Dynamic Nonparametric Bayesian Models And the Birth-Death Process

Eric P. Xing

December 2005
CMU-CALD-05-114

Center for Automated Learning & Discovery
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

When modeling longitudinal data using a set of hidden processes such as state-space models, a common assumption is that the number of hidden processes is fixed, and all hidden processes have the same life span (i.e., all start at the onset of the data stream and terminate at the end of the data stream). In this report I outline a framework of modeling complex longitudinal data using a birth-death process, in which hidden processes and emerge, evolve, and extinct over time. The model is built on top of a temporally evolving Dirichlet process, and thus allow the total number of hidden processes to be unbounded. I also derive a Gibbs sampling algorithm for inference on this model.

Keywords:
Dirichlet Process, nonparametric Bayesian models, birth-death process, Kalman filter, state-space models, longitudinal data analysis.

1 Introduction

Our goal is to design a temporal mixture model, in which the number of mixture components is unbounded, the component can retain, die out or emerge over time, and (the actual configuration or parameterization of) each component can also evolve over time in a Markovian fashion. This means that for multiple consecutive time points, the data observed over these time points can come from a set of commons themes living through this period (which correspond to the mixture components being retained over this period) or themes that partially whose life span overlap partially with this period (i.e., mixture components that die out or emerge during this period). We also assume that the conditional distribution of observed data under the same (lasting) component at different time would be a little different, i.e., the parameterization of the retained components will evolve over time. This scenario is not uncommon for real world sequential data. For example, consider modeling the temporal stream of news articles on a, say, weekly basis. Moving from week to week, some old topics could fade out (e.g., the election is now over in US), while new topics could appear over time (e.g., the Pope is being hospitalized). The specific content of the lasting topics could also (slowly) change over time (i.e., war in Iraq is developing with some slight shift of focus ...). Note that given the (number and parameterization of) components of the mixture model at each time point, we are free to design appropriate likelihood models for the data, e.g., the mixture membership model (i.e., the LDA model), the exponential harmonium model, or other extended models.

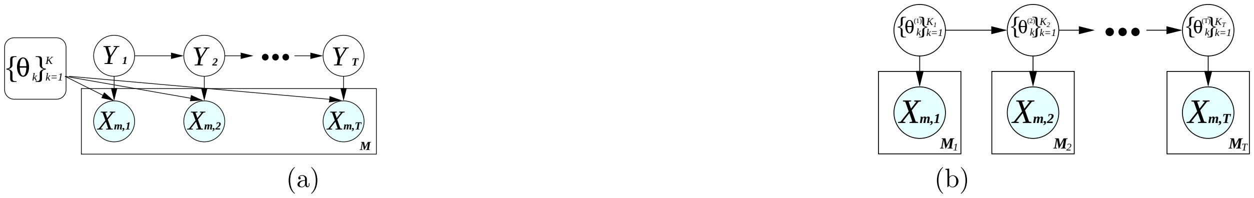

Figure 1: (a) HMM. (b) DNBM.

It is important to distinguish the temporal mixture model concerned here from an HMM, which is sometimes also (confusingly) understood as a temporal sequence of mixture models. In a typical HMM one assumes the presence of a stable (time invariant) finite mixture model, and the stochastic sequence of hidden states are used to model choices of different mixture components over time for sequences of observed data (typically a single data point at a time) emitted from the chosen component at each time point. Generally, multiple observed sequences are assumed to be *iid* samples from the same HMM (Fig 1a). The infinite HMM model allows the number of mixture components in such a mixture to be unbounded, but it is still a time invariant mixture, and the hidden states sequences choose one component at a time in the space of all components. In the model to be described in the following, we concern ourselves not only with the (dynamic) choices of mixture components at each time for each data point in a corpus, but also the evolution of the mixture model itself (e.g., its cardinality, centroids, birth and death events, etc.) over time, which defines time-specific mixture models for data in a sequence of corpora. Specifically, the output our model are temporal collections of data sets (e.g., weekly collections of corpora), and at each time point, all the elements in the corpus of that time are (marginally) *iid* samples of the entire mixture model of that time (Fig 1b).

Here is the plan for the rest of the paper. In section 2, I will present the general probabilistic structure of a *dynamic nonparametric Bayesian model* (DNBM) with Dirichlet process mixing and Bernoulli sieving, using three different generative schemes. I will first use a sequential Chinese restaurant process (CRP) to illustrate a constructive definition of DNBM. Then I will give another construction based on the infinite limit of a finite dimensional, dynamically evolving mixture model, which connects DNBM to the more familiar state-space model for (finite) topic evolution. Then I will give a stick-breaking construction of DNBM which can be used to device more efficient sampling algorithms or variational methods for inference. In section 3, I will give detailed description of a nonparametric state space model instantiated from DNBM, which can be used to model the aforementioned temporally evolving mixtures. Finally I will develop a Gibbs sampling algorithm for approximate inference on such models.

2 The Model

2.1 A sequential CRP construction

A Chinese restaurant process can be described by the following metaphor. We have a Chinese restaurant with infinite number of tables. The first customer comes in and deterministically sit at the first table and orders a dish θ_1 from distribution G_0 for this table. After having $i-1$ customers, suppose we have K_i nonempty tables, each table, e.g., table k, is occupied by n_k customers and serves a shared dish θ_k randomly sampled from a distribution G_0. When an i-th customer comes in, he will either choose an occupied table, say, table k, with probability $\frac{n_k}{i-1+\tau}$ and share the dish θ_k on that table; or he can sit at a new table K_i+1, with probability $\frac{\tau}{i-1+\tau}$, and order a new dish θ_{K_i+1} for this table. The set of θ_k's generated from this process are discrete random measures following a Dirichlet process $DP(\tau, G_0)$. A DP can be used as a prior distribution for a infinite mixture model, with the θ_k on each table serving as the parameter of a mixture component, and the customers sitting at that table corresponding to the data points in the corresponding mixture components admitting $p(\cdot|\theta_k)$.

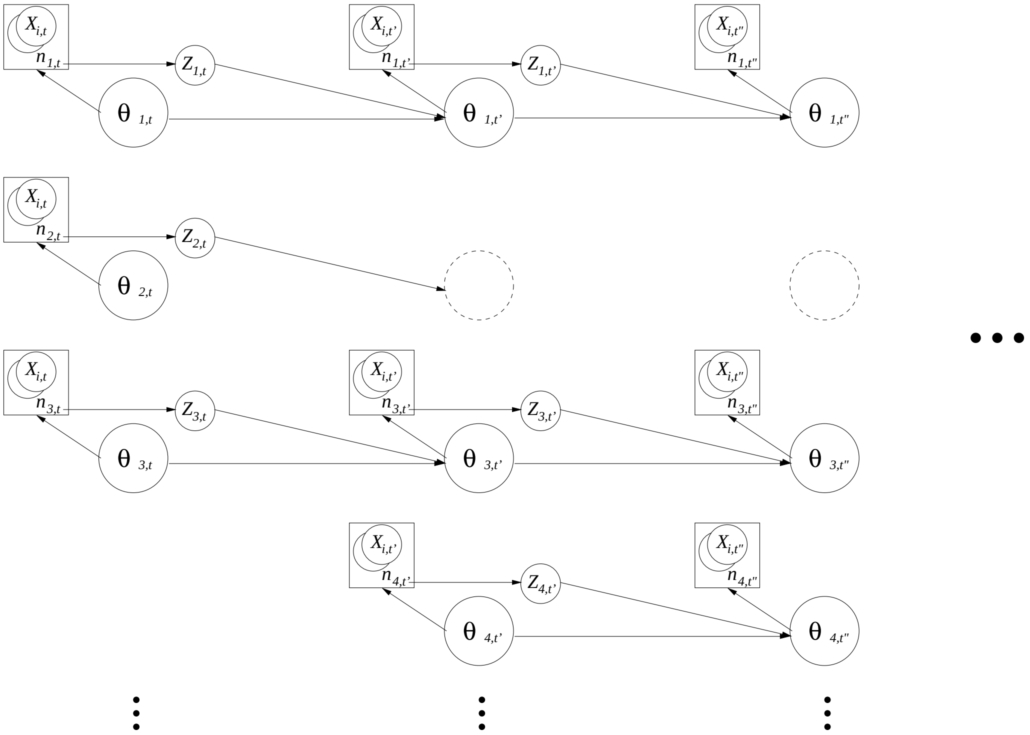

Figure 2: CRP. (Just for illustration, not all the dependencies are depicted.)

2.1.1 The temporally dependent CRP

Now we construct a series of temporally dependent CRP using the following metaphor. Consider at time t, we have a Chinese restaurant with K nonempty tables (there are infinite number of unoccupied tables available for more customers if necessary, but we do not explicitly represent them unless then are taken), each table (i.e., each component in a mixture), for example, table k, serves a distinctive dish θ_k (the parameters of this component), and is occupied by n_k customers denoted by $x_{i,t}$'s. Now suppose we are moving to the next time point $t+1$. We assume that the same restaurant carries on, and all the non-empty tables are to be retained. For each of the retained tables, we inherit the dishes on these tables. In a document clustering scenario, the dishes may correspond to topic-specific multinomial parameter vector of word frequencies $\vec{\boldsymbol{\theta}}$. To introduce evolution of the topic contents over time, we further assume that, for the retained tables, we have $\vec{\boldsymbol{\theta}}_{k,t+1} \sim p(\cdot|\vec{\boldsymbol{\theta}}_{k,t})$. One handy conditional model for this purpose is the state-space model for stochastic dynamic evolutions of multivariate Gaussian variables, of which the inference can be solved in close-form via a Kalman filter or stochastically using particle filtering. As described in section 3, each contiguous sequences of dishes corresponding to one table over time (i.e., $\vec{\boldsymbol{\theta}}_{k,1}, \ldots, \vec{\boldsymbol{\theta}}_{k,T_k}$) can be modeled by a single SSM. topic-specific dynamics can be easily introduced if desired. Unfortunately, the multinomial vector $\vec{\boldsymbol{\theta}}_{k,t}$ for a topic is not multivariate Gaussian (e.g., they are subject to the normalization constrain). A heuristic surrogate is to treat $\ln \vec{\boldsymbol{\theta}}_{k,t}$ as multivariate Gaussian and subject it to the state-space model. At the next time point, we recover $\vec{\boldsymbol{\theta}}_{k,t+1}$ from the (linear Gaussian) transformed $\ln \vec{\boldsymbol{\theta}}_{k,t+1}$ (for simplicity, in the sequel we use θ, i.e., a generic "dish" for the table, instead of $\vec{\boldsymbol{\theta}}$, to denote the parameter of a mixture component). Dave Blei has shown that this trick induces well-behaving transitions within a simplex. It is possible to define a more direct random walk model on multinomial simplexes. Now, although the dish of a retained table is defined by the aforementioned state-space model, to return to the non-parametric nature of the overall model, for tables that are newly instantiated at time $t+1$, we still sample the dishes from the base measure G_0.

Putting everything together, to receive new customers at time $t+1$, each customer use the usual CRP strategy, pick an occupied table according to the present plus the immediate previous occupancy of this table, and have the dish on that table, or with probability $\frac{\tau}{\sum_k n_{k,t+1} + \sum_{k'} n_{k',t} + \tau}$, pick a new table and sample a new dish from G_0.

The above-described model can be understood as a sequence of conditional CRPs for each time point, given the CRPs in the immediate previous time points. If we assume that the distribution of customer-sittings at time t is governed by a Dirichlet process, then without the death process (i.e., all tables are retained over time) and the topic evolution process (i.e., all dishes are non-evolving over time), then the distribution of customer-sittings at time $t+1$ is essentially the posterior distribution of Dirichlet process $DP(\tau, G_0)$ given the customer-sittings at time t:

$$\begin{aligned} & \theta_{i,t+1} | \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, \{n_{k,t}, \theta^*_{k,t}\}_{k=1}^{K_t}, G_0, \tau \\ \sim \; & \sum_{k=1}^{K_i} \frac{n_{k,t+1} + \sum_{k'}^{K_t} n_{k',t} \mathrm{Pre}(\cdot, \theta^*_{k',t})}{i-1+N_t+\tau} \delta_{\theta^*_{k,t+1}}(\cdot) + \frac{\alpha_0}{i-1+N_t+\tau} G_0, \end{aligned} \tag{1}$$

where $\theta^*_{k,t+1}$ (resp. $\theta^*_{k,t}$) denotes the k-th unique value of θ at time point $t+1$ (resp. time t), and $N_t = \sum_{k=1}^{K_t} n_{k,t}$ is the total number of customers at time t, and $\mathrm{Pre}(a,b)$ denotes an indicator function that equals to 1 when b is the predecessor of a under a state-space model and 0 otherwise. According to Ferguson, the posterior of a Dirichlet process is also a Dirichlet process.

2.1.2 The Bernoulli sieve

To model the phenomena that topics usually last a finite amount of time, now we assume that when transitioning from time t to $t+1$, the same restaurant carries on, but the fate of each table from time t (identifiable because of its associated distinct dish) depends on its popularity at time t (Fig 2). Specifically, each table is associated with a "extinction probability" $1-\alpha_k$, and a Bernoulli trial is to be performed based on this probability for each table. To related the table's retention probability with the occupancy of the table, we let:

$$\alpha_k = \frac{1}{1+\exp\{-\eta(n_k/\sum_k n_k - 1/K)\}} \tag{2}$$

Essentially, if the occupancy rate (i.e., $n_k/\sum_k n_k$) of a table k is much larger than the average occupancy rate (i.e., $1/K$), the table has a high probability to be retained; o/w, it is more likely to disappear in the following time point. To further impose reasonable regularization of such a Bernoulli process, e.g., encouraging retention of most of the reasonably occupied tables and proposing removal of only extremely unpopular tables, we control the shape of the logistic function by determining η using the following three-point logistic regression scheme.

We assume that the table with the biggest occupancy (or if desirable, the k_h-th most occupied table) has a retention probability $\alpha_{\max}$ (say 0.99), the table with the smallest occupancy (or if desirable, the k_l-th lest occupied table) has a retention probability $\alpha_{\min}$ (say 0.01). Finally, we assume that the most populous table in the lowest quartile of all table has a retention probability of $\alpha_{1/4} = 0.5$. Thus, letting $y = [\alpha_{\min}, \alpha_{1/4}, \alpha_{\max}]^T$, and $x = [n_l, n_m, m_h]^T$ (the occupancies of the reference tables), we have the following linear regression problem:

$$\begin{aligned} & \eta x = 1 - \log(1/y - 1) = y' \\ \Rightarrow \quad & \eta = (x^T x)^{-1} x^T y' \end{aligned} \tag{3}$$

If desirable, we can perform a further transformation of $\alpha = \{\alpha_1, \ldots, \alpha_K\}$ by letting $\beta = A\alpha$, which introduces coupling of the extinction rates between tables. Note that for each time point t, we will recompute the α's based on the occupancy at time t. Thus we have a non-stationary, conditional birth-death process for each (non-empty) table, parameterized by $\{\alpha_{k,t} \ : \ t = 1, \ldots, T; k|t = 1, \ldots, K_t\}$.

Now, for every table k at time t, we define a Bernoulli indicator $z_{k,t}$ with parameter $\alpha_{k,t}$ (or $\beta_{k,t}$ if a coupling of transformation is used). Given $\mathbf{z}_t = \{z_{k,t}\}_{k=1}^{K_t}$, a subset of tables occupied at time t will be eliminated and the rest are inherited to time $t+1$, together with the dishes $\{\theta_{k,t} \ : z_{k,t} = 1, \ k = 1, \ldots, K_t\}$ served on the tables (with some modifications described bellow). This means that when these inherited tables are picked for the first time at $t+1$, we do not need to go to the base measure to sample the dish. Furthermore, to reflect the popularities of dishes inherited from t, we associate these retained tables with a pseudo count of the customers equal to the occupancy number of the corresponding tables at time t. This is analogous to the scenario that the restaurant actually keep these tables as "seed" tables already instantiated before seeing any customer and the incoming customers will see both the present and immediate previous popularities of the tables, according to which they pick their seats. To briefly summarize, now we have a dynamic model for *birth-death* of mixture components, and the evolution of the popularities of the components.

With the Bernoulli process and the linear-Gaussian state space model for dish evolution, the conditional distribution of the dish for the i-th costumer at time $t+1$, given configuration of the restaurant at time t, and customer-sittings at time $t+1$ up to the $(i-1)$-th costumer (we record both the dish each customer had, $\theta_{i,t+1}$, and an auxiliary indicator $c_{k,t+1} \in \{1, \ldots, K_t\}$ indicating from which dish at time t a distinct $\theta^*_{k,t+1}$ is evolved):

$$\begin{aligned} & \theta_{i,t+1} | \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, c_{1,t+1}, \ldots, c_{K_i,t+1}, \{n_{k,t}, \theta^*_{k,t}, z_{k,t}\}_{k=1}^{K_t}, G_0, \tau \\ \sim \quad & \sum_{k=1}^{K_i} \frac{n_{k,t+1} + \sum_{k'=1}^{K_t} z_{k',t} n_{k',t} \delta_{k'}(c_{k,t+1})}{i - 1 + \sum_k z_{k,t} n_{k,t} + \tau} \delta_{\theta^*_{k,t+1}}(\cdot) + \sum_{k'=1}^{K_t} \frac{z_{k',t} n_{k',t} \mathbb{I}(k' \notin \mathcal{C}_{K_i,t+1})}{i - 1 + \sum_k z_{k,t} n_{k,t} + \tau} p(\cdot | \theta^*_{k',t}) \\ & + \frac{\tau}{i - 1 + \sum_k z_{k,t} n_{k,t} + \tau} G_0, \end{aligned} \tag{4}$$

where $\mathcal{C}_{K_i,t+1}$ denotes the set of values been taken by $c_{1,t+1}, \ldots, c_{K_i,t+1}$ (i.e., inherited dishes from time t that has been ordered up to the $(i-1)$-th customer), and $\mathbb{I}(k' \notin \mathcal{C}_{K_i,t+1})$ is an indicator function indicating that the k'-th dish from time t is not yet ordered at time $t+1$. I believe this is still a Dirichlet measure. The stick-breaking construction makes this more explicit.

Given $\theta_{i,t+1}$, which is a random measure, the probability of a data item $x_{i,t+1}$ is given by the likelihood function $F(x_{i,t+1} | \theta_{i,t+1})$.

2.2 Infinite limit of conditional finite mixture models

We introduce auxiliary index variable $Y_{i,t+1}$ to denote the index of distinctive value taken by $\theta_{i,t+1}$ (i.e., $\theta^*_{y_{i,t+1},t+1}$, $y_{i,t+1} \in \{1, 2, \ldots, L\}$), and we assume that the indexes of the inherited mixture components are the same as their original indexes in the previous time point. Define $Y_{i,t+1}$ as a multinomial variable with parameter $\boldsymbol{\pi}_{t+1}$.

$$\begin{aligned}
\boldsymbol{\pi}_{t+1}|\tau, \{n_{k,t}, z_{k,t}\}_{k=1}^{K_t} &\sim \mathrm{Dir}(\frac{\tau}{L} + z_{1,t}n_{1,t}, \ldots, \frac{\tau}{L} + z_{K_t,t}n_{K_t,t}, \frac{\tau}{L}, \ldots, \frac{\tau}{L}) \\
y_{i,t+1}|\boldsymbol{\pi}_{t+1} &\sim \mathrm{Multinomial}(\boldsymbol{\pi}_{t+1}) \\
\theta_{k,t+1}|G_0, \theta_{k,t}, z_{k,t} &\sim G_0^{1-z_{k,t}}(\cdot) \times p(\cdot|\theta_{k,t})^{z_{k,t}} \\
x_{i,t+1}|y_{i,t+1}, \{\theta_{k,t+1}\}_{k=1}^{L} &\sim F(\cdot|\theta_{y_{i,t+1},t+1}). \qquad (5)
\end{aligned}$$

Integrating out $\boldsymbol{\pi}_{t+1}$ and let L go to infinity, we find that the conditional probability defining the prior for $y_{i,t+1}$ reach the following limit:

$$\begin{aligned}
& y_{i,t+1}|y_{1,t+1}, \ldots, y_{i-1,t+1}, \{n_{k,t}, z_{k,t}\}_{k=1}^{K_t}, G_0, \tau \\
\sim & \sum_{k=1}^{K_i} \frac{n_{k,t+1} + \sum_{k'=1}^{K_t} z_{k',t}n_{k',t}\delta_{k'}(\cdot)}{i - 1 + \sum_k z_{k,t}n_{k,t} + \tau}\delta_k(\cdot) + \frac{\tau}{i - 1 + \sum_k z_{k,t}n_{k,t} + \tau}, \qquad (6)
\end{aligned}$$

which leads to the same prior for θ's as the one defined by the afore-described dynamic CRP.

2.3 Stick-breaking construction

Under a stick-breaking construction, a Dirichlet process can be expressed by the following formula:

$$p(\cdot) = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}(\cdot)$$

where $\theta_i \sim G_0(\cdot)$ denotes a discrete random measure, $\pi_i = \gamma_i \prod_{i'=1}^{i-1}(1-\gamma_{i'})$ denote the weight of the measure and $\gamma_{i'} = Beta(1, \tau)$. For each sample ψ_n from $p(\cdot)$, let $y_n \in \{1, 2, \ldots\}$ denote the index of the atom that ψ_n equals to, i.e., $\psi_n = \theta_{y_n}$. It is easily seen that y_n admits a infinite dimensional multinomial distribution represented by the aforementioned stick-breaking construction:

$$p(y) = \sum_{i=1}^{\infty} \pi_i \delta_i(y). \qquad (7)$$

Suppose that we have sampled N instances of ψ_n, which contains K unique θ_k, each with occupancy number n_k. Now we suppose that our subsequent samples are generated from a posterior probability distribution of the y_{N+1}, given the stick-breaking prior and the previous samples of y_n's. In the following we derive a stick-breaking construction of this posterior. First let's consider the posterior probability of $y_{N+1} = k$,

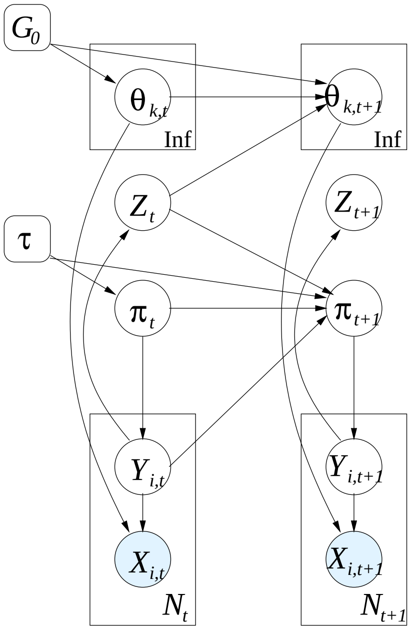

Figure 3: The graphical model of the stick-breaking representation.

where $1 \le k \le K$:

$$\begin{aligned}
& p(y_{N+1} = k | \pi_1, \ldots, \pi_K; y_1, \ldots, y_N) \\
\propto \ & p(y_{N+1} = k | \pi_1(\gamma), \ldots, \pi_K(\gamma); n_1, \ldots, n_K) p(\gamma) \\
\propto \ & p(y_{N+1} = k, n_1, \ldots, n_{k-1}, n_k, n_{k+1}, \ldots, n_K | \pi(\gamma)) p(\gamma) \\
= \ & \prod_{i=1}^{k-1} \pi_i^{n_i} \times \pi_k^{n_k+1} \times \prod_{i=k+1}^{K} \pi_i^{n_i} \times \prod_i \theta_i^0 (1-\theta_i)^{\tau-1} \\
= \ & (1-\gamma_1)^{\tau-1} \prod_{i=1}^{k-1} \gamma_i^{n_i} (1-\gamma_i)^{\sum_{j=i+1}^K n_j + \tau} \times \gamma_k^{n_k+1} (1-\gamma_k)^{\sum_{j=k+1}^K n_j + \tau - 1} \prod_{i=k+1}^{K} \gamma_i^{n_i} (1-\gamma_i)^{\sum_{j=i+1}^K n_j + \tau - 1} \\
= \ & \gamma_k (1-\gamma_{k-1}) \ldots (1-\gamma_1) \times (1-\gamma_1)^{\tau-1} \prod_{i=1}^{K} (\gamma_i)^{n_i} (1-\gamma_i)^{\sum_{j=i+1}^K n_j + \tau - 1} \\
= \ & \pi_i(\gamma_1, \ldots, \gamma_k) \prod_i Beta(\gamma_i | 1 + n_i, \sum_{j=i+1}^{K} n_j + \tau) \qquad (8)
\end{aligned}$$

It is easy to show that when $k > K$,

$$p(y_{N+1} = k | \pi_1, \ldots, \pi_K; y_1, \ldots, y_N) \propto \pi_i(\gamma_1, \ldots, \gamma_k) \prod_{i=1}^{K} Beta(\gamma_i | 1 + n_i, \sum_{j=i+1}^{K} n_j + \tau) \prod_{i=K+1}^{k} Beta(\gamma_i | 1, \tau).$$

Thus, given $\pi_1, \ldots, \pi_K$ and $y_1, \ldots, y_N$, the posterior distribution of y has the following stick-breaking construction:

$$p(y) = \sum_{i=1}^{\infty} \pi_i \delta_i(y). \qquad (9)$$

where $\pi_i = \gamma_i \prod_{i'=1}^{i-1} (1-\gamma_{i'})$ denote the weight of the random measure $theta_i$, $\gamma_{i \le K} = Beta(1+n_i, \sum_{j=i+1}^{K} n_j +$

τ), and $\gamma_{i>K} = Beta(1, \tau)$. Assuming no transformation of the discrete random measure θ_i, we will still have $\theta_i \sim G_0(\cdot)$. Thus we complete the stick-breaking representation of the posterior of a standard DP.

Using this construction to the DNBM, given extinction indicator $z_1, \ldots, z_K$ for $\pi_1, \ldots, \pi_K$, we have sieved occupancy counts $\{n'_k = z_k n_k \ : \ k = 1, \ldots, K\}$. Thus the stick-breaking construction for the distribution of π_i remains the formally the same, except that in the Beta prior the n_i's are replaced by n'_i's. The process for topic evolution over time translates to make the original base measure G_0 into a conditional base measure.

$$\theta_{k,t+1}|G_0, \theta_{k,t}, z_{k,t} \quad \sim \quad G_0^{1-z_{k,t}}(\cdot) \times p(\cdot|\theta_{k,t})^{z_{k,t}}. \tag{10}$$

Note that to simplify the definition, we introduced $z_k = 0$ for all $k > K$, so that the above formula apply to all k's.

It is easy to show that by taking the infinite limit of the condition finite mixture model, we can also get the stick-breaking construction described above.

3 A nonparametric Bayesian state-space model for sequence of Gaussian mixtures

Given different choices of the likelihood model $F(\cdot|\theta)$ for data, the base measure $G_0(\cdot)$ for the initial prior distribution of the mixture components, and the transition model $T(\theta_t|\theta_{t-1})$ for the evolution of mixture components, the dynamic nonparametric Bayesian model described above readily applies to a wide range of dynamic systems involving non-trivial temporal/spatial behaviors, such as multiple birth-death processes of subpopulations of samples in the system, temporal/spatial evolution of different subpopulations, etc..

Now we demonstrate such an application in the context of Bayesian density estimation of times series data from a dynamically evolving mixture. Specifically, we assume that $F(\cdot|\theta_t^{(k)})$ is a multivariate normal distribution for data from mixture component k at time t, $T(\theta_t^{(k)}|\theta_{t-1}^{(k)})$ is a linear dynamic model defined on the contiguous series of evolving parameters of a particular components k in the dynamic system. At each time t, data $\mathbf{x} = \{x_i\}$ are sampled from an infinite mixture of $\{\theta_t^{(k)}\}$ whose components are either inherited from $t-1$ or newly emerged at time t. The reason that we are interested in this dynamic Gaussian mixture model is not only because of its popularity and algebraic manipulability, but more because of its direct relevance to a number of real world application of our interest. For example, it readily defines a dynamic log-normal model for topic-specific word frequencies which, together with the dynamic DP model from topic popularities, leads to a dynamic extensions of the LDA model for time series of document corporas. It can also be used to define dynamic mixtures models for image pixels for tracking and detecting objects from video streams.

To proceed, we need to specify the prior G_0 for $\theta = \{\mu, W\}$ (to simplify notation, here and in the sequel we omit the superscript "(k)" that indexes the relevant mixture component when we describe properties or formulas applied to all components), the mean and precision matrix of the multivariate Gaussian random variables. A convenient form is the normal-Wishart conjugate to the normal sampling model. Thus, under $G_0(\cdot)$, we assume that $\mu \sim \mathcal{N}(\nu, (\alpha_\mu W)^{-1})$ is a multivariate-normal with mean ν and precision matrix $\alpha_\mu W$; and we assume that $W = \Sigma^{-1} \sim \text{Wishart}(T, \alpha_w)$:

$$\begin{aligned} p(W|T, \alpha_w) &= c(n, \alpha_w)|T|^{\alpha_w/2}|W|^{(\alpha_w - d - 1)/2} \exp\{\frac{1}{2}\text{tr}\{TW\}\} \\ &\equiv \text{Wishart}(W|\alpha_w, T), \end{aligned} \tag{11}$$

where T is a positive definite scale matrix, α_w is the degree of freedom of the normal-Wishart, d is the dimensionality of the precision matrix, and $c(d, \alpha_w)$ is a normalization constant given by

$$c(d, \alpha_w) = \Big[2^{\frac{\alpha_w d}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\big(\frac{\alpha_w + 1 - i}{2}\big)\Big]^{-1}. \tag{12}$$

For now, we assume that the prior parameters $\alpha_w, \alpha_\mu, T, \nu$ are specified.

Let $F_{G_0}(\cdot) = \int p(\cdot|\theta)G_0(\theta)d\theta$ represent the (marginal) likelihood of a single sample from a normal distribution under the normal-Wishart prior and given no other samples, it can be shown that the likelihood of a single data point x_i is a d-dimensional multivariate Student-t distribution with $\gamma = \alpha_w - d + 1$ degree of freedom, and having mode ν and precision matrix $\Upsilon = \frac{\alpha_\mu(\alpha_w-d+1)}{\alpha_\mu+1}T^{-1}$:

$$F_{G_0}(x_i) = \frac{\Gamma((\alpha_w+1)/2)}{(\alpha_w-d+1)/2((\alpha_w-d+1)\pi)^{d/2}}|\Upsilon|^{1/2}\Big(1+\frac{1}{\alpha_w-d+1}(x_i-\nu)^T\Upsilon(x_i-\nu)\Big)^{(\alpha_w+1)/2} \quad (13)$$

Computing this quantity can be difficult and one can approximate it by a multivariate-normal distribution having the same mean and covariance as the multivariate t-distribution (i.e., a moment matching approximation):

$$F_{G_0}(x_i) \cong \mathcal{N}(x_i|\mu_0, W_0^{-1}), \quad (14)$$

where $\mu_0 = \nu$ and the covariance matrix $W_0^{-1} = \frac{\gamma-2}{\gamma}\Upsilon^{-1} = \frac{\alpha_\mu+1}{\alpha_\mu(\alpha_w-d-1)}T$.

Due to conjugacy, the posterior distribution of $\{\mu, W\}$ under normal-Wishart prior and given observations $\mathbf{x}_{\setminus i}$ (say, containing M data points) is still a normal-Wishart. Specifically, let $\bar{x}$ denote the sample mean and S denote the sample covariance of data $\mathbf{x}_{\setminus i}$, we have $\mu \sim \mathcal{N}(\nu', (\alpha'_\mu W)^{-1})$ with updated mean vector $\nu' = \frac{\alpha_\mu\nu+M\bar{x}}{\alpha_\mu+M}$ and scale parameter $\alpha'_\mu = \alpha_\mu + M$, and $W \sim \text{Wishart}(\alpha'_w, T')$ with updated degree of freedom $\alpha'_w = \alpha_w + M$ and scale matrix $T' = T + MS + \frac{\alpha_\mu M}{\alpha_\mu+M}(\nu-\bar{x})(\nu-\bar{x})^T$. We denote this posterior distribution by $H_{G_0}(\cdot|\mathbf{x}_{\setminus i})$ for later reference. Following Eqs. (13) and (14), the posterior likelihood of a single data (given $\mathbf{x}_{\setminus i}$) is also a multivariate t-distribution and again can be approximated by a multivariate-normal distribution having the mean $\mu_0 = \nu'$ and covariance $W_0^{-1} = \frac{\alpha'_\mu+1}{\alpha'_\mu(\alpha'_w-d-1)}T'$.

$$H_{G_0}(x_i|\mathbf{x}_{-i}) \cong \mathcal{N}(x_i|\mu_0, W_0^{-1}). \quad (15)$$

Note that the normal-Wishart distribution only defines the prior for $\{\mu, W\}$ of a component when it is first instantiated. Now we briefly describe the stochastic dynamic model for a contiguous series of normal parameters, $\{(\mu_{t_1}, \Sigma_{t_1}^{-1}), \ldots, (\mu_{t_L}, \Sigma_{t_L}^{-1})\}$, of a normal component which is instantiated at time t_1 from the normal-Wishart base measure, and contiguously inherited until time t_L. We define the following state space model for data $\{x_{i,t}\}$:

$$\begin{aligned} \mu_{t+1} &= A\mu_t + Gw_t \quad (16) \\ x_{i,t} &= C\mu_t + v_t, \forall i, \quad (17) \end{aligned}$$

where $w_t \sim \mathcal{N}(0, Q)$ represents normal transition noise with zero mean and covariance matrix Q; and $v_t \sim \mathcal{N}(0, \Sigma_t^{-1})$ represents normal observation noise. Essentially, this is a Bayesian SMM in which the backbone mean vector μ_t's follow a linear dynamic model with initial distribution $\mathcal{N}(\nu, \alpha_\mu W)$ and transition dynamic Eq. (16) (where, for simplicity, we let $A = I$ so that the transition is a random walk), and the emission precision metrics $\Sigma_t^{-1} \sim \text{Wishart}(\alpha_w, T)$ (It is possible to construct a time-specific Wishart, e.g., defined by the posterior resulted from samples in the previous time point, but for simplicity, we deffer such elaborations). Overall, for $t > 0$, the prior distribution of $\{(\mu_t, \Sigma_t^{-1})\}$ is NOT a canonical normal-Wishart, but a product of a Gaussian defined by the LDS and a plain Wishart, and is therefore not conjugate prior. Note that in our setting, the output of our SSM at each time is not a single data point $x_{i,t}$, but a set of data points $\mathbf{x}_t$. It is well known that under an SSM, the posterior distribution of the centroid μ_t given the entire observed sequence and fixed precision metrics $\{\Sigma_t^{-1}\}$ is still a normal distribution, of which the mean and covariance matrix can be readily estimated using the Kalman filtering (KF) and Rauch-Tung-Striebel (RTS) smoothing algorithms. Here we give the modified Kalman filter "measurement-update" equations that take into account multiple rather than single output data points [1]. The RTS equations and the "time-update"

[1] This can be derived using the factor that the posterior mean and covariance matrix of the mean of a normal distribution $\mathcal{N}(\mu, \Sigma)$ given data $\mathbf{x}$ and prior of the mean $\mathcal{N}(\mu_0, \Sigma_0)$ is:

$$\Sigma_p = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}, \qquad \mu_p = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}(n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0) \quad (18)$$

equations of KF is the identical to the standard case for single output and hence omitted:

$$\begin{aligned}\hat{\mu}_{t+1|t+1} &= \hat{\mu}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1|t}C^T + \Sigma_t/n)^{-1}(\bar{x}_{t+1} - C\hat{\mu}_{t+1|t}) \\ P_{t+1|t+1} &= P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1|t}C^T + \Sigma_t/n)^{-1}CP_{t+1|t}, \end{aligned} \tag{19}$$

where $\hat{\mu}_{t+1|t+1}$ denotes the mean of μ_{t+1} conditioned on the partial sequence $x_1, \ldots, x_{t+1}$; $\hat{\mu}_{t+1|t}$ denotes the mean of μ_{t+1} conditioned on the partial sequence $x_1, \ldots, x_t$; $P_{t+1|t+1}$ and $P_{t+1|t}$ are the covariance matrices of μ_{t+1} conditioned of partial sequences $x_1, \ldots, x_{t+1}$ and $x_1, \ldots, x_t$, respectively; and $\bar{x}_{t+1}$ denote the sample mean of observations at time $t+1$ (from this SSM).

Now let $\nu_t = \hat{\mu}_{t|T}$ denote the mean vector and $\Phi_t = P_{t|T}$ denote the covariance matrix of μ_t resulted from the SSM, i.e., $\mu_t \sim \mathcal{N}(\nu_t, \Phi_t)$. We also know that $\Sigma_t^{-1} \sim \text{Wishart}(\alpha_w, T)$. Thus the marginal likelihood of a datum $x_{i,t}$ is:

$$F_{\text{SSM}}(x_{i,t}) = \int p(x_{i,t}|\mu_t, \Sigma_t^{-1})dp(\mu_t|\nu_t, \Phi_t)dp(\Sigma_t^{-1}|\alpha_w, T). \tag{20}$$

Note that due to non-conjugacy of the prior, this term can not be computed in close form. In stead, we can approximate it using MCMC:

$$\begin{aligned}F_{\text{SSM}}(x_{i,t}) &\cong \int \frac{1}{M}\sum_{m=1}^{M} p(x_{i,t}|\mu_t, \Sigma_{t,m}^{-1})dp(\mu_t|\nu_t, \Phi_t) \\ &= \frac{1}{M}\sum_{m=1}^{M} p(x_{i,t}|\nu_t, (\Sigma_{t,m} + \Phi_t)^{-1}) \end{aligned} \tag{21}$$

where $\Sigma_{t,m}^{-1} \sim \text{Wishart}(\alpha_w, T)$.

Similarly

$$\begin{aligned}H_{\text{SSM}}(x_{i,t}|\mathbf{x}_{-i,t}^{(k)}) &\cong \int \frac{1}{M}\sum_{m=1}^{M} p(x_{i,t}|\mu_t, \Sigma_{t,m}^{-1})dp(\mu_t|\nu_t, \Phi_t, \mathbf{x}_{-i,t}^{(k)}) \\ &= \frac{1}{M}\sum_{m=1}^{M} p(x_{i,t}|\nu'_t, (\Sigma_{t,m} + \Phi'_t)^{-1}) \end{aligned} \tag{22}$$

where $\Sigma_{t,m}^{-1} \sim \text{Wishart}(\alpha_w, T, \mathbf{x}_{-i,t}^{(k)}, \mu_t) = \text{Wishart}(\alpha_w + n_{t,k}, T + \sum_{j=1}^{n_{t,k}}(x_{j,t}^{(k)} - \mu_t^{(k)})(x_{j,t}^{(k)} - \mu_t^{(k)})^T)$. Note that the posterior of the Wishart is still a Wishart, which is dependent not only on $\mathbf{x}_{-i,t}^{(k)}$, but also on a sample of $\mu_t^{(k)}$. Posterior mean and variance of the Gaussian ν'_t and Φ'_m can be estimated from the footnote. Typically, for efficiency, we set $M = 1$, i.e., one sample of $\Sigma_{t,m}$. So essentially, given Σ_t^{-1}, we can compute the posterior of $\mu_t^{(k)}$ under the LDS; given a sample of $\mu_t^{(k)}$, we can also sample a Σ_t^{-1} from its posterior— a Gibbs-like procedure.

Note the $\Sigma_{t,m}^{-1}$ at every time point is conditionally independent given the Wishart prior. So when drawing $\Sigma_{t,m}^{-1}$ for a time point we do not need to consider other samples of x in other time points. Thus the likelihood of data of an inherited class k under an SSM for the mean and Wishart prior for the covariance is: $F_{\text{SSM}}(\mathbf{x}_t^{(k)}) = \prod_i F_{\text{SSM}}(x_{i,t}^{(k)})$, whereas the likelihood of data of an new-born class under a standard normal-Wishart prior is: $F_{\text{G}_0}(\mathbf{x}_t^{(k)}) = \prod_i F_{\text{G}_0}(x_{i,t}^{(k)})$.

4 A Gibbs sampling algorithm

Now we precede to describe a Gibbs sampling scheme for this model.

To encourage efficient mixing, it is useful to associate each data point $x_{i,t}$ (i.e., a random customer) with a auxiliary indicator RV $y_{i,t}$, denoting the index of the unique value of the random measure $\theta_{i,t}$ used to define the likelihood of $x_{i,t}$ (i.e., the index of the table sit by customer i at time t). With this auxiliary RV, every time the unique value of $\theta^*_{k,t}$ is resampled, all samples of $\theta_{i,t}$ pertaining to this value will have their value changed. In our dynamic CRP construction the value of $y_{i,t}$ depends on $y_{[-i],t}$, the value of mixture component indicators of all other samples at time t; $y_{\cdot,t-1}$, all such indicators at time at time $t-1$; $\mathbf{z}_{\cdot,t}$, the survival indicators at time t of all components from the previous time. It also directly influence the distribution of $x_{i,t}$, the sample whose distribution is defined by the draw of $\theta_{i,t}$, and $z_{\cdot,t+1}$ the distribution of the survival indicators of time t components at time $t+1$. Our sampling scheme is as follows:

- For each t, sample the inheritance indicators $\mathbf{z}_t$; and also the index of descendants of each inherited component. For each t, let $\mathbf{c}_t$ denote the vector bookkeeping the indices of the ancestors of components at $t-1$, if component k at time time has no ancestor, let $c_{k,t}=0$.
- Given $\mathbf{z}$, sample class indicator $\mathbf{y}$.
- Given class mean μ (from previous round) and class labels $\mathbf{y}$, sample the precision matrix W for each class at each time.
- Given the precision matrix W and and class labels, run Kalman filter to compute posterior for the class means at each time, sample class means μ accordingly.

Now we proceed to sample the survival RV $z_{k,t}$. RV $z_{k,t}$ is related to the occupancy pattern at time t by a regression function $p(\cdot|\mathbf{n}_t)$. It also directly influence (together with $\mathbf{n}_t$) the distribution of all $y_{i,t+1}$; and the distribution of all $\theta_{k,t+1}$. The Gibbs predictive distribution of $y_{i,t}$ can be written as follows:

$$\begin{aligned} p(z_{k,t}|\mathbf{n}_t,\boldsymbol{\theta}_t,\mathbf{z}_{-k,t},\boldsymbol{\theta}_{t+1},\mathbf{n}_{t+1},\mathbf{c}_{t+1}) &\propto p(z_{k,t}|\mathbf{n}_t)p(\mathbf{n}_{t+1},\boldsymbol{\theta}_{t+1}|z_{k,t},\mathbf{z}_{-k,t},\mathbf{n}_t,\boldsymbol{\theta}_t,\mathbf{c}_{t+1}) \\ &= p(z_{k,t}|\mathbf{n}_t)p(\mathbf{n}_{t+1}|z_{k,t},\mathbf{z}_{-k,t},\mathbf{n}_t)p(\boldsymbol{\theta}_{t+1}|z_{k,t},\mathbf{z}_{-k,t},\boldsymbol{\theta}_t,\mathbf{c}_{t+1}) \quad (23) \end{aligned}$$

The conditional probability of $\mathbf{n}_{t+1}$ and $\boldsymbol{\theta}_{t+1}$ can be computed as follows:

$$\begin{aligned} & p(\mathbf{n}_{t+1},\boldsymbol{\theta}_{t+1}|\mathbf{z}_t,\mathbf{n}_t,\boldsymbol{\theta}_t) \\ = & \int_{\boldsymbol{\pi}} \mathrm{Dir}(\boldsymbol{\pi}|\frac{\tau}{L_t}+\mathbf{z}_t \bullet \mathbf{n}_t) \prod_{l}^{L_{t+1}} \pi_l^{n_{l,t+1}} d\boldsymbol{\pi} \times \prod_{l}^{L_{t+1}} G_0(\theta_{l,t+1})^{\mathbf{1}(c_{l,t+1}=0)} H_{\mathrm{SSM}}(\theta_{l,t+1}|\theta_{c_{l,t+1},t})^{\mathbf{1}(c_{l,t+1}\neq 0)} \\ = & \frac{\Gamma(\tau+N_t(\mathbf{z}_t))}{\Gamma(\tau+N_t(\mathbf{z}_t)+N_{t+1})} \prod_{l=1}^{L_{t+1}} \frac{\Gamma(\tau/L_{t+1}+z_{l,t}n_{l,t}+n_{l,t+1})}{\Gamma(\tau/L_{t+1}+z_{l,t}n_{l,t})} G_0(\theta_{l,t+1})^{\mathbf{1}(c_{l,t+1}=0)} H_{\mathrm{SSM}}(\theta_{l,t+1}|\theta_{c_{l,t+1},t})^{\mathbf{1}(c_{l,t+1}\neq 0)} \end{aligned} \quad (24)$$

where $\mathbf{1}(\cdot)$ represents an indicator function of the true/false outcome of its argument, "$\bullet$" denote the Hadamard (i.e., elementwise) product of two vectors, L_t denote the total number of non-empty clusters at time t, and $N_t(\mathbf{z}_t)=\sum_{l'=1}^{L_t} z_{l',t}n_{l',t}$, which can understood as the "$\mathbf{z}_t$-sieved" total occupancy at time t. For convenience, we also define $N_t^{(k)}(\mathbf{z}_t)=\sum_{l'=1}^{L_t} z_{l',t}n_{l',t}-z_{k,t}n_{k,t}$ to be the "$\mathbf{z}_t$-sieved" total occupancy except for table k at time t.

Now back to the predictive distribution to $z_{k,t}$. Instead of sampling $z_{k,t}$ directly, we sample an auxiliary variable d which represents an indicator of the descendant of component k at time t. If $d_{k,t}=0$, then components k has no descendant at time $t+1$ and therefore $z_{k,t}=0$; if $d_{k,t}=l$, then component l at time $t+1$ is the descendant of component k at time t, and we can set $z_{k,t}=1$ and $c_{l,t+1}=k$. The proposal distribution of $d_{k,t}$ can be written as follows:

$$\begin{aligned} & p(d_{k,t}=l,l\neq 0|\mathbf{n}_t,\boldsymbol{\theta}_t,\mathbf{z}_{-k,t},\boldsymbol{\theta}_{t+1},\mathbf{n}_{t+1},\mathbf{c}_{t+1}) \\ = & C\times p(z_{k,t}=1|\mathbf{n}_t)p(\mathbf{n}_{t+1},\boldsymbol{\theta}_{t+1}|\mathbf{z}_t,\mathbf{n}_t,\boldsymbol{\theta}_t,\mathbf{c}_{-l,t+1},c_{l,t+1}=k,) \\ = & C\times\alpha_k \frac{\Gamma(\tau+N_t^{(k)}(z_t)+n_{k,t})}{\Gamma(\tau+N_t^{(k)}(\mathbf{z}_t)+n_{k,t}+N_{t+1})} \frac{\Gamma(\tau/L_{t+1}+n_{k,t}+n_{l,t+1})}{\Gamma(\tau/L_{t+1}+n_{k,t})} H_{\mathrm{SSM}}(\theta_{l,t+1}|\theta_{k,t}), \quad (25) \end{aligned}$$

where the normalization constant can be obtained by summing the unnormalized term above over all $l \in \mathcal{D}_k$, where $\mathcal{D}_k$ represents the set of all possible values l can take, including 0, the index of the current descendant of component k of time t, and the indices of those components at time $t+1$ currently having no ancestor:

$$\begin{aligned} C &= \sum_{l' \in \mathcal{D}_k} \alpha_k \frac{\Gamma(\tau + N_t^{(k)}(z_t) + n_{k,t})}{\Gamma(\tau + N_t^{(k)}(\mathbf{z}_t) + n_{k,t} + N_{t+1})} \frac{\Gamma(\tau/L_{t+1} + n_{k,t} + n_{l',t+1})}{\Gamma(\tau/L_{t+1} + n_{k,t})} H_{\text{SSM}}(\theta_{l',t+1}|\theta_{k,t}) \\ &\quad +(1-\alpha_k) \frac{\Gamma(\tau + N_t^{(k)}(z_t))}{\Gamma(\tau + N_t^{(k)}(\mathbf{z}_t) + N_{t+1})} \frac{\Gamma(\tau/L_{t+1} + n_{l,t+1})}{\Gamma(\tau/L_{t+1})} G_0(\theta_{l,t+1}) \end{aligned} \tag{26}$$

Note that $\theta_{k,t}$ and $\theta_{l,t+1}$ can be integrated out under both the base measure and the SMM.

$$\begin{aligned} & p(d_{k,t} = l|\mathbf{z}_{-k,t}, \mathbf{n}_t, \mathbf{y}, \mathbf{x}, \mathbf{c}) \\ = \ & C \times p(z_{k,t} = 1|\mathbf{n}_t) p(\mathbf{n}_{t+1}|\mathbf{n}_t, z_{k,t} = 1, \mathbf{c}_{-l,t+1}, c_{l,k+1} = k) p(\mathbf{x}_{t+1}^{(k)}|\mathbf{y}, \mathbf{x} \setminus \mathbf{x}_{t+1}^{(k)}, \mathbf{c}_{-l,t+1}, c_{l,k+1} = k, \mathbf{c}_{-(t+1)}) \\ = \ & \frac{F_{\text{SSM}}(\mathbf{x}_{t+1}^{(l)}|\hat{\theta}_{t+1|T}, P_{t+1|T})}{\sum_{l' \in \mathcal{D}_k} F_{\text{SSM}}(\mathbf{x}_{t+1}^{(l')}|\hat{\theta}_{t+1|T}, P_{t+1|T}) + \frac{1-\alpha_k(\mathbf{n}_t)}{\alpha_k(\mathbf{n}_t)} r(\mathbf{n}_t, \mathbf{n}_{t+1}, \mathbf{z}_t) F_{G_0}(\mathbf{x}_{t+1}^{(l)})} \end{aligned} \tag{27}$$

where $\mathbf{x}_{t+1}^{(k)}$ denotes the subset of $\mathbf{x}_{t+1}$ belonging to component k; $F_{\text{SSM}}(\mathbf{x}_{t+1}^{(k)}|\hat{\theta}_{t+1|T}, P_{t+1|T})$ denotes the likelihood of data of class k under a Gaussian prior of the parameters of the likelihood model, which is estimated from the SMM model on component k; $F_{\text{G}_0}(\cdot)$ denote the marginal likelihood of data using the base measure as the prior of the likelihood model (both derived in the previous section); and $r(\mathbf{n}_t, \mathbf{n}_{t+1}, \mathbf{z}_t)$ is a function determined by the occupancy pattern under possible values of $z_{k,t}$ and $d_{k,t}$:

$$\begin{aligned} r(\mathbf{n}_t, \mathbf{n}_{t+1}, \mathbf{z}_t) &= \frac{\Gamma(\tau + N_t^{(k)}(\mathbf{z}_t) + n_{k,t} + N_{t+1}) \Gamma(\tau/L_{t+1} + n_{k,t}) \Gamma(\tau + N_t^{(k)}(z_t)) \Gamma(\tau/L_{t+1} + n_{l,t+1})}{\Gamma(\tau + N_t^{(k)}(z_t) + n_{k,t}) \Gamma(\tau/L_{t+1} + n_{k,t} + n_{l,t+1}) \Gamma(\tau + N_t^{(k)}(\mathbf{z}_t) + N_{t+1}) \Gamma(\tau/L_{t+1})} \\ &= \frac{\Phi_{\tau+N_t^{(k)}+N_{t+1}}^{\tau+N_t^{(k)}+N_{t+1}+n_{k,t}-1}}{\Phi_{\tau+N_t^{(k)}}^{\tau+N_t^{(k)}+n_{k,t}-1}} \times \frac{\Phi_{\tau/L_{t+1}}^{\tau/L_{t+1}+n_{l,t+1}-1}}{\Phi_{\tau/L_{t+1}+n_{k,t}}^{\tau/L_{t+1}+n_{k,t}+n_{l,t+1}-1}}, \end{aligned} \tag{28}$$

where $\Phi_b^a, a > b$ represents a partial factorial $a \times (a-1) \times, \ldots, \times (b+1) \times b$.

The Gibbs predictive distribution of $y_{i,t}$ can be written as follows:

$$\begin{aligned} & p(y_{i,t}|\mathbf{y}_{[\setminus i],t}, \mathbf{y}_{t-1}, \mathbf{z}_{t-1}, x_{i,t}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \mathbf{c}_t) \\ \propto \ & p(y_{i,t}|\mathbf{y}_{[\setminus i],t}, \mathbf{y}_{t-1}, \mathbf{z}_{t-1}) p(x_{i,t}|y_{i,t}, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t, \mathbf{c}_t, G_0) \\ = \ & \begin{cases} \frac{n_{[\setminus i],k,t} + n_{l,t-1}}{N_{t-1}+i-1+\tau} p(x_{i,t}|\theta_{k,t}) \quad \text{if } \ k = y_{i'} \text{ for some } i' \neq i, \text{and } c_{k,t} = l. \\ \frac{n_{l,t-1}}{N_{t-1}+i-1+\tau} \int p(x_{i,t}|\theta_{k,t}) dH_{\text{SSM}}(\theta_{k,t}|\theta_{l,t-1}) \quad \text{if } \ k \neq y_{i'} \text{ for all } i' \neq i, \text{and } c_{k,t} = l. \\ \frac{n_{[\setminus i],k,t}}{N_{t-1}+i-1+\tau} p(x_{i,t}|\theta_{k,t}) \quad \text{if } \ k \neq y_{i'} \text{ for some } i' \neq i, \text{and } c_{k,t} = 0. \\ \frac{\tau}{N_{t-1}+i-1+\tau} \int p(x_{i,t}|\theta_t) dG_0(\theta) \quad \text{if } \ k \neq y_{i'} \text{ for all } i' \neq i, \text{and } c_{k,t} = 0. \end{cases} \end{aligned} \tag{29}$$

In a context of using a conjugate G_0 prior for newly instantiated and an SSM for surviving θ_t, we could in some cases integrate the θ to speed up mixing.

$$\begin{aligned} & p(y_{i,t}|\mathbf{y}_{[\setminus i],t}, \mathbf{y}_{t-1}, \mathbf{z}_{t-1}, x_{i,t}) \\ = \ & \begin{cases} b \frac{n_{[\setminus i],k,t} + n_{l,t-1}}{N_{t-1}+N_t-1+\tau} F_{\text{SSM}}(x_i|\mathbf{x}_{-i,t}^{(k)}) \quad \text{if } \ k = y_{i'} \text{ for some } i' \neq i, \text{and } c_{k,t} = 1. \\ b \frac{n_{k,t-1}}{N_{t-1}+N_t-1+\tau} F_{\text{SSM}}(x_i) \quad \text{if } \ k \neq y_{i'} \text{ for all } i' \neq i, \text{and } c_{k,t} = 1. \\ b \frac{n_{[\setminus i],k,t}(-i)}{N_{t-1}+N_t-1+\tau} (1-\alpha_k) F_{G_0}(x_i|\mathbf{x}_{-i,t}^{(k)}) \quad \text{if } \ k \neq y_{i'} \text{ for some } i' \neq i, \text{and } c_{k,t} = 0. \\ b \frac{\tau}{N_{t-1}+N_t-1+\tau} (1-\alpha_k) F_{G_0}(x_i) \quad \text{if } \ k \neq y_{i'} \text{ for all } i' \neq i, \text{and } c_{k,t} = 0. \end{cases} \end{aligned} \tag{30}$$

where b is a normalization constant, and $H_{\text{SSM}}(\theta_k|\cdot)$ and $H_{\text{SSM}}(\theta_k|\cdot, \mathbf{x}_{\backslash i,t})$ are the posterior distribution of θ_k in an SMM of topic k over time given observations from times other than time t and all observations except $x_{i,t}$, respectively.

When Gibbs sampling for $y_{i,t}$ choose a value not equal to any other $y_{j,t}$ at time t, a value for $\theta^{(y_{i,t})}$ is chosen from $H_{G_0}(\cdot|x_{i,t})$, which is the posterior distribution of θ based on the prior G_0 and the single observation $x_{i,t}$. It will be used also as the initial hidden state of the SSM that govern temporal evolution of topic k.

Now given class mean μ (from previous round) and class labels $\mathbf{y}$, we can sample the precision matrix W for each class at each time from its posterior: $\text{Wishart}(\alpha_w + n_{t,k}, T + \sum_{j=1}^{n_{t,k}}(x_{j,t}^{(k)} - \mu_t^{(k)})(x_{j,t}^{(k)} - \mu_t^{(k)})^T)$.

Finally, given instantiations of $\mathbf{y}$ and $\mathbf{z}$, and the emission matrix $\Sigma_t^{(k)} = W^{(k)}$, inference over the mean parameter over time μ is equivalent to working with several independent state-space models, which can be solved in close-form using the Kalman filter and RTS algorithm on each chain of retained components, as discussed in the previous section.

Note that the Kalman filter and RTS algorithm is performed under a given covariance matrix for the emission. After this step, we re-sample $\mu_t^{(k)}$ for the next round.