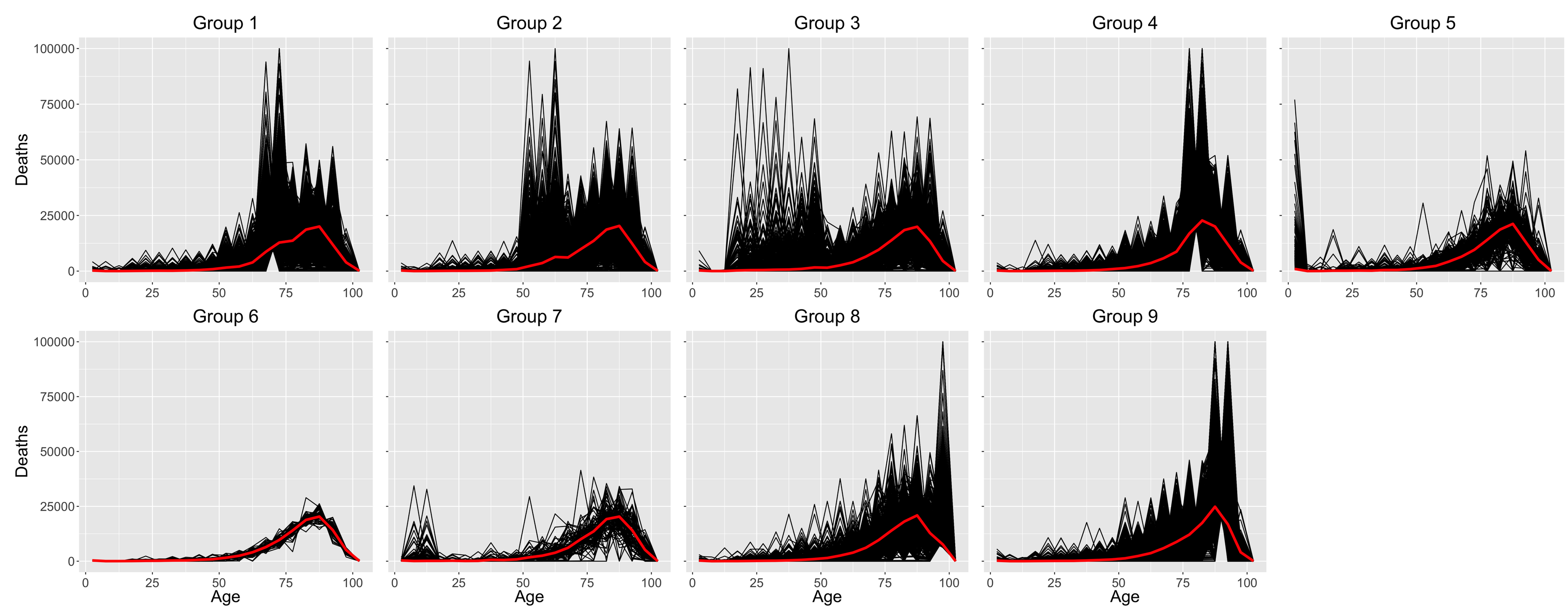

Figure 3. Detected groups and corresponding estimated curves.

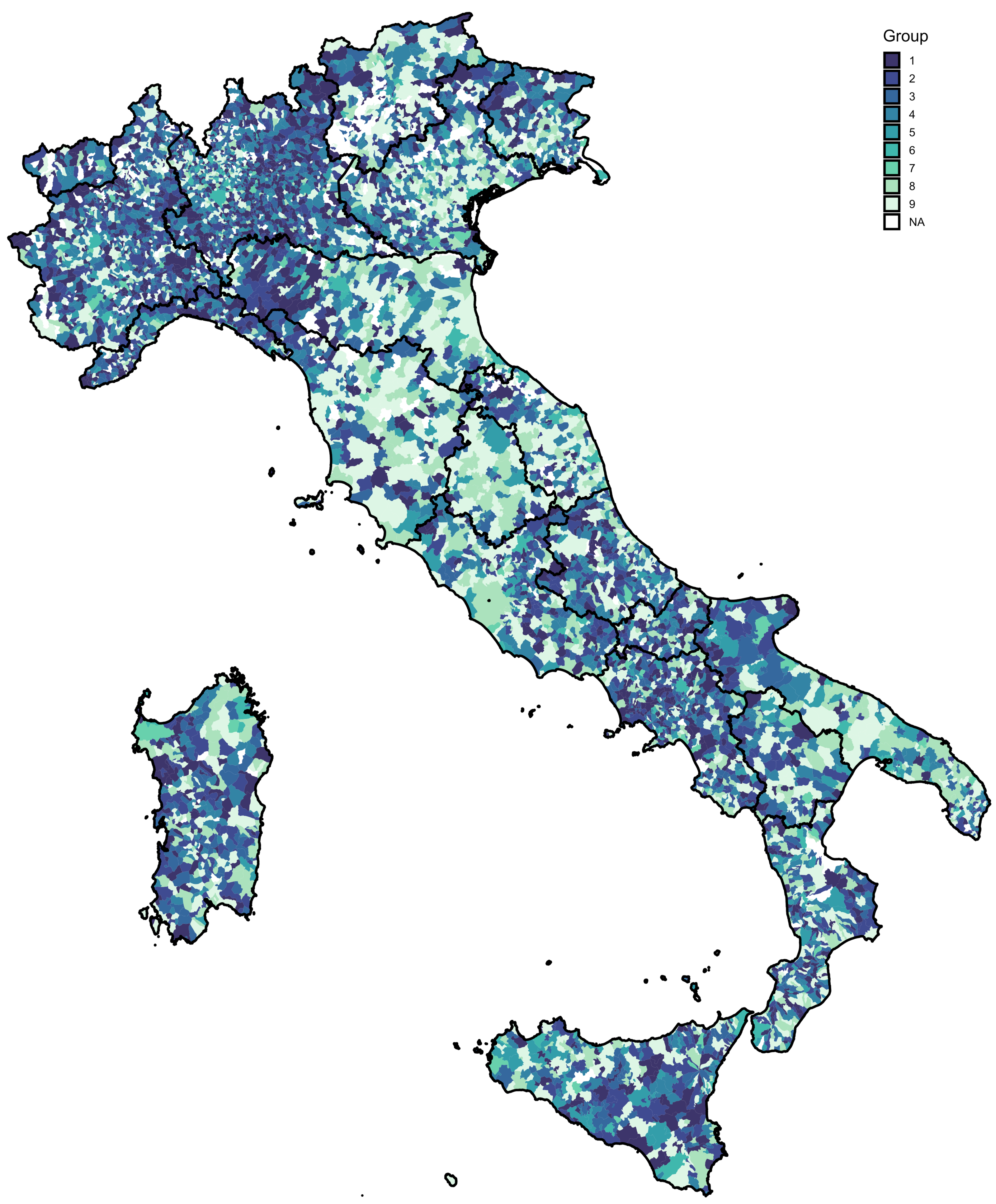

Figure 4. Group of each Italian municipality detected by the model.

References

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Hjort, N. L., Holmes, C., Müller, P., Walker, S. G. (Eds.). (2010). Bayesian nonparametrics (Vol. 28). Cambridge University Press.

Aliverti, E., Mazzuco, S., Scarpa, B. (2021). Dynamic modeling of mortality via mixtures of skewed distribution functions. *arXiv preprint arXiv:2102.01599*.

Contact information

Davide Agnoletto, PhD student
University of Padua
davide.agnoletto.2@phd.unipd.it