SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

SIS Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 1222·2022 ANNI

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

Copyright © 2022

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Contents

3 Solicited Sessions 385

4 Contributed Sessions 844

Preface

This book includes the contributions presented at the 51st Scientific Meeting of the Italian Statistical Society (SIS) held in Caserta at the Università della Campania "Luigi Vanvitelli", from the 22nd to 24th of June, 2022.

The conference has registered more than 300 presentations, including 4 keynotes in plenary invited sessions and 9 presentations in 3 guest sessions, 48 presentations collected in 16 specialized sessions and 68 presentations in 17 solicited sessions, all dealing with specific themes in methodological and/or applied statistics and demography. Furthermore, more than 200 contributions, with one or more authors, have been spontaneously submitted to the Program Committee and arranged in 43 contributed sessions.

The high number of contributions and the large participation at the conference show that researchers have met the challenge of pursuing working even in the face of the pandemic period from which we are only now emerging. The research activity in our field therefore has never stopped, and the desire to participate in scientific events, as a place for exchange and discussion on new developments in our field, remains a living characteristic of our scientific community.

With the publication of this book, we wish to offer to all members of the Italian Statistical Society, all international academics, researchers, Ph.D. students, and all interested practitioners, a good snapshot of the on-going research in the statistical and demographic fields. We deeply thank all contributors for having submitted their works to the conference and all the researchers for their remarkable job in acting as referees accurately and timely. We also would like to thank the International Biometric Society (IBS) – Italian region, the European Network for Business and Industrial Statistics (ENBIS) and the Italian Society of Statistical Physics (SIFS) we had the pleasure of hosting. A special thanks is addressed to the Scientific and Organizational Committees for their great efforts devoted to all the organizational aspects, to the Università della Campania "Luigi Vanvitelli" and to the Department of Mathematics and Physics who made this event possible, as well as to the Municipality of the Town of Caserta who has patronized the event and to all the funders for their supports.

Finally, we wish to express our gratitude to the publisher Pearson Italia for all the support received.

Bayesian mixtures of semi-Markov models

Misture Bayesiane di modelli semi-Markov

Rosario Barone and Andrea Tancredi

Abstract In this paper we propose a clustering technique for continuous-time semi-Markov models in order to take account of groups of individuals having similar process realizations. In fact fitting standard parametric models in presence of heterogeneity between population groups may produce biased inferences for relevant process feautres. To model individual heterogeneity we consider a Dirichlet process mixture (DPM) of semi-Markov continuous-time models. We also consider the case of discretely observed trajectories of continuous time processes, providing an algorithm which clusterize the observations after having reconstructed the continuous-time paths between the observed points. Full MCMC inference is performed with an application to a real dataset.

Abstract *In questo articolo proponiamo un metodo di clustering per modelli semi-Markov a tempo continuo al fine di tenere conto di gruppi di individui con realizzazioni del processo simili. Infatti, l'utilizzo di modelli parametrici standard in presenza di eterogeneità può produrre inferenze distorte. Per modellare l'eterogeneità individuale consideriamo una mistura di processi di Dirichlet (DPM) di modelli semi-Markov a tempo continuo. Consideriamo anche il caso di processi a tempo continuo osservati in punti discreti, presentando un algoritmo che consente di raggruppare le osservazioni dopo aver simulato le traiettorie tra i punti discreti. nel lavoro verrà presentato un algoritmo MCMC e un'applicazione con dati reali.*

Key words: Dirichlet process prior, Multi-state models, MCMC, Time series clustering.

Rosario Barone
University of Rome Tor Vergata, e-mail: rosario.barone@uniroma2.it.

Andrea Tancredi
Sapienza University of Rome, e-mail: andrea.tancredi@uniroma1.it.

1 Introduction

In a panel data set with individuals observed in time, clustering techniques may be useful for finding groups of similar individuals. With exception for the Markov case [1], all the mixtures approaches for multi-state models are related to finite mixtures and consider completely observed processes [2]. In this paper we tackle the clustering problem for the class of semi-Markov processes. In general, note that inference for semi-Markov models may present computational difficulties when observations are at discrete time points, so that the process is not completely observed. In fact, the likelihood function is not available and approximation methods are required. [3] reconstruct the likelihood function by simulating the trajectories between the observed points with a Metropolis-Hastings step based on Markovian proposals drawn from the uniformization algorithm [4]. Here, we extend their approach by considering a Bayesian nonparametric mixture model for both completely and discretely observed continuous time semi-Markov models. More specifically, by defining infinite mixtures of semi-Markov, with a Dirichlet process (DP) prior [5] on the mixing measure, we get a Dirichlet Process Mixture (DPM) [6] of semi-Markov models.

2 Semi-Markov multi-state models

Let us consider a continuous time process $Y(\cdot) = \{Y(t), t \geq 0\}$ with discrete state space $\mathscr{S} = \{1, \dots S\}$. We assume that the process $Y(\cdot)$ is semi-Markov. This is equivalent to say that the instantaneous transition rates $q_{rs}(t, \mathscr{F}_t)$, conditionally on the past history of the process, depend only on the time spent in the current state, i.e.

$$q_{rs}(t, \mathscr{F}_t) = \lim_{\delta t \to 0} \frac{P\{Y(t+\delta t) = s | X(t) = r, T^* = t - u\}}{\delta t}$$

where T^* denotes the entry time in the last state assumed before time t. Hence, semi-Markov models can be obtained by defining the transition functions $q_{rs}(u)$ and setting

$$P\{Y(t+\delta t) = s | Y(t) = r, T^* = t - u\} = \begin{cases} q_{rs}(u)\delta t + o(\delta t) & s \neq r \\ 1 - \sum_{l \neq r} q_{rl}(u)\delta t + o(\delta t) & s = r \end{cases}$$

Notice that a semi-Markov process $Y(t)$ can be also defined as the result of a state sequence generated by a Markov chain with transition probabilities p_{rs} and sojourn times having distribution functions F_{rs}, that is depending only on the departure and arrival states. To specify the functions $q_{rs}(u)$ we can also proceed directly by fixing the transition probabilities p_{rs} and the conditional sojourn distributions F_{rs}. By doing so, the resulting hazard functions turn out to be $q_{rs}(u) = p_{rs}F'_{rs}(u)/(1 - F_r(u))$ where $F_r(u) = \sum_{l \neq r} p_{rl} F_{rl}(u)$. Several parametric distributions can be proposed for $q_{rs}(u)$ or $F_{rs}(u)$. Assuming for example cause-specific hazards proportional to those of a distribution on positive values with parameters depending only on the initial state, i.e. $q_{rs}(u) = p_{rs}q(u;\phi_r)$, the transition probabilities are p_{rs}, and the density of trajectory y on the interval $[0,T]$ can be generally written as

$$p_\theta(y) = p_\theta(s,z) = \left(\prod_{i=1}^{n} p_{s_{i-1}s_i} q(z_i - z_{i-1};\phi_{s_{i-1}}) e^{-\int_0^{z_i - z_{i-1}} q(u;\phi_{s_{i-1}},)du}\right) \times e^{-\int_0^{T-z_n} q(u;\phi_{s_n})du},$$

where $z = (z_1,\ldots,z_n)$ is the sequence of jump times, $s = (s_1,\ldots,s_n)$ is the sequence of visited states and $\theta \in \Theta$ is the vector with all the process parameters.

3 Dirichlet Process Mixture of semi-Markov models

In this section we introduce the notation of the DPM model. Let $p_\theta(y)$ be the probability density function of a semi-Markov trajectory $y(t) = (s,z)$. Let G be a probability distribution defined on the parameter space Θ. We define the density function of an infinite mixture of semi-Markov models p_G with respect to the mixing measure G as

$$p_G(s,z) = \int p_\theta(s,z)dG(\theta).$$

By assuming a $DP(M,G_0)$ on the mixing measure G, we get a DPM of semi-Markov models. Let $y_i(t) = (s,z)_i$, for $i = 1,\ldots,N$, be N fully observed paths on $[0,T_i]$. Note that N represents the number of sample individuals. We may rewrite the model in a hierarchical form:

$$\begin{aligned} y_i(t)|\theta_i &\overset{ind}{\sim} p_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned}$$

4 MCMC sampling

We now outline the MCMC algorithm. Let $\mathbf{y} = (\mathbf{s},\mathbf{z})$ be the set of N observed trajectories of a continuous time multi-state model defined on the state space $\mathscr{S}$. Let θ_h^*, $h = 1,\ldots,k$ denote the $k \leq N$ unique values among the N individual parameters and let $\Psi_h = (i : \theta_i = \theta_h^*)$. An important property of the Dirichlet process is in fact to produce discrete distributions for G so that the trajectories y_i are clustered on the base of the unique values generated by G. We represent the clustering by an equivalent set of cluster membership indicators, $\psi_i = h$ if $i \in \Psi_h$, i.e. the i-th observation belongs to the h-th cluster. Let $\mathbf{y}_h^*$ indicate the trajectories arranged by clusters, i.e. the set of trajectories belonging to the h-th cluster. Let the multiset $\rho_n = \{\Psi_1,\ldots,\Psi_k\}$ be a random partition of the trajectories $\{1,\ldots,N\}$. The DPM implies a model on a random partition ρ_n of the experimental units. Let $\pi(\rho_n)$ be the Pólya urn prior on the random partition ρ_n. The posterior conditional densities $\pi(\psi_i = h|\psi_{-i}, \mathbf{y})$ are derived as follows:

$$\pi(\psi_i = h|\psi_{-i}, \mathbf{y}) \propto \begin{cases} N_h^- p((s,z)_i|\mathbf{y}_h^{*-}) & \text{for } h = 1,\dots,k^- \\ M p((s,z)_i) & \text{for } h = k^- + 1 \end{cases},$$

where N_h^- is the number of paths in the h-th cluster with exclusion of the i-th observation, M represents the precision parameter of the DP, k^- indicates the unique values of θ with exclusion of the i-th observation, $p((s,z)_i) \equiv \int p_{\theta_h^*}((s,z)_i) G_0(d\theta_h^*)$ and $p((s,z)_i|\mathbf{y}_h^{*-}) = \int p_{\theta_h^*}((s,z)_i) d\pi(\theta_h^*|\mathbf{y}_h^{*-})$. However, with the semi-Markov kernel density there is not conjugate centring measure G_0. Therefore, approximation methods are required in order to evaluate the integrals.
Since there is no conjugacy between the base measure G_0 and the semi-Markov kernel density, we adopt the approach of [7], by defining a valid Markov chain algorithm by drawing $(N-k)$ auxiliary values for θ from the centring measure G_0. More in detail, if $N_h > 1$, the posterior model on the random partition ρ_n is defined as:

$$\pi(\psi_i = h|\psi_{-i}, \mathbf{y}) \propto \begin{cases} N_h^- p_{\theta_h^*}((s,z)_i) & \text{for } h = 1,\dots,k^- \\ \frac{M}{k^-+1} p_{\theta_{k^-+1}^*}((s,z)_i) & \text{for } h = k^- + 1 \end{cases}, \tag{1}$$

while if $N_h = 1$, $\psi_i = h$ is imputed to form a singleton cluster. Therefore, with probability $(k-1)/k$ leave ψ_i unchanged, otherwise remove ψ_i from the h-th cluster, relabel the θ_h^* to comply with the no-gaps rule, and then update ψ_i. By considering $y_i = (s,z)_i$, $i = 1,\dots,N_h$, the density of the data lying inside the h-th mixture component conditional on the partition and on the kernel parameters is

$$\pi(\mathbf{y}|\theta_h, \psi) = \prod_{i=1}^{N_h} p_{\theta_h}((s,z)_i|\psi_i = h),$$

by Bayes theorem we get the posterior density of the h-th mixture component parameters:

$$\pi(\theta_h|\psi, \mathbf{y}) \propto \pi(\mathbf{y}|\theta_h, \psi) \cdot G_0(\theta_h). \tag{2}$$

4.1 Discretely observed semi-Markov processes

We now extend the DPM of semi-Markov models to the the case of discretely observed trajectories, that is when the exact jump times are unknown and the density function $p_\theta(y)$ is not available. In particular we use the algorithm proposed by [3], which allows to reconstruct the trajectories between the discretely observed points for each observed individual via a Metropolis-Hastings step based on a Markovian approximation of the semi-Markov process. By indicating with $Q(s,z|\theta,x)$ the conditional distribution of a whole sample path conditionally on an observed set of points summarized by the vector x, the MCMC algorithm for the discretely observed case may be summarized as follows:

- *Trajectory reconstruction*: for $i = 1\ldots,N$ draw $(s,z)_i \sim Q(s,z|\theta_i,x_i)$;
- *Clustering*: for $i = 1\ldots,N$ draw ψ_i from (1);
- *Updating cluster parameters*: for $h = 1,\ldots,k$ draw θ_h from (2).

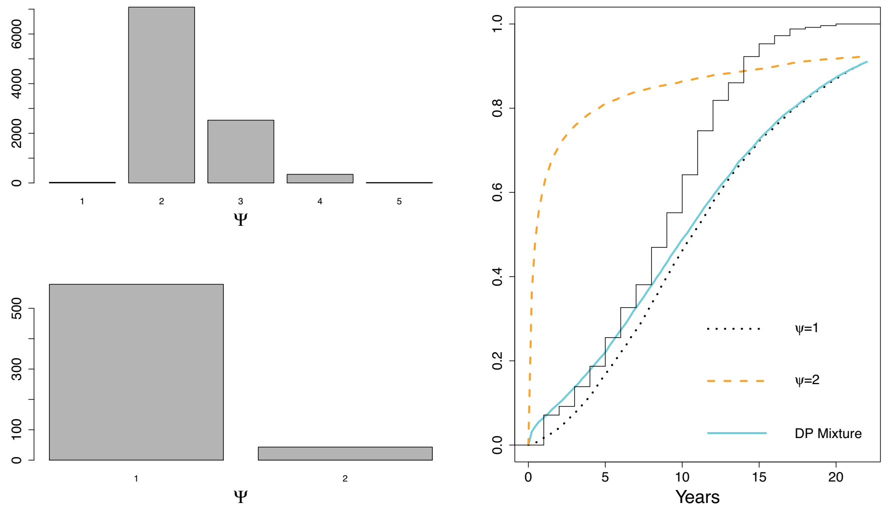
5 Application

As a real data application, we analyze the progression of coronary allograft vasculopathy (CAV) [8] with a data set available with the R package `msm`, see [9]. The data provides the disease status (CAV-free (1), mild CAV (2) and moderate or severe CAV (3)) observed approximately each year after transplant for a set of 622 subjects followed up until their most recent visit if alive at the end of the observation period or until death (state (4)). Death times are exactly observed. Moreover, since data comprises apparent transitions from higher to lower states, which are in fact the results of a misclassification since the deterioration of the arterial walls is an irreversible process, we recode all the reverse transitions as remaining in the higher of the two states and permit transitions only to the adjacent states or to the death state. To specify the semi-Markov model we assume Weibull sojourn times. The parameter set is $\theta = (p,\gamma,\alpha)$ where γ and α are the vector with the rate and shape parameters of the Weibull sojourn times and p is the matrix with the transition probabilities. For the Dirichlet process we chose a precision parameter $M = 1$ and defined the centering measure to be the product between Dirichlet distributions for the rows of p, Gamma distributions for the rate parameters and log Normal distributions for the sahpe parameters, that is $G_0 \equiv \mathrm{Dir}_{\mathscr{S}-1} \times \mathrm{Gamma}_{\mathscr{S}-1} \times \log \mathrm{N}_{\mathscr{S}-1}$. We run the MCMC sampler for 10000 iterations with a burnin of 2000. The model gives evidence of the existence of two groups of observations with different features inside the sample. In Table 1 and Figure 1 we show the results by reporting some posterior summaries for the proposed model.

Table 1 CAV data: DPM of semi-Markov.

	ψ	γ_1	γ_2	γ_3	α_1	α_2	α_3	p_{12}	p_{14}	p_{23}	p_{24}
$E(\cdot|Y,\psi)$	1	0.13	0.24	0.25	1,47	1.46	1.10	0.71	0.29	0.72	0.28
$SD(\cdot|Y,\psi)$	1	0.01	0.02	0.03	0.10	0.18	0.12	0.04	0.04	0.08	0.08
$q_{0.025}(\cdot|Y,\psi)$	1	0.12	0.19	0.20	1.30	1.16	0.88	0.64	0.21	0.56	0.12
$q_{0.975}(\cdot|Y,\psi)$	1	0.14	0.29	0.30	1.65	1.85	1.34	0.79	0.36	0.88	0.44
$E(\cdot|Y,\psi)$	2	6.27	1.42	1.30	0.66	0.89	4.45	0.71	0.29	0.16	0.84
$SD(\cdot|Y,\psi)$	2	3.62	1.45	1.01	0.19	1.32	4.45	0.36	0.36	0.23	0.23
$q_{0.025}(\cdot|Y,\psi)$	2	0.16	0.02	0.08	0.39	0.17	0.37	0.02	0.00	0.00	0.13
$q_{0.975}(\cdot|Y,\psi)$	2	13.08	4.30	3.81	1.10	4.95	15.85	1.00	0.98	0.87	1.00

Fig. 1 CAV data: on the top left panel the maximum number of mixture components observed for each iteration and on the bottom left the distribution of the observations across the estimated components; on the right panel the cumulative posterior predictive distributions of the death time.



References

1. Luo, Yu, David A. Stephens, and David L. Buckeridge. Bayesian clustering for continuous-time hidden Markov models. *Canadian Journal of Statistics*, 2021.
2. Sylvia Frühwirth-Schnatter and Christoph Pamminger. Model-based clustering of categorical time series. *Bayesian Analysis*, 2010
3. Rosario Barone and Andrea Tancredi. Bayesian inference for discretely observed continuous time multi-state models. *preprint arXiv: 2202.12447*, 2022.
4. Asger Hobolth and Eric A Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204–1231, 2009.
5. Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.
6. Albert Y Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357, 1984.
7. Steven N MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.
8. Linda D Sharples, Christopher H Jackson, Jayan Parameshwar, John Wallwork, and Stephen R Large. Diagnostic accuracy of coronary angiography and risk factors for post–heart-transplant cardiac allograft vasculopathy. *Transplantation*, 76(4):679–682, 2003.
9. Christopher H Jackson. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011.