Clustering mortality data with nonparametric Bayesian approach

Ozan Evkaya, The University of Edinburgh, oevkaya@ed.ac.uk

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

INTRODUCTION

Recently, there has been an increasing interest on modeling the mortality curve by using non-parametric Bayesian methodology. Within this context, the use of Dirichlet Process (DP) is very active for modeling the mortality. In this regard, the primary goal is to investigate the benefits of DP with a suitable base measure for describing the age-at-death distribution. To establish a more flexible model, the correlation across both the years and ages should be incorporated. Specifically, the impact of both correlations among the years and ages are taken into account with a suitably defined separable covariance function. The use of the separable covariance function is showcased to model the mortality via CRP based posterior calculation, relying on a Blocked Gibbs sampling. To illustrate, the proposed method is considered for clustering the mortality curves of Germany over limited set of years. Specifically, a well-designed MCMC tool for this purpose is essential to overcome the computation time problem.

KEY CONCEPTS

Dirichlet Process (DP)

Generally, a Dirichlet Process (DP) is a kind of stochastic process that describes the "distributions over distributions" with two main parameters;

- Concentration parameter: $\alpha > 0$
- Base measure: G_0

In this setting, each draw from a DP is a random probability measure G over Θ (a measure space), such that for a finite measurable partition of $(A_1, \ldots, A_k)$ on Θ, the random vector defined as $(G(A_1), \ldots, G(A_k))$ is distributed as a finite dimensional Dirichlet distribution with parameters $(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$, ie. $(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$. Here notice that G is distributed based on a DP with a base distribution G_0 such as $G \sim DP(\alpha, G_0)$ where α is the concentration parameter. Modeling via DP is equivalent to assume that the prior over G is DP and G is generated from DP, ie. $G \sim DP(\alpha, G_0)$. In this respect, one can summarize DP as,

$$\widehat{\theta}_i | G \sim G, \forall i \in 1, \ldots, n$$

$$G | \alpha, G_0 \sim DP(\alpha, G_0)$$

Chinese Restaurant Process (CRP)

In general, CRP representation simply based on the metaphor of the seemingly infinite supply of tables at a certain Chinese restaurant. Based on this analogy; the tables are like clusters and the customers denote the observations. CRP is an exchangeable process and it is more convenient to work with the cluster indices z_i (unknown cluster labels) for $i = 1, \ldots, n$, an indicator random variable to represent the label of unique value of θ_{z_i} of θ'_i s.t. $\theta'_i = \theta_{z_i}$ for all $i \in 1, \ldots, n$ to specify which value of θ_k should be used for parameter updates.

Here, the CRP provides a distribution on the finite partitions of the data. A simple visualization of the CRP can be seen as follows,

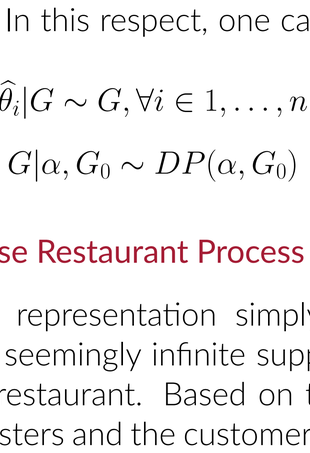

Figure 1. CRP Illustration

In CRP mixture model, the prior term $CRP(z_{1:(i-1)}; \alpha)$ is completed with a likelihood function with parameters θ_k for each cluster k, and a prior distribution G_0 for the parameters. In that respect, the CRP model can be summarized according to the following generating mechanism;

$$z_i \sim CRP(z_{1:(i-1)}; \alpha)$$

$$\theta_{z_i} | G_0 \sim G_0$$

$$\mathbf{x_i} | \theta_{z_i} \sim p(.|\theta_{z_i})$$

When $\alpha \to 0$, the number of clusters, $Num_c \to 1$. When $\alpha \to \infty$, the number of clusters, $Num_c \to N$ where N is the total number of customers (each customer choose a new table)

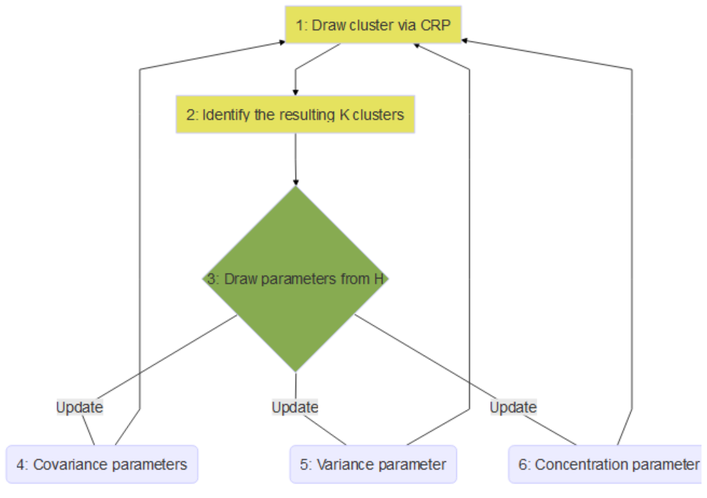

Figure 2. Steps of CRP model

The CRP model steps follow the Blocked Gibbs sampling methodology under the umbrella of two main algorithms given below.

METHODOLOGY

By following the CRP construction of DP mixture model, the proposed methodology is relied on the use of 6 steps Blocked Gibbs sampling for overall posterior calculations. The details can be summarized below by two jointly working algorithms;

▪ ALGORITHM 1: CRP for Partitioning

Inputs: Mortality Data Set ($MDS_{T \times d}$), Concentration parameter α, Priors μ_0, Σ_0 and Σ_y, and initial cluster membership c^0 (set to the same value as an initial cluster)

Output: Given the all parameters and the data set above, obtain a new cluster membership value c^1 and related probabilities via CRP

1. for $i \leftarrow 1$ to n do
   - Remove the year y_i from cluster c_i to draw a new membership for it
   - If the previous step results in empty cluster, then remove the cluster and its corresponding parameters and rearrange the order of the clusters into contiguous $1, 2, \ldots, K$
   - Draw $c_i | c_{-i}, y$ from:
   for $k \leftarrow 1$ to $K + 1$ do
   Calculate the probability of $c_i = k$ using $p(c_i | c_{-i}, \mu_k, \Lambda_0, \Lambda_k, \alpha)$
   Calculate the probability of $c_i = k + 1$ using $p(c_i \neq c_k \forall i \neq j | c_{-i}, \mu_0, \Lambda_0, \alpha)$
   if $c_i = k$ for some $j \neq i$ then
   Update $\tilde{y}_k$, n_k and Λ_k based on $c_i = k$
   else
   $c_i = K + 1$, a new cluster is created.
   Append it to the vector of non-empty clusters
   end
   end
2. Retrieve $c^1 = c_i$ and the probabilities $p(c_i)$
3. Return $(c^1, p(c^1))$

▪ ALGORITHM 2: Blocked Gibbs Sampling

Inputs: Mortality Data Set ($MDS_{T \times d}$), Concentration parameter α, Distance matrices (D_T and D_I) for Covariance matrix calculation, Priors μ_0, Σ_0 and Σ_y (with predefined set of fixed parameters), initial cluster membership c^0 and the total number of iterations ($R = 10000$). In this setting, the first 2500 iterations are reserved for burn-in.

Output: Given the all parameters and the data set above, obtain the constructed mortality curves and their clusters

1. for $t \leftarrow 1$ to R do
   - Covariance matrices for years and ages: are calculated and stored for further computation
   - Apply **Algorithm 1** to get $(c^1, p(c^1))$ for cluster membership via CRP
   - Draw new atoms from multivariate normal with updated parameters , ie. $(\Theta_h | -) \sim \mathbf{N}(\tilde{\mu}_h, \tilde{\Sigma}_h)$
   - Update the covariance parameters $\kappa_1, \kappa_2, \kappa_3$ via MCMC
   - Update variance parameter via $(\sigma^{-2} | -) \sim Gamma(a_\sigma + \frac{nT}{2}, b_\sigma + \frac{1}{2} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_i(t) - \Theta_h)^2)$
   - Update concentration parameter via $(\alpha | -) \sim Gamma(a_\alpha + N - 1, b_\alpha + \frac{1}{2} \sum_{i=1}^{n} \sum_{t=1}^{T} \log(p(c_i)))$
   end
2. Return $(cluster_{out}, \theta_{out}, \kappa_{out}, \sigma^2_{out}, \alpha_{out})$

AN EXAMPLE

The age-at-death distribution across the years (2008-2017) for Germany can be observed in Figure 3.

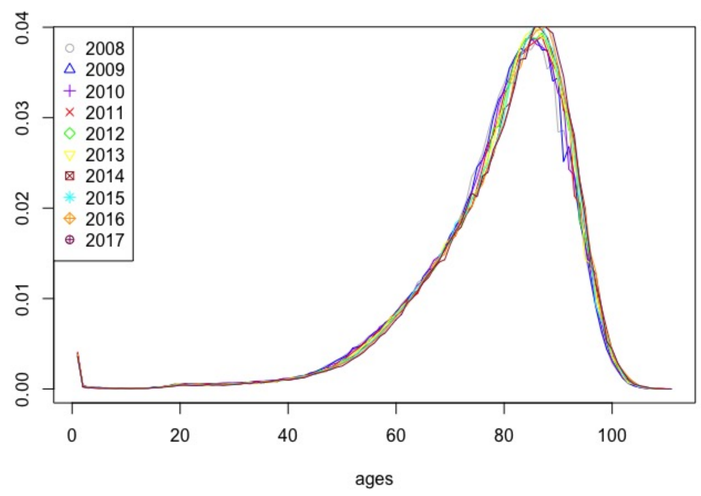

Figure 3. Mortality curves for Germany, 2008-2017

The mean representative (dashed red curve) of all the simulated mortality curves (gray curves), lying in one cluster, in Figure 4.

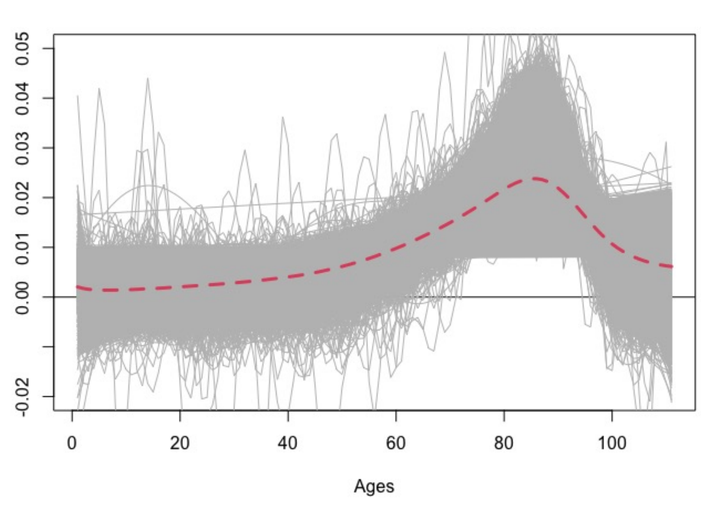

Figure 4. Simulated Mortality Curves for Germany

REFERENCES

Yuelin Lia, Elizabeth Schofield and Mithat Gönen (2019) A tutorial on Dirichlet Process mixture modeling. *Journal of Mathematical Psychology*, **91**, 128 – 144.

Sofia Curzio (2019) Analisi Funzionale di Curve di Mortalita con un Approccio Bayesiano non Parametrico *MSc Thesis*.