Bayesian Analysis (2022) **TBA**, Number TBA, pp. 1–23

Seemingly Unrelated Multi-State Processes: A Bayesian Semiparametric Approach*

Andrea Cremaschi[†] Raffaele Argiento[‡] Maria De Iorio[§,†,¶] Cai Shirong[†] Yap Seng Chong[§,†] Michael Meaney[†,§,‖] and Michelle Kee[†]

Abstract. Many applications in medical statistics and other fields can be described by transitions between multiple states (e.g. from health to disease) experienced by individuals over time. In this context, multi-state models are a popular statistical technique, in particular when the exact transition times are not observed. The key quantities of interest are the transition rates, capturing the instantaneous risk of moving from one state to another. The main contribution of this work is to propose a joint semiparametric model for several possibly related multi-state processes (Seemingly Unrelated Multi-State, SUMS, processes), assuming a Markov structure for the transitions over time. The dependence between different processes is captured by specifying a joint prior distribution on the transition rates of each process. In this case, we assume a flexible distribution, which allows for clustering of the individuals, overdispersion and outliers. Moreover, we employ a graph structure to describe the dependence among processes, exploiting tools from the Gaussian Graphical model literature. It is also possible to include covariate effects. We use our approach to model disease progression in mental health. Posterior inference is performed through a specially devised MCMC algorithm.

Keywords: Multi-State Models, Normalised Point Processes, Graphical Models, Mixture Models, Markov Chain Monte Carlo.

1 Introduction

Biomedical data are characterised by a high number of different variables, in many cases mostly categorical and recorded on a (nowadays often large) set of subjects. This is mainly due to the practice in clinical settings to record the absence/presence of symptoms and/or to use ordinal scales to represent disease markers. Typically, we only observe clinical variables at fixed time points (usually corresponding to follow up or

*The GUSTO research is supported by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Programme and administered by the Singapore Ministry of Health's National Medical Research Council (NMRC), Singapore – NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014. Additional funding is provided by the Singapore Institute for Clinical Sciences, Agency for Science Technology and Research (A*STAR). This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2019-T2-2-100. Michael J. Meaney is supported by funding from the JPB Research Foundation and the Jacob's Foundation. Dr. Argiento is grateful to A*STAR, Singapore for the funding provided.

[†]Singapore Institute for Clinical Sciences, A*STAR, Singapore, cremaschia@sics.a-star.edu.sg
[‡]Department of Economics, Universitá di Bergamo, Bergamo, Italy
[§]Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[¶]Department of Statistical Science,University College London, UK
[‖]Department of Psychiatry, McGill University, Montreal, Canada

 https://doi.org/10.1214/22-BA1326

hospital visits), and as such these variables are interval-censored (i.e., panel data). The objective of clinical studies is often to model a patient's *disease progression*, as captured by the evolution over time of one or more responses of interest, e.g. representing the disease status and associated clinical markers. A popular approach is to use multi-state models describing the transition of individuals among multiple states in continuous time (see, for instance, Cook, 1999; Sung et al., 2007; Jackson et al., 2011; van den Hout et al., 2015; De Iorio et al., 2018). In this framework, it is straightforward to include time-homogeneous covariates and time varying continuous covariates (leading to a Markov regression model).

In this work, we propose a joint modelling approach for several categorical variables evolving simultaneously through time. More in details, our approach is based on a combination of ideas from multi-state models, seemingly unrelated regression (SUR) (Zellner, 1963; Wang, 2010), Gaussian Graphical models (GGM) and Product Partition Models with Covariates (PPMx) (Müller et al., 2011). In a Bayesian framework, we define a joint model for several multi-state processes, which represent the evolution of, for instance, clinical markers of interest as in the disease progression application of Section 3. The main idea is that the different markers provide complementary information on the underlying health status and, as such, they are regarded as stochastic processes defined on a finite state-space evolving in continuous time according to *dependent* Markov processes. We link the different Markov processes through the specification of a flexible prior distribution on the instantaneous transition rates, specifically a mixture distribution with random number of components (Argiento and De Iorio, 2019). In this way, we build a robust modelling strategy, which leads to covariate-driven clustering of the subjects and enables the inclusion of different types of covariates/responses in a natural and efficient way (Barcella et al., 2017). Each multi-state process is then, conditionally on the vector of instantaneous transition rates, independent from the other processes, resembling the seemingly unrelated regressions setting of Zellner (1963). Furthermore, we allow the dependence structure between the transition rates to be encoded into a random graph, which is also object of posterior inference, as it is done in sparse SUR models (SSUR, Wang, 2010). Thus, the nature of the dependence is learnt from the data, spanning from independence to full inter-dependence. As such, we refer to our model as *Seemingly Unrelated Multi-State* (SUMS) processes. Briefly, the proposed model allows for: (i) multiple responses; (ii) processes with more than two states; (iii) patient- and process-specific times of observation; (iv) inclusion of mixed-type covariates; (v) covariate-driven clustering of the subjects; (vi) missing initial state information.

One of the main advantages of our modelling strategy is that the relationship between different multi-state processes is encoded into a graph structure. Indeed, if there is an edge linking two processes, it means that they are conditionally dependent, while the absence of an edge implies conditional independence. This gives insight into the co-regulatory mechanisms of the different processes. This is relevant in many applications as often it is also of interest to identify important factors affecting disease progression, for better prognosis and therapeutic choices. Moreover, the model allows for the inclusion of time-homogeneous covariates (of any type) and time-varying continuous covariates in a regression component, for which standard variable selection techniques (e.g. shrinkage, spike and slab priors) can be employed.

The manuscript is organised as follows: Section 2 introduces the SUMS model, by presenting how its key components – the joint multi-state model, the mixture prior with unknown number of components and the graphical structure describing the dependence among processes – interrelate, as well as the specifically designed Markov Chain Monte Carlo (MCMC) algorithm. Section 3 presents an application of the model to the analysis of mental health indicators obtained from the GUSTO cohort study. Section 4 concludes the work. In Supplementary Material (Cremaschi et al., 2022), we include a detailed description of the algorithm and of the GUSTO dataset, a sensitivity analysis and a simulation study, as well as further results from the analysis of the GUSTO data.

2 SUMS: Seemingly Unrelated Multi-State Processes

2.1 Modelling of Multi-State Processes

Multi-state models can be used to describe how an individual moves between a set of states in continuous time. In this work, we focus on multi-state processes for panel data, where the states of several processes are observed only at certain time points, and their exact transition times are not known. For each $h = 1, \dots, p$, let $\{Y^{(h)}(t), t \in \mathbb{R}^+\}$ be a continuous time Markov process, where $Y^{(h)}(t)$ represents the state of the h-th process over time, with state-space $\mathcal{S}^{(h)} = \{1, \dots, d^{(h)}\}$ of dimension $d^{(h)}$, i.e. $Y^{(h)}(t) \in \mathcal{S}^{(h)}$. The elements of $\mathcal{S}^{(h)}$ represent the states that the h-th process can visit between transitions. The exact times of transition of the processes $Y^{(h)}(t)$ are not known, but in applications the processes are observed on a discrete set of time points, $\boldsymbol{t}_i^{(h)} = (t_{i1}^{(h)}, \dots, t_{in_i^{(h)}}^{(h)})$, where $n_i^{(h)}$ denotes the number of observed time points for the i-th individual and h-th process. Notice that the times of observation and their number are both process- and subject-specific. We indicate with $Y_{ij}^{(h)} = Y^{(h)}(t_{ij}^{(h)})$ the value of the h-th process $Y^{(h)}(t)$ at the j-th observed time $t_{ij}^{(h)}$ for the i-th subject. Hence, for each subject $i = 1, \dots, N$, we observe the random vector $\boldsymbol{Y}_i^{(h)} = (Y_{i1}^{(h)}, \dots, Y_{in_i^{(h)}}^{(h)})$, corresponding to the h-process and whose distribution is the finite-dimensional law of the process $Y^{(h)}(t)$ at the times of observation. The probability law of the h-th process $Y^{(h)}(t)$ is assigned via the matrix of instantaneous transition rates $\boldsymbol{Q}^{(h)} = [\lambda^{(h)}(r,s)]_{r,s}$. The instantaneous transition rates $\lambda^{(h)}(r,s)$, with $r, s \in \mathcal{S}^{(h)}$, represent the instantaneous risk of moving from one state to the other. The vector $\boldsymbol{\lambda}^{(h)} = \{\lambda^{(h)}(r,s) : r, s \in \mathcal{S}^{(h)}, r \neq s\}$, of dimension $d^{(h)}\left(d^{(h)} - 1\right)$, corresponds to the off-diagonal transition rates of the matrix $\boldsymbol{Q}^{(h)}$, concatenated by row from top to bottom.

Let $\epsilon_{ij}^{(h)} = t_{ij}^{(h)} - t_{ij-1}^{(h)}$ indicate the length of the j-th time interval, for $j = 2, \dots, n_i^{(h)}$, subject i and process h. Chapman-Kolmogorov equations can be solved to obtain the transition probabilities $\boldsymbol{p}_{ij}^{(h)}(\boldsymbol{\lambda}^{(h)}, \epsilon_{ij}^{(h)}) = \{p_{ij}^{(h)}(r, s; \boldsymbol{\lambda}^{(h)}, \epsilon_{ij}^{(h)}) : r, s \in \mathcal{S}^{(h)}\}$, where $\boldsymbol{\lambda}^{(h)}$ denotes the vector of transition rates for the h-process (see Ross et al., 1996). When $d^{(h)} = 2$, closed-form solutions are readily available (Cox and Miller, 1977), while problems involving more than three states are usually tackled numerically (Moler and Van Loan, 2003). It can be shown that for each process h a unique stationary distribution exists (Grimmet and Stirzaker, 2001). The stationary distribution can be used

as marginal distribution for modelling the state of the processes at time $j = 1$, given the vectors of instantaneous transition rates $\boldsymbol{\lambda}^{(h)}$, in contrast to the general practice in multi-state modelling of specifying the model conditionally on the initial state. This is important, as it allows Bayesian imputation of missing observations at time one, since they are treated as unknown parameters in the model. This aspect is particularly useful in our application, where the initial time presents a non-negligible missing rate. For each process, we assume that the Markov property holds, i.e. conditionally on current and past events, future transitions only depend on the current state. Moreover, we assume that the transition rates are also subject specific and can vary with time, as discussed in Section 2.2. To highlight time and subject dependence, we use the notation $\boldsymbol{\lambda}_{ij}^{(h)}$. This implies that the transition probabilities $\boldsymbol{p}_{ij}^{(h)}(\boldsymbol{\lambda}_{ij}^{(h)}, \epsilon_{ij}^{(h)})$ as well as the stationary distribution $\pi_{ij}^{(h)}(k; \boldsymbol{\lambda}_{ij}^{(h)})$, with $k \in \mathcal{S}^{(h)}$, are functions of the time- and subject-specific instantaneous transition rates (see Ross et al., 1996, for details) and the law of $\boldsymbol{Y}_i^{(h)}$ is given by:

$$\prod_{j=2}^{n_i^{(h)}} \left(p_{ij}^{(h)} \left(Y_{ij-1}^{(h)}, Y_{ij}^{(h)}; \boldsymbol{\lambda}_{ij}^{(h)}, \epsilon_{ij}^{(h)} \right) \right) \pi_{i1}^{(h)} \left(Y_{i1}^{(h)}; \boldsymbol{\lambda}_{i1}^{(h)} \right).$$

The aim of this work is to jointly model the processes $Y^{(h)}(t)$, capturing their time evolution and possible dependencies. In what follows, for simplicity, we indicate a transition of the h-th process between different states of $\mathcal{S}^{(h)}$ with the notation $r \to s$. The joint likelihood for the vector of observed states $\boldsymbol{Y}_i^{(h)}$, for $i = 1, \ldots, N$ and $h = 1, \ldots, p$, is then:

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\lambda}^{Y}\right) = \prod_{i=1}^{N} \prod_{h=1}^{p} \prod_{j=2}^{n_i^{(h)}} \left(p_{ij}^{(h)} \left(Y_{ij-1}^{(h)}, Y_{ij}^{(h)}; \boldsymbol{\lambda}_{ij}^{(h)}, \epsilon_{ij}^{(h)} \right) \right) \pi_{i1}^{(h)} \left(Y_{i1}^{(h)}; \boldsymbol{\lambda}_{i1}^{(h)} \right), \qquad (1)$$

where $\boldsymbol{Y}$ and $\boldsymbol{\lambda}^{Y}$ indicate the multi-dimensional arrays containing the observation vectors $\boldsymbol{Y}_i^{(h)}$ and the instantaneous transition rate vectors $\boldsymbol{\lambda}_{ij}^{(h)}$, while $p_{ij}^{(h)}$ denotes the transition probabilities and $\pi_{i1}^{(h)}$ is the (stationary) distribution at time one. Dependence across processes is captured through the specification of a joint distribution on the vectors of transition intensities $\boldsymbol{\lambda}_{ij}^{(h)}$.

2.2 Model for Transition Intensities

The instantaneous transition rates $\lambda_{ij}^{(h)}(r, s)$ can be made covariate-dependent by specifying a Cox proportional hazard model. This allows the inclusion of both time-homogeneous covariates as well as time-varying continuous covariates. Alternatively, a semi-proportional intensity model can be easily specified for the covariates as in Kim et al. (2012). Note that the decision of including either type of covariates is process-specific. The time-homogeneous covariates are straightforwardly incorporated in the model, and we denote them here by $\boldsymbol{X}_i^{(h)} = (X_{i1}^{(h)}, \ldots, X_{ig^{(h)}}^{(h)})$, for the i-th individual and h-th process. On the other hand, the time-varying continuous ones, denoted by $\boldsymbol{Z}_i^{(h)}(t) =$

$(Z_{i1}^{(h)}(t), \ldots, Z_{iq^{(h)}}^{(h)}(t))$, require additional assumptions. They are usually included via a piece-wise constant effect over each interval of observations (Andersen et al., 2012), or by modelling them as longitudinal processes, linking their distribution to the ones of the multi-state processes via the inclusion of suitable random effects (Ferrer et al., 2016). The first option has a clear computational advantage, while the latter has the potential to yield better inference on the overall disease progression. The code provided with this manuscript allows for the implementation of the first method. This assumption leads to a piecewise constant model for the instantaneous transition rates $\lambda_{ij}^{(h)}(r,s)$, with $r, s \in \mathcal{S}^{(h)}$, and consequently for the matrix of transition intensities. The model for the instantaneous log-transition rates is then:

$$\log\left(\lambda_{ij}^{(h)}(r,s)\right) = \phi_i^{(h)}(r,s) + \boldsymbol{X}_i^{(h)}\boldsymbol{\beta}_{rs}^{(h)} + \boldsymbol{Z}_{ij}^{(h)}\boldsymbol{\gamma}_{rs}^{(h)}, \quad j = 1, \ldots, n_i^{(h)}, h = 1, \ldots, p, \quad (2)$$

where $\phi_i^{(h)}(r,s)$ represents the baseline transition rate (on log scale) of transition $r \to s$. The parameters $\boldsymbol{\beta}_{rs}^{(h)} \in \mathbb{R}^{g^{(h)}}$ and $\boldsymbol{\gamma}_{rs}^{(h)} \in \mathbb{R}^{q^{(h)}}$ are the vectors of regression coefficients for the h-th process and the $r \to s$ transition.

Let $\boldsymbol{\phi}_i^{(h)} = \{\phi_i^{(h)}(r,s) : r \to s\}$ be the vector of baseline log-transition rates for process h and subject i. A key component of our modelling strategy is the specification of the distribution of the vector $\boldsymbol{\phi}_i = (\boldsymbol{\phi}_i^{(1)}, \ldots, \boldsymbol{\phi}_i^{(p)})$, containing the baseline log-transition rates of all the p processes for each subject i. To this end, we borrow ideas from the SUR framework of Zellner (1963), where p different regression models are linked by specifying a joint error distribution, usually multivariate normal. The SUR methodology is one of the main techniques for handling multiple responses and offers a way to share information between models which are *seemingly* unrelated, since they describe different data-generating processes. However, since these are observed for the same set of subjects and measurements are taken on often related processes, the study of their interdependency is of great interest in most applications. For this reason SUR-type models have gained vast popularity in different fields, such as Phenomics (Houle et al., 2010; Banterle et al., 2018). In our application, for instance, where we deal with several processes associated to different aspects of maternal mental health (depression, anxiety, sleep quality), it is important to understand the relationships between such processes in order to have a comprehensive view of the phenomenon under study. As in the SUR framework, in our context each process is modelled by its own *seemingly unrelated* multi-state Markov process, but then they are *related* through the joint prior distribution on $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)$. Motivated by this parallelism, we name the proposed model Seemingly Unrelated Multi-State (SUMS) processes.

To capture the inter-individual heterogeneity and allow for clustering of the subjects, we choose as prior distribution for $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ a mixture prior with random number of components, where the distribution of the weights is given by the normalisation of a finite point process, as proposed by Argiento and De Iorio (2019). Finite mixture models with random number of components have been extensively studied in the literature, with a particular interest towards computational aspects (Richardson and Green, 1997; Stephens, 2000; Nobile, 2004) as they involve trans-dimensional moves. Recently, Miller and Harrison (2018) discussed the analogies between a finite mixture with random

number of components and a Dirichlet process mixture model (Lo, 1984), proposing a marginal algorithm for posterior inference for a model equipped with Dirichlet weights. The approach adopted in this paper and firstly proposed by Argiento and De Iorio (2019) is more general, as it allows the specification of different distributions for the unnormalised weights as well as a conditional algorithm to speed up inference. Argiento and De Iorio (2019) show that a finite mixture model is simply a realisation of a stochastic process whose dimension is random and has an infinite dimensional support. This leads to flexible distributions for the weights of the mixture, which is given by the normalisation of a finite point process. They refer to their construction as *normalised independent finite point process*. This approach has several advantages, allowing for flexible modelling of the weights in the mixture as well as efficient posterior computations (e.g. as compared to traditional reversible jump algorithms for mixture models). In more details, we assume the following mixture prior:

$$\begin{aligned} \boldsymbol{\phi}_i &= \boldsymbol{\phi}^\star_{c_i}, \quad i = 1, \dots, N, \\ \boldsymbol{\phi}^\star_1, \dots, \boldsymbol{\phi}^\star_M \mid M &\overset{\text{iid}}{\sim} P_0\left(\boldsymbol{\phi}^\star \mid \boldsymbol{\theta}\right), \\ \mathbb{P}(c_i = m) &\propto S_m, \quad i = 1, \dots, N, \\ S_1, \dots, S_M \mid M &\overset{\text{iid}}{\sim} \text{Gamma}\left(\gamma_S, 1\right), \\ M &\sim \text{Poi}_1\left(\Lambda\right), \end{aligned} \tag{3}$$

where we denote by $\text{Gamma}(a, b)$ the Gamma distribution with mean a/b, and by $\text{Poi}_1(\Lambda)$ the Poisson distribution shifted by one unit with parameter Λ. Note that in the work of Argiento and De Iorio (2019) different distributions for the unnormalised weights $\boldsymbol{S} = (S_1, \dots, S_M)$ are considered. Here, we opt for the Gamma distribution as it leads to the standard mixture model with Dirichlet weights. The variables $\boldsymbol{c} = (c_1, \dots, c_N)$ indicate the component allocation of the subjects and their corresponding prior probabilities are proportional to the unnormalised weights $\boldsymbol{S}$. Finally, the vectors $\boldsymbol{\phi}^\star_1, \dots, \boldsymbol{\phi}^\star_M$ are a finite sequence of locations for the mixture distribution and are, conditionally on the number of components M, i.i.d. from the base measure P_0. As shown by Argiento and De Iorio (2019), posterior computations in this setting are greatly simplified via the introduction of a latent variable, conditionally on which the unnormalised weights of the mixture in (3) become independent. The specification of a joint prior distribution for $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N$ in model (3) and the choice of P_0 are crucial in our modelling strategy, as it will be shown in Section 2.3, since this allows inference on the shared dependence structure among the components of the vectors $\boldsymbol{\phi}^\star_m$, for $m = 1, \dots, M$ and, consequently, on the dependence structure among the p different processes. As an alternative, we could have opted for a Bayesian nonparametric prior, such as the Dirichlet process (De Iorio et al., 2018) and the beta-Dirichlet process prior (Kim et al., 2012), or, taking a complete different approach, flexible modelling of the baseline transition intensities can be achieved using penalised splines (Kneib and Hennerfeind, 2008).

2.3 Gaussian Graphical model

We use tools from the Gaussian Graphical models literature to describe the dependence among the p processes. Referring to model (3), we assume that $\boldsymbol{\phi}^\star_1, \dots, \boldsymbol{\phi}^\star_M \mid M \overset{\text{iid}}{\sim} P_0 =$

$\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}_G)$, where the key modelling feature is the specification of the prior on the precision matrix $\boldsymbol{\Omega}_G$ conditional on a graph G, which captures the conditional dependence structure among the baseline log-transition rates. The novelty of our modelling strategy is that G is modelled conditionally on another random graph G_0, which describes the dependence structure among processes and is one of the main objects of our inference.

In details, consider the graph $G_0 = (V_0, E_0)$, defined over the set of nodes $V_0 = \{1, \ldots, p\}$, i.e. each node in the graph corresponds to a multi-state process $Y^{(h)}(t)$. The edge set E_0 is formed of the pairs $E_0 \subseteq \{(h,k) \in V_0 \times V_0 : h < k\}$. We consider only simple graphs, i.e. undirected graphs, without self-loops nor multiple edges. It is important to highlight that G_0 describes the dependence structure at process level (but not at observation level), which then needs to be translated in dependence among the components of the vector $\boldsymbol{\phi}^{\star}_m = (\boldsymbol{\phi}^{\star,(1)}_m, \ldots, \boldsymbol{\phi}^{\star,(p)}_m)$, with $\boldsymbol{\phi}^{\star,(h)}_m = \{\phi^{\star,(h)}_m(r,s) : r \to s; r, s \in \mathcal{S}^{(h)}\}$ for $h = 1, \ldots, p$. To this end, we define a second graph G whose structure is determined by G_0. In particular, we let $G = (V, E)$ be the graph whose nodes are the indices of the vector $\boldsymbol{\phi}^{\star}_m$, i.e. $V = \{1, \ldots, D_p\}$, with $D_p = \sum_{h \in V_0} d^{(h)}(d^{(h)} - 1)$. G is a deterministic function of G_0 specified as follows. First, we assume there exists an edge in G between transition rates of the same process, i.e. between the components of $\boldsymbol{\phi}^{\star,(h)}_m$. Therefore, an empty graph G_0 (with no edges) corresponds to a graph G with p cliques, one for each process. Second, if there is an edge between nodes h and k in G_0 (i.e., $(h,k) \in E_0$), then there is and edge between all the possible pairs formed by an element of $\boldsymbol{\phi}^{\star,(h)}_m$ and one of $\boldsymbol{\phi}^{\star,(k)}_m$. An illustration for the case of three processes, which can assume only two possible states, is given in Figure 1. We write $G = f(G_0)$, f being the *deterministic* transformation described above. Note that f is bijective and, as such, the specification of a prior on G_0 implies a prior on G. This construction is advantageous in terms of dimension reduction, as the dimension of the graph space where G_0 is defined can be significantly smaller than the one of G, leading to more efficient exploration of the posterior space. Note that G_0 has p nodes, while G has D_p nodes.

Following the literature on GGMs, the conditional independence structure of the multivariate Gaussian vectors $\boldsymbol{\phi}^{\star}_m \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}_G)$, for $m = 1, \ldots, M$, is described by constraining the elements of the precision matrix $\boldsymbol{\Omega}_G$ (Dempster, 1972). Namely, two elements of the vector $\boldsymbol{\phi}^{\star}_m$ are, conditionally on the others, independent if and only if there is a zero in the corresponding entry of the precision matrix $\boldsymbol{\Omega}_G$. Since G is a deterministic function of G_0, it is the latter that encodes the conditional independence structure of the vectors $\boldsymbol{\phi}^{\star}_m$, for $m = 1, \ldots, M$ (see Figure 1), i.e. an edge in G_0 implies a set of edges/cliques in G as described above. The standard conjugate prior for the precision matrix $\boldsymbol{\Omega}_G$ is the G-Wishart distribution, specified conditionally on the graph structure G (Roverato, 2002). The last component needed to fully specify this part of the model is the prior distribution for the graph G, which is simply implied, through the bijection f, by the prior on the graph G_0:

$$\pi(G_0 \mid \eta) \propto \eta^{|E_0|} (1-\eta)^{\binom{p}{2} - |E_0|}, \quad \eta \in (0,1),$$

where $|E_0|$ is the number of edges in graph G_0 (i.e., the size of E_0), while $\binom{p}{2}$ is the number of possible graphs with nodes $V_0 = \{1, \ldots, p\}$. This prior is equivalent to assuming a Bernoulli prior with probability of success (here inclusion) η on each edge

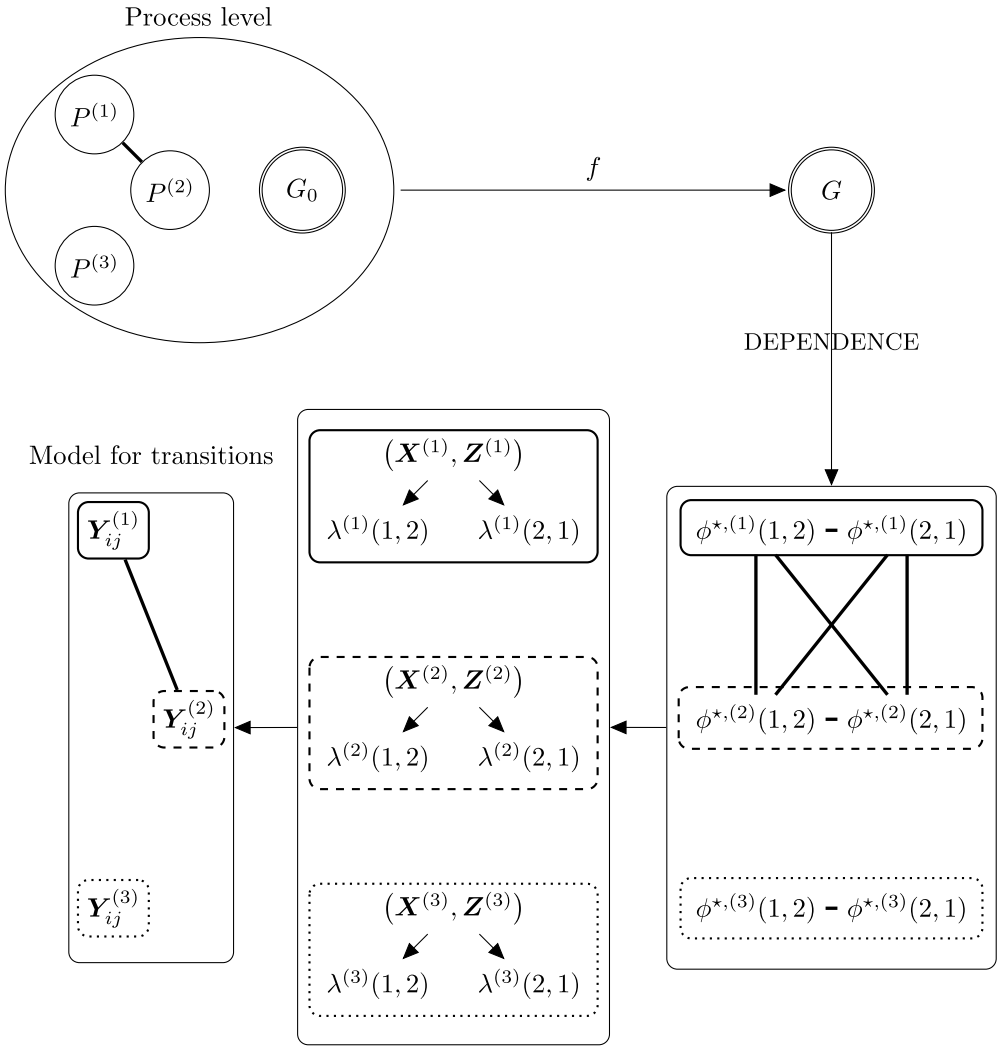

Figure 1: Example with three processes, which can assume only two possible states $\{1,2\}$. The graph G_0 describes the conditional independence between the processes in $V_0 = \{P^{(1)}, P^{(2)}, P^{(3)}\}$, with $P^{(1)}$ and $P^{(2)}$ conditionally dependent. The graph G is obtained as a deterministic function f from G_0. The graph G models the dependence among the baseline log-transition rates $\boldsymbol{\phi}^\star$ (solid lines). These, together with available covariates, contribute to determine the distribution of the transition rates $\boldsymbol{\lambda}$. Finally, the $\boldsymbol{\lambda}$s determine the transition probabilities of the three processes, which govern the model for the observed transitions between states.

of the graph G_0, independently across edges. Small values of η favour sparser graphs (Armstrong et al., 2009). Finally, we point out that, while the prior for the graph G_0 is defined over all possible graphs, including the non-decomposable ones, the resulting prior distribution on G is defined on a restricted space due to the clique constraints imposed on the transitions of the same process which need to be fully connected.

2.4 Full Model Specification

Here we combine the strategies described above and present the full SUMS model:

$$
\begin{aligned}
&\boldsymbol{Y}_i^{(1)}, \ldots, \boldsymbol{Y}_i^{(p)} \mid \left\{\boldsymbol{\lambda}_i^{(h)}, \boldsymbol{\beta}^{(h)}, h = 1, \ldots, p\right\} \overset{\text{ind}}{\sim} \prod_{h=1}^{p} p\left(\boldsymbol{Y}_i^{(h)} \mid \boldsymbol{\lambda}_i^{(h)}, \boldsymbol{\beta}^{(h)}\right), \quad i = 1, \ldots, N, \\
&\log\left(\lambda_i^{(h)}(r,s)\right) = \phi_{c_i}^{\star,(h)}(r,s) + \boldsymbol{X}_i^{(h)} \boldsymbol{\beta}_{rs}^{(h)}, \; r \to s, \; r,s \in \mathcal{S}^{(h)}, \; h = 1, \ldots, p, \; i = 1, \ldots, N, \\
&\boldsymbol{\beta}^{(h)} \mid \boldsymbol{U}_{\boldsymbol{\beta}^{(h)}}, \boldsymbol{V}_{\boldsymbol{\beta}^{(h)}} \sim \text{MN}_{g^{(h)} \times d^{(h)}(d^{(h)}-1)}\left(\boldsymbol{0}, \boldsymbol{U}_{\boldsymbol{\beta}^{(h)}}, \boldsymbol{V}_{\boldsymbol{\beta}^{(h)}}\right), \quad h = 1, \ldots, p, \\
&\boldsymbol{\phi}_m^{\star} = (\boldsymbol{\phi}_m^{\star,(1)}, \ldots, \boldsymbol{\phi}_m^{\star,(p)}) \mid M, \boldsymbol{\mu}, \boldsymbol{\Omega}_G \sim P_0 = \text{N}_{D_p}\left(\boldsymbol{\mu}, \boldsymbol{\Omega}_G\right), \quad m = 1, \ldots, M, \\
&\boldsymbol{\mu}, \boldsymbol{\Omega}_G \mid G, \boldsymbol{m}_\mu, k_0, \nu, \boldsymbol{\Psi} \sim \text{N}_{D_p}\left(\boldsymbol{\mu} \mid \boldsymbol{m}_{\boldsymbol{\mu}}, k_0 \boldsymbol{\Omega}_G\right) G\text{-Wishart}_G\left(\boldsymbol{\Omega}_G \mid \nu, \boldsymbol{\Psi}\right), \\
&k_0 \mid a_{k_0}, b_{k_0} \sim \text{Gamma}\left(a_{k_0}, b_{k_0}\right), \\
&\mathbb{P}\left(c_i = m \mid S_m\right) \propto S_m, \quad m = 1, \ldots, M, \quad i = 1, \ldots, N, \\
&S_1, \ldots, S_M \mid M, \gamma_S \overset{\text{iid}}{\sim} \text{Gamma}\left(\gamma_S, 1\right), \\
&M \mid \Lambda \sim \text{Poi}_1(\Lambda), \\
&G = f\left(G_0\right), \quad p\left(G_0 \mid \eta\right) \propto \eta^{|E_0|}\left(1-\eta\right)^{\binom{p}{2}-|E_0|},
\end{aligned}
\tag{4}
$$

where we indicate with $\boldsymbol{\phi}_m^{\star}$ the vectors of unique baseline log-transition rates for the m-th component in the model and M is the unknown number of components in the mixture. Note that we do not include in the expression for the log-transition rates the term containing the time-varying continuous covariates $\boldsymbol{Z}_{ij}^{(h)}$ (see (2)) as they are not available in the main application. Their inclusion is straightforward by specifying a prior distribution on the regression coefficients $\boldsymbol{\gamma}_{rs}^{(h)}$ similar to the one used for $\boldsymbol{\beta}_{rs}^{(h)}$. Here $\boldsymbol{c} = (c_1, \ldots, c_N)$ represents the allocation vector, i.e. it specifies to which component the i-th observation is assigned to, with $\boldsymbol{\phi}_i = \boldsymbol{\phi}_{c_i}^{\star}$. The probability of c_i being equal to the m-th component of the mixture is proportional to the unnormalised weight S_m, for $m = 1, \ldots, M$. Therefore, due to the discreteness of the mixing measure, the parameters $\boldsymbol{\phi}_i$ are assigned to K_N different clusters, with $K_N \leq M$. In Section 3, we discuss prior choices for the application under study. We refer to Argiento and De Iorio (2019) for a thorough discussion on prior specification in mixture models with unknown number of components. However, we point out that the mixture component of the model is specified conditionally on the graph structure G, and the graph is common to the components of the mixture. Finally, $\text{MN}_{n \times p}(\boldsymbol{0}, \boldsymbol{U}, \boldsymbol{V})$ is the matrix-variate Normal distribution of dimension $n \times p$ centred on the null matrix $\boldsymbol{0}$ and with covariance matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ of dimensions $n \times n$ and $p \times p$, respectively. Note that if some processes have the same covariates, depending on the application, we could easily assume a dependent prior for the regression coefficients, or we could cluster individuals also on covariate effects, including the regression coefficients in the mixture prior.

2.5 Relationship with PPMx Models

The SUMS model has wide applicability in biomedical research as some processes can be regarded as responses and some as covariates. Indeed, time-varying categorical covariates, usually representing symptoms, are very common in the field of medical research, for instance in association with the monitoring of a patient's disease status over time. In the application to disease progression in Section 3, some processes represent mental health outcomes of interest, while others correspond to categorical clinical markers (e.g. sleep quality), and the goal of the analysis is to model the joint evolution of outcomes and clinical predictors. Handling of time-varying multivariate categorical information can be problematic in several applications. In a Bayesian framework, De Iorio et al. (2018) discuss possible solutions and propose an approach based on a latent health function borrowing ideas from Item Response Theory (Thissen and Steinberg, 2009). This approach, although computationally efficient, does not allow for a direct quantification of the covariate effect on the clinical response of interest and it may lead to identifiability problems. A simpler and more common approach to deal with time-varying categorical covariates is to introduce appropriate dummy variables, considerably increasing the number of parameters to be estimated, resulting in slower computations and lower effectiveness in high dimensional problems. Another computational effective solution is to summarise the covariates into an often arbitrary time-varying score, but at the cost of losing information and interpretability. Here, we propose the SUMS model, which overcomes the above problems, as simply some of the processes $Y^{(h)}$ can correspond to responses of interests and some others to clinical markers.

When some of the multi-state processes are interpreted as covariates, the SUMS model has interesting connections with Product Partition Models with Covariates (PPMx), a popular class of models in the Bayesian nonparametric literature (Müller et al., 1996, 2011). Analogously, our approach induces a covariate-driven clustering structure on the subjects and enables the inclusion of different types of covariates/responses in a natural and efficient way (Barcella et al., 2017). We now give details about the relationship between SUMS and PPMx. The main modelling idea behind the PPMx is to include covariates into the partition model (e.g., into a Dirichlet Process Mixture model framework) by modifying the prior on the partition induced by the mixture model via a similarity function summarising the covariate information. In this work, building upon Müller et al. (1996), we specify the PPMx model by modelling the covariates and the responses jointly, thus treating also the covariates as random variables. In this framework, inference is performed on an augmented probability space, i.e. the joint space of covariates and responses. Specifically, let $(Y^{(1)}(t), \dots, Y^{(p_Y)}(t))$ be the response processes and denote by $(H^{(1)}(t), \dots, H^{(p_H)}(t))$ the explanatory factors (covariate processes). Then we can rewrite the model in (4) for the processes $\{(Y^{(h)}(t), H^{(l)}(t)); h = 1, \dots, p_Y; l = 1, \dots, p_H; t \in \mathbb{R}^+\}$, with $p_Y + p_H = p$. Simplifying the notation for ease of explanation:

$$\boldsymbol{Y}_i^{(1)}, \dots, \boldsymbol{Y}_i^{(p_Y)}, \boldsymbol{H}_i^{(1)}, \dots, \boldsymbol{H}_i^{(p_H)} \mid \left\{ \boldsymbol{\phi}_i^{Y,(h)}, \boldsymbol{\phi}_i^{H,(l)}, h = 1, \dots, p_Y, l = 1, \dots, p_H \right\} \overset{\text{ind}}{\sim}$$
$$\prod_{h=1}^{p_Y} p\left(\boldsymbol{Y}_i^{(h)} \mid \boldsymbol{\phi}_i^{Y,(h)}\right) \prod_{l=1}^{p_H} p\left(\boldsymbol{H}_i^{(l)} \mid \boldsymbol{\phi}_i^{H,(l)}\right), \quad i = 1, \dots, N.$$

Let $\boldsymbol{c}$ be the vector of allocation variables introduced in (3), and let ρ_N be the partition of the indices $\{1,\dots,N\}$ induced by $\boldsymbol{c}$. We indicate by C_j the set of indices belonging to the j-th cluster, i.e. $C_j = \{i \in \{1,\dots,N\} \mid c_i = j\}$, and by $n_j = |C_j|$ its numerosity. Thus a partition with K_N clusters corresponds to $\rho_N = \{C_1,\dots,C_{K_N}\}$. Marginalising with respect to $S_1,\dots,S_M$ in (3), we obtain the following exchangeable partition probability function for ρ_N (Argiento and De Iorio, 2019):

$$p(\rho_N) = V(N,K_N)\prod_{j=1}^{K_N}\frac{\Gamma(\gamma_S+n_j)}{\Gamma(\gamma_S)}, \tag{5}$$

where $V(N,K_N)$ is a constant depending only on the sample size N and the number of clusters K_N.

Let $\boldsymbol{\phi}_j^\star$ be the location parameter corresponding to cluster C_j, for $j=1,\dots,K_N$. We can partition the vector $\boldsymbol{\phi}_j^\star = (\boldsymbol{\phi}_j^{\star,Y}, \boldsymbol{\phi}_j^{\star,H}) = (\boldsymbol{\phi}_j^{\star,(1)},\dots,\boldsymbol{\phi}_j^{\star,(p_Y)},\boldsymbol{\phi}_j^{\star,(p_Y+1)},\dots,\boldsymbol{\phi}_j^{\star,(p)})$, i.e. in a sub-vector corresponding to the location parameters of the response processes and one corresponding to the parameters of the covariate processes. In Supplementary Material Section 1 we show that, starting with the joint model on $(\boldsymbol{Y},\boldsymbol{H},\boldsymbol{\phi}^{\star,Y},\boldsymbol{\phi}^{\star,H},\rho_N,M)$ (see Section 2.4), conditioning first on the partition ρ_N and $\boldsymbol{H}$ and then marginalising with respect to $\boldsymbol{\phi}_j^{\star,H}$ (and other parameters), we obtain the PPMx representation of the SUMS model:

$$\begin{aligned}
&\boldsymbol{Y}_i^{(1)},\dots,\boldsymbol{Y}_i^{(p_Y)} \mid \boldsymbol{\phi}_1^{\star,Y},\dots,\boldsymbol{\phi}_{K_N}^{\star,Y},\rho_N \overset{\text{ind}}{\sim} \prod_{h=1}^{p_Y} p\left(\boldsymbol{Y}_i^{(h)} \mid \boldsymbol{\phi}_{c_i}^{\star,(h)}\right), \quad i=1,\dots,N,\\
&p\left(\rho_N \mid \boldsymbol{H}_1^\star,\dots,\boldsymbol{H}_{K_N}^\star\right) \propto V(N,K_N)\prod_{j=1}^{K_N}\frac{\Gamma(\gamma_S+n_j)}{\Gamma(\gamma_S)}\mathcal{G}(\boldsymbol{H}_j^\star),\\
&\boldsymbol{\phi}_j^{\star,Y} \mid K_N, \sim P_0^Y, \quad j=1,\dots,K_N,
\end{aligned} \tag{6}$$

where P_0^Y is the marginal distribution of $\boldsymbol{\phi}_j^{\star,Y}$ obtained from P_0. In model (6), $\frac{\Gamma(\gamma_S+n_j)}{\Gamma(\gamma_S)}$ is referred to as *cohesion*, while $\mathcal{G}(\boldsymbol{H}_j^\star)$ is the *similarity*, i.e. a function of the array of covariates corresponding to the subjects in cluster j, denoted as $\boldsymbol{H}_j^\star := \{\boldsymbol{H}_i : i \in C_j\}$, for $j=1,\dots,K_N$. The cohesion expresses prior information about the partition, such as the average size of a cluster, while the similarity function $\mathcal{G}$ captures the contribution of the covariates to the clustering structure. The term $\mathcal{G}$ in (6) allows subjects with similar covariates to be assigned to the same cluster with higher probability, and is obtained by marginalising the law of the processes $\boldsymbol{H}$ with respect to the corresponding cluster-specific parameters $\boldsymbol{\phi}_j^{\star,H}$ (see Supplementary Material Section 1 for a proof). Note that the prior on the partition in (5) belongs to the class of Gibbs-type priors (De Blasi et al., 2013), and presents a product partition structure (Quintana and Iglesias, 2003), which allows to derive model (6) starting from (4). We point out that the similarity function $\mathcal{G}$ is not known in closed form (see (5) in Supplementary Material), differently from the common PPMx specification, where the similarity function is usually obtained from a conjugate model for the covariate vector to simplify computations. In the proposed approach, the evaluation of $\mathcal{G}$ would require an expensive numerical approximation. For this reason, we resort to a conditional MCMC algorithm analogous to the one proposed by Argiento and De Iorio (2019), not requiring the evaluation of the integral in (6).

2.6 MCMC Algorithm

Posterior inference is performed through a MCMC algorithm, described in detail in Supplementary Material Section 2. The numerous non-conjugate updates required by the proposed model are tackled using adaptive Metropolis-Hastings sampling schemes (Haario et al., 2001; Atchadé et al., 2005), which need an additional short burn-in period. Additionally, inference under the proposed model is challenging given the presence of the graphs G and G_0. We adopt the birth-and-death approach of Mohammadi et al. (2015), and extend their algorithm to accommodate for MCMC moves on cliques instead of single edges, recalling that each edge in G_0 corresponds to a clique in G through the map f. Indeed, the original algorithm of Mohammadi et al. (2015) is based on theoretical results from the GGM literature (see Wang et al., 2012), which can be extended to our modelling settings. In Supplementary Material Section 4, we also compare the performance of our model with the approach of De Iorio et al. (2018) and with two alternative versions of the proposed model (i.e., Dirichlet Process (DP) and parametric versions). The results of the comparison show that the proposed model outperforms the parametric approach, as well as the nonparametric competitors in terms of clustering, leading to comparable results with respect to the estimation of regression coefficients.

3 Application to the GUSTO Study

The GUSTO study (Growing Up in Singapore Towards healthy Outcomes, Soh et al., 2014) is a longitudinal birth cohort study started in 2009 and involving Singaporean mothers and their children. The study is one of the most carefully phenotyped parent-offspring cohorts, focusing on the roles of foetal, developmental and epigenetic factors involved in early body composition as well as neuro-development. In this work we consider data on $N = 301$ mothers, followed during pre- and post-natal periods, starting from three months before childbirth. The main focus of the analysis is understanding the relationship among five psychometric indicators obtained from specific questionnaires: the Beck's Depression Inventory II (BDI II, Beck et al., 1961); the Edinburgh Postnatal Depression Scale (EPDS, Matthey et al., 2006); the State-Trait Anxiety Inventory (STAI, Spielberger et al., 1983) that can be decomposed into two different scores describing the anxious states (STAI-s), reflecting characteristics that can vary with time, and the anxiety traits (STAI-t), reflecting more stable characteristics; and the Pittsburgh Sleep Quality Index (PSQI, Buysse et al., 1989). Following the literature (Beck et al., 1961; Matthey et al., 2006; Meaney, 2018; Spielberger et al., 1983; Buysse et al., 1989), the score ranges of these questionnaires are discretised to obtain clinically relevant categories. The scores are recorded at different time points, as reported in Supplementary Material Table 3. These five processes represent time-varying categorical observations and are modelled jointly via SUMS, to capture significant relationships between them (see also Meaney et al., 2018). In our setting, the four mental health indicators (BDI, EPDS, STAI-s, STAI-t) represent the main clinical responses of interest, while the sleep quality indicator (PSQI) can be considered as a time-varying categorical covariate. For all processes, we assume missingness at random and impute missing values at the first time of observation from their full conditionals (see Section 2.2 of Supplementary Material). We are also provided with information regarding socio-demographic and clinical

markers, as well as scoring obtained from additional questionnaires measuring personality traits. In particular, we have individual scores for the Big Five Inventory (BFI, John et al., 1999) (including the scores for *Extraversion*, *Agreeableness*, *Conscientiousness*, *Neuroticism*, *Openness*, and *Liking*) and for the Maternal Childhood Adversity (MCA, Bouvette-Turcot et al., 2015). Many of the remaining covariates are time-homogeneous categorical, while no time-varying continuous covariates are available. The time-homogeneous continuous covariates are centred and scaled so that each column has null mean and unitary standard deviation, thus estimating the corresponding regression coefficients $\boldsymbol{\beta}^{(h)}$ on the same scale across processes. The full set of covariates (which is 17-dimensional, including dummy coding for the categorical ones) is described in more details in Supplementary Material Table 4, and is included in the specification of the instantaneous transition rates of the four psychometric processes, but not of PSQI.

Hyper-Prior Elicitation We fit the model described in Section 2.4 to the GUSTO data. We need to specify the hyperparameters for the priors in the three components of the model: the transition rates, the mixture model with random number of components and the graphical model.

In order to induce sparsity in the graph structure and identify meaningful relationship between the SUMS processes, we set the a-priori probability of edge inclusion to $\eta = 0.1$. The hyperparameters of the centring measure P_0 are: $\boldsymbol{m_\mu} = \mathbf{0}$, $k_0 \sim \text{Gamma}(1,1)$, $\nu = D_p + 2$ and $\boldsymbol{\Psi} = \mathbb{I}_{D_p}/\nu$, where $\mathbb{I}_p$ is the identity matrix of size p. In the case of a full graph G, the latter corresponds to $\mathbb{E}(\boldsymbol{\Omega}_G \mid G) = \mathbb{I}_{D_p}$. The regression coefficients $\boldsymbol{\beta}^{(h)}$ are a-priori independent and identically distributed, i.e. $\boldsymbol{U}_{\boldsymbol{\beta}^{(h)}} = \boldsymbol{V}_{\boldsymbol{\beta}^{(h)}} = \mathbb{I}_{g^{(h)}d^{(h)}(d^{(h)}-1)}$, for $h = 1, \ldots, 4$. Recall that covariates are not included in the model of the transition intensities corresponding to PSQI. The mixture prior for the baseline log-transition rates $\boldsymbol{\phi}_1^\star, \ldots, \boldsymbol{\phi}_M^\star$ is controlled by the hyperparameters Λ and γ_S. These parameters determine the distribution of the number of components and the corresponding allocation of the subjects, and are the object of an extensive sensitivity analysis presented in Supplementary Material Section 3. In this application we fix these parameters to $\Lambda = 0.01$ and $\gamma_S = 0.1$.

Posterior Inference We run the MCMC algorithm described in Section 2.6 for 50000 iterations, after an initial burn-in period of 1000 iterations used to initialise the adaptive Metropolis-Hastings, discarding 40000 iterations as burn-in and thinning every 2, obtaining a final sample of 5000. The algorithm is run on a Dell workstation with Intel Xeon W-2223 Processor (base frequency 3.60GHz) and takes approximately 2.5 hours. The computational time employed by this MCMC scales reasonably well with the dimension of the problem, as shown in the simulation study in Supplementary Material Section 4.1.

We explore the relationship between the multi-state processes as captured by a graph (see Section 2.3). Inference on the posterior distribution of the graphical structure G_0 is obtained by reporting the posterior edge inclusion probability for each pair of nodes. In Figure 2 we report the posterior median graph, obtained by including only those edges with posterior edge inclusion probability greater than 0.5 (Barbieri et al., 2004). The four

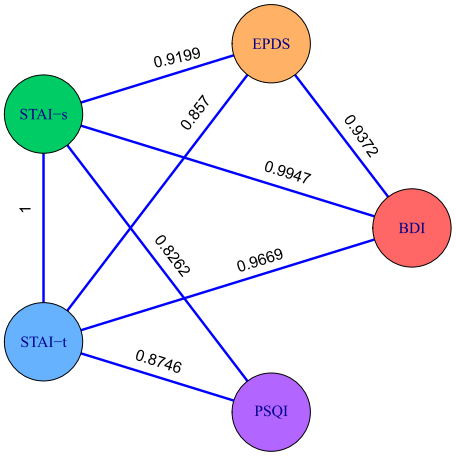

Figure 2: GUSTO study. Posterior median graph of G_0, each edge included in the graph has posterior edge inclusion probability greater than 0.5.

clinical mental health indicators BDI, EPDS, STAI-s and STAI-t are strongly associated, presenting a clique in the posterior median graph. Interestingly, the sleep quality index PSQI is only related to the anxiety indices STAI-s and STAI-t, forming a clique as well. Links between probable anxiety and sleeping quality have been reported in previous studies (Swanson et al., 2011; Ibrahim and Foldvary-Schaefer, 2012), and it is confirmed by our findings. Moreover, as previously reported, poor sleep quality may feed into poor emotional and mental health states (Ruiz-Robledillo et al., 2015; Osnes et al., 2019).

Another important aspect of the proposed model is the possibility of including covariates in the specification of the transition rates via (2). Posterior inference on the coefficient $\boldsymbol{\beta}^{(h)}$, for $h = 1, \ldots, 4$ is not trivial, due to the high number of parameters involved. The importance of each covariate can be assessed through Bayes Factors (BF), defined as the ratio of the marginal contributions derived from the model with the corresponding regression coefficient set to zero versus the full model (Kass and Raftery, 1995). Closed form expressions for the Bayes Factor under the SUMS model are not available, and thus we use the Savage-Dickey density ratio method (Wagenmakers et al., 2010; Verdinelli and Wasserman, 1995). The applicability of this method is guaranteed by the component-wise assumption of independence a-priori for the regression coefficients $\boldsymbol{\beta}^{(h)}$, for $h = 1, \ldots, 4$ (see the full model specification in (4)). For each process h, the values of $-\log_{10}(\mathrm{BF}_{jk}^{(h)})$ are reported in the heatmap of Figure 3, for $j = 1, \ldots, g^{(h)}$ and $k = 1, \ldots, d^{(h)}(d^{(h)} - 1)$. The magnitude of $-\log_{10}(\mathrm{BF}_{jk}^{(h)})$ measures the evidence in favour of the full model (Kass and Raftery, 1995). The majority of the coefficients is characterised by a low value of $-\log_{10}(\mathrm{BF}_{jk}^{(h)})$, supporting the hypothesis of no association, particularly in the case of the STAI processes. However, some coefficients are characterised by $-\log_{10}(BF)$ values above 1 or 2, indicating strong evidence in support of the inclusion of the corresponding covariate in the specific process. Of particular

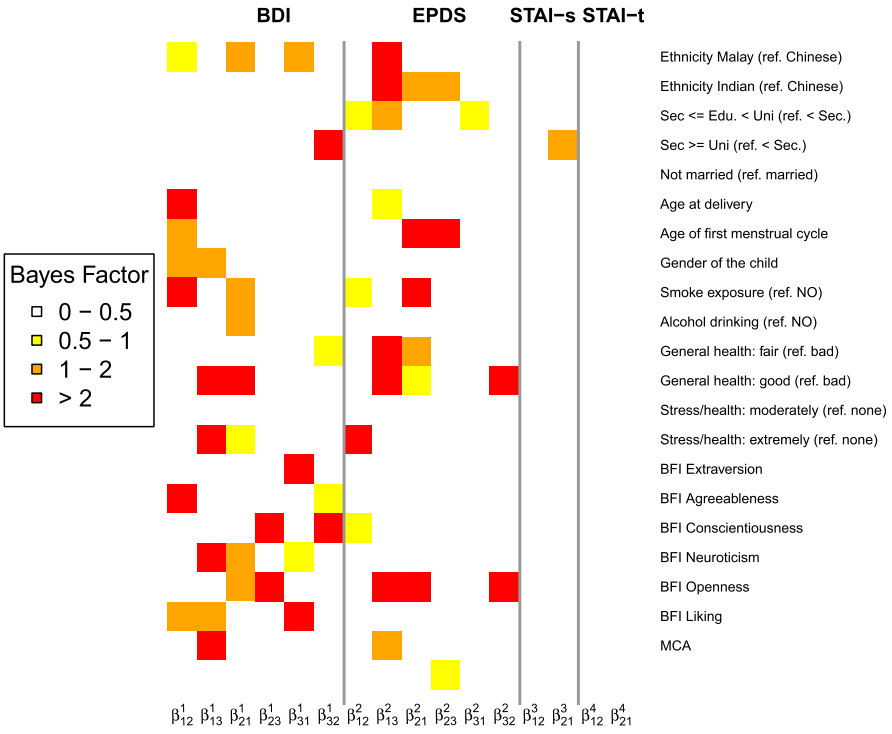

Figure 3: GUSTO study. Heatmap of Bayes Factors ($-\log_{10}(\mathrm{BF}^{(h)}_{jk})$) for the individual regression coefficients $\beta^{(h)}_{jk}$, for $j = 1, \ldots, g^{(h)}$, $k = 1, \ldots, d^{(h)}(d^{(h)} - 1)$ and $h = 1, \ldots, 4$. Each row refers to a different covariate included in the model. Each column is associated with a possible transition for each process, excluding PSQI which is modelled as an explanatory factor.

interest are the coefficients relative to the BFI and MCA scores, representing different traits of personality, trauma and parental relationship. We present the posterior mean and 95% credible intervals of the regression coefficients relative to BFI and MCA in detail in Figure 12 of Supplementary Material Section 6.

The personality traits of the mothers as described by the BFI scores have been previously associated with increased likelihood for both antenatal and postnatal mood disorder traits (Ritter et al., 2000; Leigh and Milgrom, 2008). Our analysis supports this as BFI traits have a relevant impact on both BDI and EPDS (95% credible interval does not contain zero). An interesting result appears through the estimates of the BFI's *Neuroticism* dimension, which characterises transitions $1 \to 3$ (deterioration, positive regression coefficient) and $2/3 \to 1$ (improvement, negative regression coefficient) in both BDI and EPDS scores, indicating that higher *Neuroticism* scores are associated with higher depressive symptoms during the peripartum period (Kitamura et al., 1993; O'hara and Swain, 1996). A similar result holds for the STAI-t process with respect

to *Neuroticism*. We also notice the effects of BFI's *Extraversion* and *Agreeableness* differ for BDI and EPDS's transitions. This could be explained by the fact that the social behaviours associated with *Extraversion* and *Agreeableness* are distinct (Tobin et al., 2000; Jensen-Campbell and Graziano, 2001). Extraverts tend to actively seek out social interactions, whereas people scoring high on *Agreeableness* prefer harmonious relationships. Maternal history of developmental adversity is linked to increased risk for depression (Leigh and Milgrom, 2008), of which childhood abuse is a strong risk factor (Seng et al., 2014), as highlighted by the importance of the MCA covariate for the transition $3 \to 2$ (improvement) in BDI and EPDS (see Figure 3 and Supplementary Material Figure 12). This result is also confirmed by Mandelli et al. (2015), who reports that women who are victims of childhood neglect or abuse are at least twice as likely to suffer from depression. The quality of the relationship with their parents may also contribute to maternal developmental adversity. Mothers who received low parental care and high control during childhood are at risk for peripartum anxiety (Grant et al., 2012) and depression (McMahon et al., 2005).

The choice of the mixture model (3) as prior distribution for the vector of log-transition rates $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ allows for clustering of the subjects. Inference on the random partition is shown in Supplementary Material Figure 13, where the posterior distributions of the number of clusters, components and of the co-clustering probabilities are reported. An estimate of the random partition induced on the subjects under study is obtained by minimizing Binder's loss function (Binder, 1978) with equal costs. We obtain a partition with three clusters, which also corresponds to the posterior mode of the number of clusters. The three clusters contain 135, 135 and 31 subjects, respectively, and are labelled according to their size in decreasing order. In Figure 4 we report the posterior distribution of $\phi^{(h)}(r, s)$ conditional on the Binder's partition, for $r \to s$ and $h = 1, \ldots, p$. Cluster-specific estimates of transition rates differ among clusters (see Supplementary Material Section 7 for a discussion). For instance, transition rates corresponding to improvement in the BDI or EPDS scores are higher in Clusters 1 and 2 rather than Cluster 3. The two-states processes (STAI-s, STAI-t and PSQI) also seem to present differences between clusters in the same direction, identifying Cluster 3 as the one most prone to a deterioration in mental health status of its subjects. A similar behaviour can be observed in Supplementary Material Figure 14, where we show posterior inference on transition probabilities, as well as predictive distributions.

4 Conclusions

Observations on time-evolving related processes are very common in biomedical applications and beyond. In this work we present a Bayesian semiparametric approach for joint modelling of several multi-state Markov processes, describing an individual's transitions between different states in continuous time. The proposed model builds on the multi-state Markov models, GGM and PPMx literature. The different multi-state processes are linked by imposing a flexible prior distribution for the instantaneous transition rates, which allows for data-driven clustering of the subjects. The dependence among the processes is captured by a graph and posterior inference is performed through a tailored MCMC algorithm.

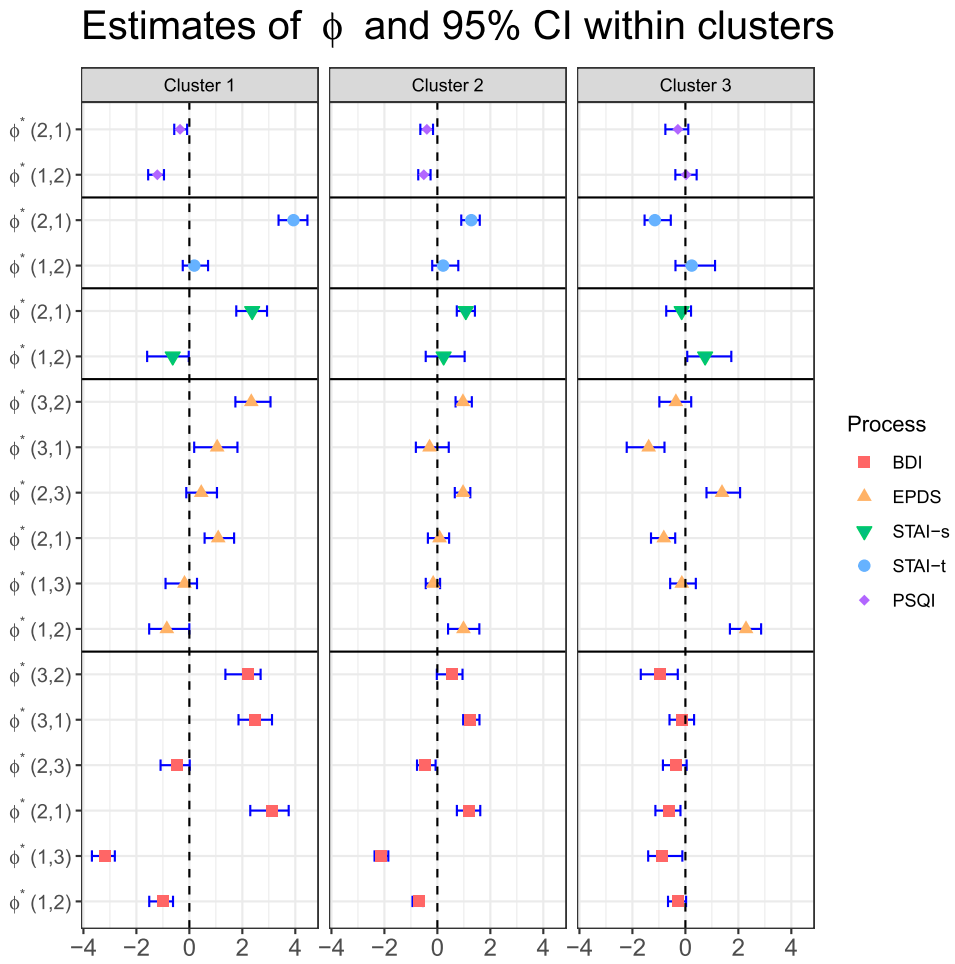

Figure 4: GUSTO study. Posterior means and 95% credible intervals of the instantaneous log-transition rates $\phi^{(h)}(r,s)$ for each process $h = 1, \ldots, p_0$. The vertical dashed lines represent the value 0, while the horizontal continuous lines divide the estimates for the five processes. The estimates are obtained by fixing the partition of the subjects to the Binder's partition, and re-running the algorithm for the conditional model. Each sub-plot refers to one of the clusters in the fixed partition.

The proposed model finds wide applicability, due to its flexibility, interpretability and relative ease of computations. In this work, we analyse data from the GUSTO cohort study with the aim of understanding the evolution and relationship between mental health indicators over time. Our findings are in agreement with existing medical literature and shed more light on the influence of childhood and parental factors on mental health progression. Potential extensions include higher order Markov dependency and joint modelling of multi-state processes and continuous longitudinal trajectories.

A possible alternative to our approach is to represent the categorical covariates with continuous Gaussian latent variables linked to the categorical outcome by thresholding (Albert and Chib, 1993), allowing for the inclusion of a time component through auto-regressive terms in the likelihood (e.g. Barcella et al., 2018). To the best of our knowledge, this strategy has not been employed in the context of multi-state models, and it represents an interesting direction for future developments. However, this formulation could suffer from limited interpretability (García-Zattera et al., 2007) and could induce further computational challenges (Zhang et al., 2006).

Supplementary Material

Seemingly Unrelated Multi-State processes: a Bayesian semiparametric approach. Supplementary Material (DOI: 10.1214/22-BA1326SUPP; .pdf). Supplementary Material file referenced throughout the manuscript. R/C++ package and simulated example available at the GitHub repository https://github.com/AndCre87/SUMS.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association*, 88(422): 669–679. MR1224394. 18

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media. MR1198884. doi: https://doi.org/10.1007/978-1-4612-4348-9. 5

Argiento, R. and De Iorio, M. (2019). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *arXiv preprint arXiv:1904.09733*. 2, 5, 6, 9, 11

Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). "Bayesian covariance matrix estimation using a mixture of decomposable graphical models." *Statistics and Computing*, 19(3): 303–316. MR2516221. doi: https://doi.org/10.1007/s11222-008-9093-8. 9

Atchadé, Y. F., Rosenthal, J. S., et al. (2005). "On adaptive Markov chain Monte Carlo algorithms." *Bernoulli*, 11(5): 815–828. MR2172842. doi: https://doi.org/10.3150/bj/1130077595. 12

Banterle, M., Bottolo, L., Richardson, S., Ala-Korpela, M., Järvelin, M., and Lewin, A. (2018). "Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression." *bioRxiv*, 467019. 5

Barbieri, M. M., Berger, J. O., et al. (2004). "Optimal predictive model selection." *The annals of statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 13

Barcella, W., De Iorio, M., and Baio, G. (2017). "A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models." *Canadian*

Journal of Statistics, 45(3): 254–273. MR3691836. doi: https://doi.org/10.1002/cjs.11323. 2, 10

Barcella, W., De Iorio, M., and Malone-Lee, J. (2018). "Modelling correlated binary variables: an application to lower urinary tract symptoms." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4): 1083–1100. MR3832265. doi: https://doi.org/10.1111/rssc.12268. 18

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). "An Inventory for Measuring Depression." *Archives of General Psychiatry*, 4(6): 561–571. doi: https://doi.org/10.1001/archpsyc.1961.01710120031004. 12

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65(1): 31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 16

Bouvette-Turcot, A.-A., Fleming, A., Wazana, A., Sokolowski, M., Gaudreau, H., Gonzalez, A., Deslauriers, J., Kennedy, J., Steiner, M., Meaney, M., et al. (2015). "Maternal childhood adversity and child temperament: An association moderated by child 5-HTTLPR genotype." *Genes, Brain and Behavior*, 14(3): 229–237. 13

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., Kupfer, D. J., et al. (1989). "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research." *Psychiatry res*, 28(2): 193–213. 12

Cremaschi, A., Argiento, R., De Iorio, M., Shirong, C., Chong, Y. S., Meaney, M., and Kee, M. (2022). "Seemingly Unrelated Multi-State processes: a Bayesian semiparametric approach. Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1326SUPP. 3

Cook, R. J. (1999). "A mixed model for two-state Markov processes under panel observation." *Biometrics*, 55(3): 915–920. 2

Cox, D. R. and Miller, H. D. (1977). *The theory of stochastic processes*, volume 134. CRC press. MR0192521. 3

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 212–229. 11

De Iorio, M., Gallot, N., Valcarcel, B., and Wedderburn, L. (2018). "A Bayesian semiparametric Markov regression model for juvenile dermatomyositis." *Statistics in medicine*, 37(10): 1711–1731. MR3787983. doi: https://doi.org/10.1002/sim.7613. 2, 6, 10, 12

Dempster, A. (1972). "Covariance selection." *Biometrics*, 28: 157–175. MR3931974. 7

Ferrer, L., Rondeau, V., Dignam, J., Pickles, T., Jacqmin-Gadda, H., and Proust-Lima, C. (2016). "Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer." *Statistics in medicine*, 35(22): 3933–3948. MR3545618. doi: https://doi.org/10.1002/sim.6972. 5

García-Zattera, M. J., Jara, A., Lesaffre, E., and Declerck, D. (2007). "Conditional independence of multivariate binary data with an application in caries research." *Com-*

putational statistics & data analysis, 51(6): 3223–3234. MR2345637. doi: https://doi.org/10.1016/j.csda.2006.11.021. 18

Grant, K.-A., Bautovich, A., McMahon, C., Reilly, N., Leader, L., and Austin, M.-P. (2012). "Parental care and control during childhood: associations with maternal perinatal mood disturbance and parenting stress." *Archives of women's mental health*, 15(4): 297–305. 16

Grimmet, G. and Stirzaker, D. (2001). *Probability and Random Processes.*. Oxford University press. MR2059709. 3

Haario, H., Saksman, E., Tamminen, J., et al. (2001). "An adaptive Metropolis algorithm." *Bernoulli*, 7(2): 223–242. MR1828504. doi: https://doi.org/10.2307/3318737. 12

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). "Phenomics: the next challenge." *Nature reviews genetics*, 11(12): 855–866. 5

Ibrahim, S. and Foldvary-Schaefer, N. (2012). "Sleep disorders in pregnancy: implications, evaluation, and treatment." *Neurologic clinics*, 30(3): 925–936. 14

Jackson, C. H. et al. (2011). "Multi-state models for panel data: the msm package for R." *Journal of statistical software*, 38(8): 1–29. 2

Jensen-Campbell, L. A. and Graziano, W. G. (2001). "Agreeableness as a moderator of interpersonal conflict." *Journal of personality*, 69(2): 323–362. 16

John, O. P., Srivastava, S., et al. (1999). "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." *Handbook of personality: Theory and research*, 2(1999): 102–138. 13

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the american statistical association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 14

Kim, Y., James, L., and Weissbach, R. (2012). "Bayesian analysis of multistate event history data: beta-Dirichlet process prior." *Biometrika*, 99(1): 127–140. MR2899668. doi: https://doi.org/10.1093/biomet/asr067. 4, 6

Kitamura, T., Shima, S., Sugawara, M., and Toda, M. (1993). "Psychological and social correlates of the onset of affective disorders among pregnant women." *Psychological medicine*, 23(4): 967–975. 15

Kneib, T. and Hennerfeind, A. (2008). "Bayesian semi parametric multi-state models." *Statistical Modelling*, 8(2): 169–198. MR2750636. doi: https://doi.org/10.1177/1471082X0800800203. 6

Leigh, B. and Milgrom, J. (2008). "Risk factors for antenatal depression, postnatal depression and parenting stress." *BMC psychiatry*, 8(1): 24. 15, 16

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The annals of statistics*, 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 6

Mandelli, L., Petrelli, C., and Serretti, A. (2015). "The role of specific early trauma in adult depression: a meta-analysis of published literature. Childhood trauma and adult depression." *European psychiatry*, 30(6): 665–680. 16

Matthey, S., Henshaw, C., Elliott, S., and Barnett, B. (2006). "Variability in use of cut-off scores and formats on the Edinburgh Postnatal Depression Scale–implications for clinical and research practice." *Archives of women's mental health*, 9(6): 309–315. 12

McMahon, C., Barnett, B., Kowalenko, N., and Tennant, C. (2005). "Psychological factors associated with persistent postnatal depression: past and current relationships, defence styles and the mediating role of insecure attachment style." *Journal of affective disorders*, 84(1): 15–24. 16

Meaney, M., van Lee, L., Cai, S., Loy, S. L., Tham, E. K., Yap, F. K., Godfrey, K. M., Gluckman, P. D., Shek, L. P., Teoh, O. H., Goh, D. Y., et al. (2018). "Relation of plasma tryptophan concentrations during pregnancy to maternal sleep and mental well-being: The GUSTO cohort." *Journal of affective disorders*, 225: 523–529. 12

Meaney, M. J. (2018). "Perinatal maternal depressive symptoms as an issue for population health." *American Journal of Psychiatry*, 175(11): 1084–1093. 12

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 5

Mohammadi, A., Wit, E. C., et al. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1): 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 12

Moler, C. and Van Loan, C. (2003). "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later." *SIAM review*, 45(1): 3–49. MR1981253. doi: https://doi.org/10.1137/S00361445024180. 3

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83(1): 67–79. MR1399156. doi: https://doi.org/10.1093/biomet/83.1.67. 10

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 2, 10

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture." *The Annals of Statistics*, 32(5): 2044–2073. MR2102502. doi: https://doi.org/10.1214/009053604000000788. 5

O'hara, M. W. and Swain, A. M. (1996). "Rates and risk of postpartum depression—a meta-analysis." *International review of psychiatry*, 8(1): 37–54. 15

Osnes, R. S., Roaldset, J. O., Follestad, T., and Eberhard-Gran, M. (2019). "Insomnia late in pregnancy is associated with perinatal anxiety: a longitudinal cohort study." *Journal of affective disorders*, 248: 155–165. 14

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574. MR1983764. doi: https://doi.org/10.1111/1467-9868.00402. 11

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4): 731–792. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 5

Ritter, C., Hobfoll, S. E., Lavin, J., Cameron, R. P., and Hulsizer, M. R. (2000). "Stress, psychosocial resources, and depressive symptomatology during pregnancy in low-income, inner-city women." *Health Psychology*, 19(6): 576. 15

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). *Stochastic processes*, volume 2. Wiley New York. 3, 4

Roverato, A. (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models." *Scandinavian Journal of Statistics*, 29(3): 391–411. MR1925566. doi: https://doi.org/10.1111/1467-9469.00297. 7

Ruiz-Robledillo, N., Canário, C., Dias, C., Moya-Albiol, L., and Figueiredo, B. (2015). "Sleep during the third trimester of pregnancy: the role of depression and anxiety." *Psychology, health & medicine*, 20(8): 927–932. 14

Seng, J. S., D'Andrea, W., and Ford, J. D. (2014). "Complex mental health sequelae of psychological trauma among women in prenatal care." *Psychological Trauma: Theory, Research, Practice, and Policy*, 6(1): 41. 16

Soh, S.-E., Tint, M. T., Gluckman, P. D., Godfrey, K. M., Rifkin-Graboi, A., Chan, Y. H., Stünkel, W., Holbrook, J. D., Kwek, K., Chong, Y.-S., et al. (2014). "Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study." *International journal of epidemiology*, 43(5): 1401–1409. 12

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., and Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory*. Consulting Psychologists Press, Palo Alto, CA. 12

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *Annals of statistics*, 40–74. MR1762903. doi: https://doi.org/10.1214/aos/1016120364. 5

Sung, M., Soyer, R., and Nhan, N. (2007). "Bayesian analysis of non-homogeneous Markov chains: Application to mental health data." *Statistics in medicine*, 26(15): 3000–3017. MR2370982. doi: https://doi.org/10.1002/sim.2775. 2

Swanson, L. M., Pickett, S. M., Flynn, H., and Armitage, R. (2011). "Relationships among depression, anxiety, and insomnia symptoms in perinatal women seeking mental health treatment." *Journal of Women's Health*, 20(4): 553–558. 14

Thissen, D. and Steinberg, L. (2009). "Item response theory." *The Sage handbook of quantitative methods in psychology*, 148–177. 10

Tobin, R. M., Graziano, W. G., Vanman, E. J., and Tassinary, L. G. (2000). "Personality, emotional experience, and efforts to control emotions." *Journal of personality and social psychology*, 79(4): 656. 16

van den Hout, A., Fox, J.-P., and Klein Entink, R. H. (2015). "Bayesian inference for an illness-death model for stroke with cognition as a latent time-dependent risk factor." *Statistical methods in medical research*, 24(6): 769–787. MR3428428. doi: https://doi.org/10.1177/0962280211426359. 2

Verdinelli, I. and Wasserman, L. (1995). "Computing Bayes factors using a generalization of the Savage-Dickey density ratio." *Journal of the American Statistical Association*, 90(430): 614–618. MR1340514. 14

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). "Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method." *Cognitive psychology*, 60(3): 158–189. 14

Wang, H. (2010). "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics." *Computational Statistics & Data Analysis*, 54(11): 2866–2877. MR2720481. doi: https://doi.org/10.1016/j.csda.2010.03.028. 2

Wang, H., Li, S. Z., et al. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. MR2879676. doi: https://doi.org/10.1214/12-EJS669. 12

Zellner, A. (1963). "Estimators for seemingly unrelated regression equations: Some exact finite sample results." *Journal of the American Statistical Association*, 58(304): 977–992. MR0157439. 2, 5

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). "Sampling correlation matrices in Bayesian models with correlated latent variables." *Journal of Computational and Graphical Statistics*, 15(4): 880–896. MR2297633. doi: https://doi.org/10.1198/106186006X160050. 18