Poisson-Gamma Dynamical Systems with Non-Stationary Transition Dynamics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bayesian methodologies for handling count-valued time series have gained prominence due to their ability to infer interpretable latent structures and to estimate uncertainties, and thus are especially suitable for dealing with *noisy* and *incomplete* count data. Among these Bayesian models, Poisson-Gamma Dynamical Systems (PGDSs) are proven to be effective in capturing the evolving dynamics underlying observed count sequences. However, the state-of-the-art PGDS still falls short in capturing the *time-varying* transition dynamics that are commonly observed in real-world count time series. To mitigate this limitation, a non-stationary PGDS is proposed to allow the underlying transition matrices to evolve over time, and the evolving transition matrices are modeled by the specifically-designed Dirichlet Markov chains. Leveraging Dirichlet-Multinomial-Beta data augmentation techniques, a fully-conjugate and efficient Gibbs sampler is developed to perform posterior simulation. Experiments show that, in comparison with related models, the proposed non-stationary PGDS achieves improved predictive performance due to its capacity to learn non-stationary dependency structure captured by the time-evolving transition matrices.

1 Introduction

In recent years, there has been an increasing interest in modeling count time series. For instance, some previous works [1, 2, 3] are concerned with how to learn the evolving topics behind text corpus (frequencies of words) over time. Some works [4, 5, 6, 7] try to predict global immigrant trends underlying international population movements. Count time series are often *overdispersed*, *sparse*, *high-dimensional*, and thus can not be well modeled by widely used dynamic models such as linear dynamical systems [8, 9]. Recently, many works [10, 11, 12, 13, 14, 15, 16] prefer to choose distributions of the gamma-Poisson family to build their hierarchical Bayesian models. In particular, these models enjoy strong explainability and can estimate uncertainty especially when the observations are *noisy* and *incomplete*. Among these works, Poisson-Gamma Dynamical Systems (PGDSs) [13] received a lot of attention because PGDS can learn how the latent dimensions excite each other to capture complicated dynamics in observed count series. For instance, a very inspiring research paper may motivate other researchers to publish papers on related topics [17]. The outbreak of COVID-19 in one state, may lead to the rapid rising of COVID-19 cases in the nearby states and vice versa [18]. In particular, PGDS can be efficiently learned with a tractable Gibbs sampling scheme via Poisson-Logarithmic data augmentation and marginalization technique [11]. Due to its strong flexibility, PGDS achieves better performance in predicting missing entities and future observations, compared with related models [9, 15].

Despite these advantages, PGDS still can not capture the time-varying transition dynamics underlying observed count sequences, which are commonly observed in real-world scenarios [19]. For instance,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

during the initial stage of the COVID-19 pandemic, the worldwide counts of infectious patients were significantly affected by various local policies, government interventions, and emergent events [20, 21, 22]. The cross transition dynamics among the different monitoring areas were also evolving as the corresponding policies and interventions changed over time. Hence, PGDS unavoidably makes a certain amount of approximation error in capturing the aforementioned non-stationary count time series, using a *time-invariant* transition kernel.

To mitigate this limitation, Non-Stationary Poisson-Gamma Dynamical Systems (NS-PGDSs), a novel kind of Poisson-gamma dynamical systems with non-stationary transition dynamics are developed. More specifically, NS-PGDS captures the evolving transition dynamics by the specifically-designed Dirichlet Markov chains. Via the Dirichlet-Multinomial-Beta data augmentation strategy, the Non-Stationary Poisson-Gamma Dynamical Systems can be inferred with a conjugate-yet-efficient Gibbs sampler. Our contributions are summarized as follows:

- We propose a Non-Stationary Poisson-Gamma Dynamical System (NS-PGDS), a novel Poisson-gamma dynamical system with time-evolving transition matrices that can well capture non-stationary transition dynamics underlying observed count series.
- Three Dirichlet Markov chains are dedicated to improving the flexibility and expressiveness of NS-PGDSs, for capturing the complex transition dynamics behind sequential count data.
- Fully-conjugate-yet-efficient Gibbs samplers are developed via Dirichlet-Multinomial-Beta augmentation techniques to perform posterior simulation for the proposed Dirichlet Markov chains.
- Extensive experiments are conducted on four real-world datasets, to evaluate the performance of the proposed NS-PGDS in predicting missing and future unseen observations. We also provide exploratory analysis to demonstrate the explainable latent structure inferred by the proposed NS-PGDS.

2 Preliminaries

Let $\boldsymbol{y}^{(t)} = \left[y_1^{(t)}, \cdots, y_V^{(t)}\right]^{\mathrm{T}} \in \mathbb{N}^V$ be a vector of nonnegative count valued observations at time t. To capture the latent dynamics underlying count sequences, some previous works [23, 24] model the observations as

$$\boldsymbol{y}^{(t)} = p\left(\boldsymbol{z}^{(t)}\right), \;\; \boldsymbol{z}^{(t)} = f^{-1}\left(\boldsymbol{x}^{(t)}\right),$$

where $p(\cdot)$ is the observation likelihood function, and $f(\cdot)$ is an invertible link function that maps the parameters of observation component to continuous-valued latent variables $\boldsymbol{x}^{(t)} \in \mathbb{R}^K$. The latent factor $\boldsymbol{x}^{(t)}$ evolves over time according to a linear dynamical system (LDS) given by $\boldsymbol{x}^{(t)} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}^{(t-1)}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{A}$ is the state transition matrix of size $K \times K$, and $\boldsymbol{\Lambda} = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_K\right)$ is the inverse covariance matrix with λ_k^{-1} determining the variance of k-th latent dimension. Han et al. [23] adopted the Extended Rank likelihood function to model count observations using LDS with time complexity $\mathcal{O}((K+V)^3)$, which prevents it from practical applications for analyzing large-scale count data.

Recently, Acharya et al. [15] and Schein et al. [13, 16] developed Poisson-gamma family models for sequential count observations. Gamma Process Dynamic Poisson Factor Analysis (GP-DPFA) [15] models count data as $y_v^{(t)} \sim \mathrm{Pois}(\sum_{k=1}^K \lambda_k \phi_{vk} \theta_k^{(t)})$, where $\theta_k^{(t)}$ represents the strength of k-th latent factor at time t, and ϕ_{vk} captures the involvement degree of k-th factor to v-th observed dimension. To ensure the model identifiability, we can impose a restriction as $\sum_v \phi_{vk} = 1$, and thus place a Dirichlet prior over $\boldsymbol{\phi_k} = [\phi_{1k}, \cdots, \phi_{Vk}]^T$ as $\boldsymbol{\phi_k} \sim \mathrm{Dir}\left(\epsilon_0, \cdots, \epsilon_0\right)$.

To capture the underlying dynamics, the latent factor $\theta_k^{(t)}$ evolves over time according to a gamma Markov chain as $\theta_k^{(t)} \sim \mathrm{Gam}(\theta_k^{(t-1)}, c_t)$, where c_t is the rate parameter of the gamma distribution to control the variance of the gamma Markov chains. Although GP-DPFA can well fit one-dimensional count sequences, it fails to learn how the latent dimensions interact with each other.

To address this concern, Schein et al. [13] developed Poisson-gamma dynamical systems to capture the underlying transition dynamics. In particular, $\theta_k^{(t)}$ evolves over time as $\theta_k^{(t)} \sim$

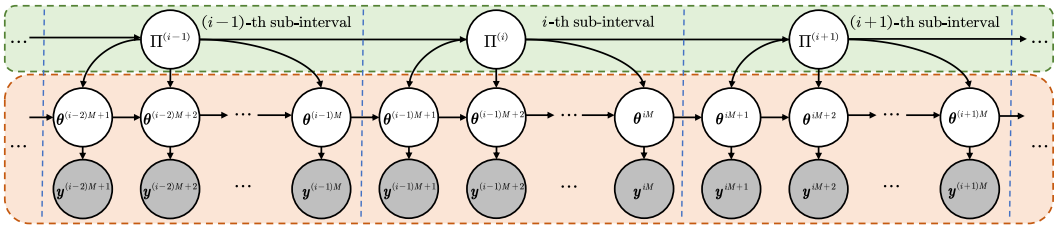

Figure 1: The graphical representation of the NS-PGDS. The time interval is divided into equally-spaced sub-intervals. Each sub-interval contains M time steps. The transition dynamics is stationary within a sub-interval. In particular, the transition matrices evolve over sub-intervals via Dirichlet Markov processes while latent factors evolve over time steps via Eq.(1).

$\mathrm{Gam}(\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}, \tau_0)$, where π_{kk_2} represents how k_2-th latent factor excites the k-th latent factor at next time step, and $\sum_{k=1}^{K} \pi_{kk_2} = 1$.

3 Non-Stationary Poisson-Gamma Dynamical Systems

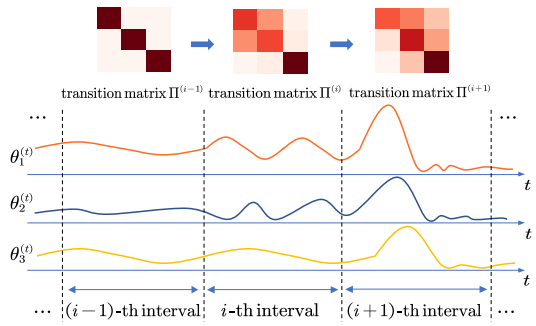

Figure 2: An example illustrates the Poisson-gamma dynamical systems with non-stationary transition kernels. The three gamma dynamic processes independently evolve over time during the $(i-1)$-th interval. During i-th interval, $\theta_1^{(t)}$ and $\theta_2^{(t)}$ gradually starts to interact with each other while $\theta_3^{(t)}$ remains independent to the other two dimensions. During $(i+1)$-th interval all the three latent components start to interact with each other.

Real-world count time sequences are often *non-stationary* because the external interventional environments are always changing over time. The stationary PGDS with a time-invariant transition kernel fails to capture such time-varying transition dynamics. For instance, the transition dynamics behind COVID-19 infectious processes are time-varying, and highly affected by various interventional policies. Hence, to mitigate this limitation, we model the count sequences as

$$y_v^{(t)} \sim \mathrm{Pois}\left(\delta^{(t)} \sum_{k=1}^{K} \phi_{vk} \theta_k^{(t)}\right),$$

in which, the latent factors are specified by

$$\theta_k^{(t)} \sim \mathrm{Gam}\left(\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{(t-1)} \theta_{k_2}^{(t-1)}, \tau_0\right), \tag{1}$$

where the multiplicative term $\delta^{(t)} \sim \mathrm{Gam}(\epsilon_0, \epsilon_0)$ and the transition matrices are time-varying as $\mathbf{\Pi}^{(t)} \equiv \left[\pi_{kk_2}^{(t)}\right]_{k,k_2=1}^{K}$. As shown in Figure 2, to model the time-varying transition dynamics, we assume the whole time interval can be divided into I equally-spaced sub-intervals. The transition kernel behind complicated dynamic counts is assumed to be *static* within each sub-interval, while evolving over sub-intervals, to capture non-stationary behaviours. In another word, the proposed model allows the latent factors to evolve over time steps while the transition matrices change over sub-intervals but assumed to be stationary within each sub-interval, as shown in Figure 1. In particular, we let each sub-interval contains M time steps, and the i-th interval contains time steps $\{t \mid t = (i-1)M + 1, \cdots, iM\}$. We define $i(t)$ as the function that maps time step t to its corresponding sub-interval.

Dirichlet-Dirichlet Markov processes. To capture how the underlying transition kernel smoothly evolves over sub-intervals, we first propose the Dirichlet-Dirichlet (Dir-Dir) Markov chain as

$$\boldsymbol{\pi}_k^{(i)} \mid \boldsymbol{\pi}_k^{(i-1)} \sim \mathrm{Dir}\left(\eta K \pi_{1k}^{(i-1)}, \cdots, \eta K \pi_{Kk}^{(i-1)}\right), \tag{2}$$

where $\boldsymbol{\pi}_k^{(i)}$ represents the k-th column of $\mathbf{\Pi}^{(i)}$, and the prior of the scaling parameter η is given by $\eta \sim \mathrm{Gam}(e_0, f_0)$.

The initial states are defined as $\theta_k^{(1)} \sim \mathrm{Gam}\left(\tau_0\nu_k, \tau_0\right)$. The prior for the transition kernel of the first sub-interval is given by $\boldsymbol{\pi}_k^{(1)} \sim \mathrm{Dir}\left(\nu_1\nu_k, \cdots, \xi\nu_k, \cdots, \nu_K\nu_k\right)$, where $\nu_k \sim \mathrm{Gam}(\frac{\gamma_0}{K}, \beta)$ and $\xi, \beta \sim \mathrm{Gam}\left(\epsilon_0, \epsilon_0\right)$. Note that the expectation and variance of the transition kernel at i-th sub-interval can be calculated as

$$\mathsf{E}\left[\boldsymbol{\pi}_k^{(i)} \mid \boldsymbol{\pi}_k^{(i-1)}\right] = \boldsymbol{\pi}_k^{(i-1)}, \quad \mathsf{Var}\left[\pi_{k_1k}^{(i)} \mid \boldsymbol{\pi}_k^{(i-1)}\right] = \frac{\pi_{k_1k}^{(i-1)}\left(1-\pi_{k_1k}^{(i-1)}\right)}{\eta K+1},$$

respectively. The transition dynamics of i-th sub-interval inherits the information of the previous sub-interval, and also adapts to the data observed in the current sub-interval. The scaling parameter η controls the variance of the transition matrices.

The prior specification defined in Eq.(2) by rescaling the transition matrix at the previous sub-interval allows the transition dynamics to change smoothly, and thus might be insufficient to capture the rapid changes observed in complicated dynamics. To further improve the flexibility of the transition structure, two modified Dirichlet Markov chains are studied to capture the correlation structure between the dimensions of the transition matrices over time.

Dirichlet-Gamma-Dirichlet Markov processes. We first introduce the Dirichlet-Gamma-Dirichlet (Dir-Gam-Dir) Markov chain to model the evolving transition matrices as

$$\begin{aligned}\boldsymbol{\pi}_k^{(i)} &\sim \mathrm{Dir}\left(\alpha_{1k}^{(i)}, \cdots, \alpha_{Kk}^{(i)}\right),\\ \alpha_{k_1k}^{(i)} &\sim \mathrm{Gam}\left(\gamma_k^{(i-1)}\textstyle\sum_{k_2=1}^{K}\psi_{kk_1k_2}^{(i-1)}\pi_{k_2k}^{(i-1)}, c_k^{(i)}\right),\end{aligned} \quad (3)$$

where we use $\psi_{kk_1k_2}^{(i-1)}$ to capture the mutation between two consecutive sub-intervals, and its prior is given by

$$\left(\psi_{k1k_2}^{(i-1)}, \cdots, \psi_{kKk_2}^{(i-1)}\right) \sim \mathrm{Dir}\left(\epsilon_0, \cdots, \epsilon_0\right),$$

and $\gamma_k^{(i)}, c_k^{(i)} \sim \mathrm{Gam}\left(\epsilon_0, \epsilon_0\right)$. Compared with the construction defined by Eq.(2), the expectation of Dirichlet-Gamma-Dirichlet Markov chain is

$$\mathsf{E}\left[\boldsymbol{\pi}_k^{(i)} \mid \boldsymbol{\pi}_k^{(i-1)}\right] = \boldsymbol{\Psi}_k^{(i-1)}\boldsymbol{\pi}_k^{(i-1)}.$$

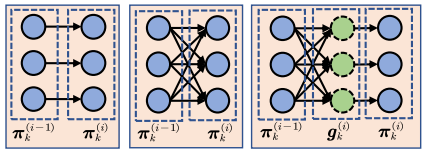

Figure 3: Diagrams of the proposed Dirichlet Markov constructions. (a) is the Dir-Dir construction. (b) is the Dir-Gam-Dir construction which takes mutation into account. (c) illustrates the PR-Gam-Dir construction which adopts Poisson randomized gamma distribution and can be equivalently represented as Eq.(5).

This construction takes interactions among components of columns into account. Hence it will dramatically improve the flexibility of our model and thus better fit more complicated dynamics, compared with Dir-Dir Markov chains that only yield smoothing transition dynamics.

Poisson-randomized-gamma-Dirichlet Markov processes. By leveraging the Poisson-randomized gamma distribution [25], we introduce another type of time-varying transition kernels, which also model the interactions among components like Dir-Gam-Dir construction but may induce different properties such as sparsity. The Poisson-randomized-gamma-Dirichlet (PR-Gam-Dir) Markov chain can be formulated as

$$\boldsymbol{\pi}_k^{(i)} \sim \mathrm{Dir}\left(\alpha_{1k}^{(i)}, \cdots, \alpha_{Kk}^{(i)}\right), \ \alpha_{k_1k}^{(i)} \sim \mathrm{RG1}\left(\epsilon^{\alpha}, \gamma_k^{(i-1)}\textstyle\sum_{k2=1}^{K}\psi_{kk_1k_2}^{(i-1)}\pi_{k_2k}^{(i-1)}, c_k^{(i)}\right), \quad (4)$$

where $\mathrm{RG1}\left(\cdot\right)$ denotes the randomized gamma distribution of the first type. Similarly, for $\psi_{kk_1k_2}^{(i-1)}$, $\gamma_k^{(i)}$, and $c_k^{(i)}$, the priors are given by

$$\left(\psi_{k1k_2}^{(i-1)}, \cdots, \psi_{kKk_2}^{(i-1)}\right) \sim \mathrm{Dir}\left(\epsilon_0, \cdots, \epsilon_0\right), \ \gamma_k^{(i)}, c_k^{(i)} \sim \mathrm{Gam}\left(\epsilon_0, \epsilon_0\right), \text{ respectively.}$$

The diagrams of three Dirichlet Markov constructions are shown in Figure 3.

4 Markov Chain Monte Carlo Inference

In this section, we present the Gibbs sampler for the proposed NS-PGDS. We only illustrate the key points of the derivation and the details can be found in the appendix.

Lemma 1 *If* $y \sim \mathrm{NB}\left(a, g\left(\zeta\right)\right)$ *and* $l \sim \mathrm{CRT}\left(y, a\right)$*, where* $\mathrm{NB}\left(\cdot\right)$ *refers to negative-binomial distribution,* $\mathrm{CRT}\left(\cdot\right)$ *represents Chinese restaurant table distribution [26], and* $g\left(z\right) = 1 - \exp\left(-z\right)$*. Then the joint distribution of* y *and* l *can be equivalently distributed as* $y \sim \mathrm{SumLog}\left(l, g\left(\zeta\right)\right)$ *and* $l \sim \mathrm{Pois}\left(a\zeta\right)$ *[11], i.e.*

$$\mathrm{NB}\left(y; a, g\left(\zeta\right)\right)\mathrm{CRT}\left(l; y, a\right) = \mathrm{SumLog}\left(y; l, g\left(\zeta\right)\right)\mathrm{Pois}\left(l; a\zeta\right),$$

where $\mathrm{SumLog}\left(l, g\left(\zeta\right)\right) = \sum_{i=1}^{l} x_i$ *and* $x_i \sim \mathrm{Log}\left(g\left(\zeta\right)\right)$ *are independently and identically logarithmic distributed random variables [27].*

Lemma 2 *Suppose* $\mathbf{n} = \left(n_1, \cdots, n_K\right)$ *and*

$$\mathbf{n} \mid n \sim \mathrm{DirMult}\left(n, r_1, \cdots, r_K\right),$$

where $\mathrm{DirMult}\left(\cdot\right)$ *refers to Dirichlet-multimonial distribution. We sample the augmented variable* $q \mid n \sim \mathrm{Beta}\left(n, r_\cdot\right)$*, where* $r_\cdot = \sum_{k=1}^{K} r_k$*. According to [28], conditioning on* q*, we have* $n_k \sim \mathrm{NB}\left(r_k, q\right)$.

Sampling $y_{vk}^{(t)}$ **:** Use the relationship between Poisson and multinomial distributions, we sample

$$\left(\left(y_{vk}^{(t)}\right)_{k=1}^{K} \mid -\right) \sim \mathrm{Mult}\left(y_v^{(t)}, \left(\frac{\phi_{vk}\theta_k^{(t)}}{\sum_{k=1}^{K}\phi_{vk}\theta_k^{(t)}}\right)_{k=1}^{K}\right).$$

Sampling $\boldsymbol{\phi_k}$ **:** Via Dirichlet-multinomial conjugacy, the posterior of $\boldsymbol{\phi_k}$ is

$$\left(\boldsymbol{\phi_k} \mid -\right) \sim \mathrm{Dir}\left(\epsilon_0 + \sum_{t=1}^{T} y_{1k}^{(t)}, \cdots, \epsilon_0 + \sum_{t=1}^{T} y_{Vk}^{(t)}\right).$$

Sampling $\theta_k^{(t)}$ **:** To sample from the posterior of $\theta_k^{(t)}$, we first sample the auxiliary variables. Setting $l_{\cdot k}^{(T+1)} = 0$ and $\zeta^{(T+1)} = 0$, we sample the augmented variables backwards from $t = T, \cdots, 2$,

$$\left(l_{k\cdot}^{(t)} \mid -\right) \sim \mathrm{CRT}\left(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)}, \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}\right),$$

$$\left(l_{k1}^{(t)}, \cdots, l_{kK}^{(t)} \mid -\right) \sim \mathrm{Mult}\left(l_{k\cdot}^{(t)}, \left(\frac{\pi_{k1}^{i(t-1)}\theta_1^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}}, \cdots, \frac{\pi_{kK}^{i(t-1)}\theta_K^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}}\right)\right).$$

Let us define $l_{\cdot k}^{(t)} = \sum_{k_1=1}^{K} l_{k_1 k}^{(t)}$ and $\zeta^{(t)} = \ln(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)})$. After sampling the auxiliary variables, then for $t = 1, \cdots, T$, by Poisson-gamma conjugacy, we obtain

$$\left(\theta_k^{(t)} \mid -\right) \sim \mathrm{Gam}\left(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)} + \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}, \tau_0 + \delta^{(t)} + \zeta^{(t+1)}\tau_0\right).$$

Sampling $\boldsymbol{\Pi}^{(i)}$ **:** We only illustrate Gibbs sampling algorithm for PR-Gam-Dir construction, sampling algorithms for other constructions can be found in the appendix. We define M as the length of each sub-interval, and I as the number of intervals. For $i = I, \cdots, 2$, because $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ and $(g_{\cdot 1k}^{(i+1)}, \cdots, g_{\cdot Kk}^{(i+1)})$ are multinomially distributed, where $l_{k_1 k}^{(i)} = \sum_{(i-1)M+1}^{iM} l_{k_1 k}^{(t)}$ refers to the summation of $l_{k_1 k}^{(t)}$ over i-th sub-interval and same notation for other variables. By the definition of Dirichlet-multinomial distribution and Lemma 2, defining $g_{k_1 k}^{(I+1)} = 0$, we sample the auxiliary variables as $(q_k^{(i)} \mid -) \sim \mathrm{Beta}(l_{\cdot k}^{(i)} + g_{\cdot k}^{(i+1)}, \alpha_{\cdot k}^{(i)})$, then we have $(l_{k_1 k}^{(i)} + g_{\cdot k_1 k}^{(i+1)}) \sim \mathrm{NB}(\alpha_{k_1 k}^{(i)}, q_k^{(i)})$. Then we further sample $(h_{k_1 k}^{(i)} \mid -) \sim \mathrm{CRT}(l_{k_1 k}^{(i)} + g_{\cdot k_1 k}^{(i+1)}, \alpha_{k_1 k}^{(i)})$. Via Lemma 1, we obtain $h_{k_1 k}^{(i)} \sim \mathrm{Pois}(-\alpha_{k_1 k}^{(i)}\ln(1 - q_k^{(i)}))$. For Dirichlet-Randomized-Gamma-Dirichlet Markov construction defined by Eq.(4), we can equivalently represent it as

$$\alpha_{k_1 k}^{(i)} \sim \mathrm{Gam}\left(g_{k_1 k}^{(i)} + \epsilon^{\alpha}, c_k^{(i)}\right),\ g_{k_1 k}^{(i)} = \mathrm{Pois}\left(\gamma^{(i-1)} \sum_{k2=1}^{K} \psi_{kk_1 k_2}^{(i-1)}\pi_{k_2 k}^{(i-1)}\right). \tag{5}$$

We define $\lambda_{k_1 k}^{(i-1)} \triangleq \gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{kk_1 k_2}^{(i-1)}\pi_{k_2 k}^{(i-1)}$ for notation conciseness. By Poisson-gamma conjugacy, we have $(\alpha_{k_1 k}^{(i)} \mid -) \sim \mathrm{Gam}(g_{k_1 k}^{(i)} + \epsilon^{\alpha} + h_{k_1 k}^{(i)}, c_k^{(i)} - \ln(1 - q_k^{(i)}))$. If $\epsilon^{\alpha} > 0$, we can

sample the posterior of $g_{k_1k}^{(i)}$ via $(g_{k_1k}^{(i)} \mid -) \sim \text{Bessel}(\epsilon^\alpha - 1, 2\sqrt{\alpha_{k_1k}^{(i)} c_k^{(i)} \lambda_{k_1k}^{(i-1)}})$, where $\text{Bessel}(\cdot)$ denotes Bessel distribution. If $\epsilon^\alpha = 0$, we sample $g_{k_1k}^{(i)}$ via

$$\left(g_{k_1k}^{(i)} \mid -\right) \sim \begin{cases} \text{Pois}\left(\frac{c_k^{(i)} \lambda_{k_1k}^{(i-1)}}{c_k^{(i)} - \ln\left(1-q_k^{(i)}\right)}\right) & \text{if } h_{k_1k}^{(i)} = 0 \\ \text{SCH}\left(h_{k_1k}^{(i)}, \frac{c_k^{(i)} \lambda_{k_1k}^{(i-1)}}{c_k^{(i)} - \ln\left(1-q_k^{(i)}\right)}\right) & \text{otherwise,} \end{cases}$$

where $\text{SCH}(\cdot)$ denotes the shifted confluent hypergeometric distribution [16]. Defining $g_{k_1k}^{(i)} = g_{k_1\cdot k}^{(i)} = \sum_{k2=1}^{K} g_{k_1k_2k}^{(i)}$, we first augment

$$\left(g_{k_11k}^{(i)}, \cdots, g_{k_1Kk}^{(i)}\right) \sim \text{Mult}\left(g_{k_1k}^{(i)}, \left(\psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}\right)_{k_2=1}^{K}\right),$$

then we obtain $g_{k_1k_2k}^{(i)} \sim \text{Pois}(\gamma^{(i-1)} \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)})$. By Dirichlet-multinomial conjugacy, we have

$$\left(\left(\psi_{k1k_2}^{(i-1)}, \cdots, \psi_{kKk_2}^{(i-1)}\right) \mid -\right) \sim \text{Dir}\left(\epsilon_0 + g_{1k_2k}^{(i)}, \cdots, \epsilon_0 + g_{Kk_2k}^{(i)}\right), \text{ and}$$
$$\left(\boldsymbol{\pi}_k^{(i-1)} \mid -\right) \sim \text{Dir}\left(\alpha_{1k}^{(i-1)} + l_{1k}^{(i-1)} + g_{\cdot 1k}^{(i)}, \cdots, \alpha_{Kk}^{(i-1)} + l_{Kk}^{(i-1)} + g_{\cdot Kk}^{(i)}\right).$$

Specifically, we have $\alpha_{k_1k}^{(1)} = \nu_{k_1}\nu_k$, if $k_1 \neq k$, and $\alpha_{k_1k}^{(1)} = \xi\nu_k$, if $k_1 = k$.

5 Related Work

Modeling count time sequences has been receiving increasing attentions in statistical and machine learning communities. Han et al. [23] adopted linear dynamical systems to capture the underlying dynamics of the data and leveraged Extended Rank likelihood function to model count observations. Some Poisson-gamma models assume that the count vector at each time step is modeled by Poisson factor analysis (PFA) [11] and leverage special stochastic processes to model the temporal dependencies of latent factors. For example, gamma process dynamic Poisson factor analysis (GP-DPFA) [15] adopts gamma Markov chains which assumes the latent factor of the next time step is drawn from a gamma distribution with the shape parameter be the latent factor of the current time step. Schein et al. [13] proposed Poisson-gamma dynamical systems (PGDSs), which take the interactions among latent dimensions into account and use a transition matrix to capture the interactions. Deep dynamic Poisson factor analysis (DDPFA) [29] adopts recurrent neural networks (RNNs) to capture the complex long-term dependencies of latent factors. Yang and Koeppl [30] applied Poisson-gamma count model to analyze relational data arising from longitudinal networks, which can capture the evolution of individual node-group memberships over time. Many modifications of PGDS have been proposed in recent years. Guo et al. [31] proposed deep Poisson-gamma dynamical systems which aim to capture the long-range temporal dependencies. Schein et al. [16] employed Poisson-randomized gamma distribution to build a new transition process of latent factors. Chen et al. [32] proposed Switching Poisson-gamma dynamical systems (SPGDS), allowing PGDS to select from several transition matrices, and thus can better adapt to nonlinear dynamics. In contrast to SPGDS, the number of transition matrices of the proposed NS-PGDS is not limited and thus can be adopted to analyze various complicated non-stationary count sequences. Filstroff et al. [33] extensively analyzed many gamma Markov chains for non-negative matrix factorization and introduced new gamma Markov chains with well-defined stationary distribution (BGAR).

6 Experiments

We conducted experiments for both predictive and exploratory analysis to demonstrate the ability of the proposed model in capturing non-stationary count time sequences. The baseline models included in the experiments are: 1) Gamma process dynamic Poisson factor analysis (GP-DPFA) [15]. GP-DPFA models the evolution of latent components as $\theta_k^{(t)} \sim \text{Gam}(\theta_k^{(t-1)}, c_t)$, in which each component evolves independently of the other components. 2) Gamma Markov chains on the rate parameter of gamma distribution (GMC-RATE) [33]. GMC-RATE adopts gamma Markov chains defined via the rate parameter of the gamma distribution to model the evolution of $\theta_k^{(t)}$

			GP-DPFA	GMC-RATE	GMC-HIER	BGAR	PGDS	NS-PGDS (Dir-Dir)	NS-PGDS (Dir-Gam-Dir)	NS-PGDS (PR-Gam-Dir)
ICEWS	MAE	S	0.259 ±0.005	0.258 ±0.005	0.256 ±0.006	0.264 ±0.006	0.215 ±0.007	0.215 ±0.008	**0.214** ±0.008	0.215 ±0.008
		F	0.176 ±0.005	0.187 ±0.003	0.185 ±0.016	0.222 ±0.043	0.185 ±0.003	**0.167** ±0.009	0.169 ±0.006	0.169 ±0.009
	MRE	S	0.125 ±0.003	0.124 ±0.002	0.122 ±0.003	0.130 ±0.004	0.102 ±0.005	**0.101** ±0.005	**0.101** ±0.005	0.102 ±0.005
		F	0.099 ±0.006	0.114 ±0.003	0.111 ±0.018	0.142 ±0.036	0.108 ±0.001	**0.094** ±0.005	0.097 ±0.004	0.097 ±0.008
NIPS	MAE	S	18.299 ±6.545	17.105 ±6.449	17.098 ±6.441	17.935 ±6.450	14.706 ±4.414	14.032 ±4.401	14.026 ±4.405	**14.014** ±4.387
		F	48.355 ±1.461	46.234 ±1.629	102.506 ±39.932	62.449 ±14.463	51.562 ±0.679	**45.979** ±1.342	46.710 ±1.152	46.582 ±1.196
	MRE	S	0.729 ±0.412	0.684 ±0.316	0.664 ±0.315	0.769 ±0.366	0.590 ±0.097	0.581 ±0.090	0.581 ±0.090	**0.580** ±0.090
		F	0.415 ±0.016	**0.387** ±0.023	0.580 ±0.148	0.465 ±0.049	0.459 ±0.006	0.399 ±0.003	0.395 ±0.006	0.397 ±0.003
USEI	MAE	S	4.681 ±0.564	4.931 ±0.872	4.748 ±0.829	5.244 ±0.939	4.703 ±0.538	4.600 ±0.542	4.608 ±0.541	**4.596** ±0.562
		F	11.665 ±0.367	9.454 ±0.809	12.423 ±1.060	21.948 ±0.133	11.118 ±0.220	7.973 ±1.222	**7.168** ±1.221	7.296 ±1.127
	MRE	S	1.458 ±0.177	1.128 ±0.189	**1.088** ±0.162	1.941 ±0.209	1.279 ±0.257	1.309 ±0.220	1.298 ±0.236	1.301 ±0.229
		F	7.473 ±0.623	6.508 ±0.571	8.929 ±2.514	13.706 ±1.268	4.238 ±0.325	2.602 ±0.455	**2.577** ±0.331	2.685 ±0.366
COVID-19	MAE	S	7.935 ±0.751	7.144 ±1.159	7.240 ±0.848	7.819 ±1.348	7.566 ±1.095	**6.969** ±1.107	6.988 ±1.056	6.981 ±1.022
		F	9.137 ±1.102	9.600 ±1.257	10.409 ±1.910	12.550 ±2.156	9.314 ±0.236	8.799 ±0.706	**8.770** ±0.438	9.033 ±0.477
	MRE	S	0.564 ±0.126	**0.493** ±0.136	0.504 ±0.109	0.769 ±0.169	0.558 ±0.130	0.523 ±0.125	0.525 ±0.124	0.526 ±0.123
		F	0.627 ±0.106	0.556 ±0.052	0.585 ±0.067	0.759 ±0.150	0.585 ±0.007	0.523 ±0.028	0.519 ±0.017	**0.513** ±0.014

Table 1: Results of predictive analysis. "S" means data smoothing and "F" means data forecasting.

as $\theta_k^{(t)} \sim \mathrm{Gam}(\alpha, \beta/\theta_k^{(t-1)})$. 3) Gamma Markov chains on the rate parameter with hierarchical auxiliary variable (GMC-HIER) [33]. GMC-HIER models the evolution of latent components with an auxiliary variables as $z_k^{(t)} \sim \mathrm{Gam}(\alpha_z, \beta_z\theta_k^{(t-1)})$ and $\theta_k^{(t)} \sim \mathrm{Gam}(a_\theta, \beta_\theta z_k^{(t)})$. 4) Autogressive beta-gamma procecss (BGAR) [34, 33]. BGAR is also a gamma Markov model. In contrast to the above models, there is a well-defined stationary distribution for BGAR. 5) Poisson-gamma dynamical system (PGDS) [13] takes interactions among latent dimensions into account, and models the evolution of $\theta_k^{(t)}$ as $\theta_k^{(t)} \sim \mathrm{Gam}(\tau_0 \sum_{k_2=1}^K \pi_{kk_2}\theta_{k_2}^{(t-1)}, \tau_0)$.

The real-world datasets used in the experiments are: 1) **Integrated Crisis Early Warning System (ICEWS)**: ICEWS is an international relations event dataset, comprising interaction events between countries extracted from news corpora. For ICEWS dataset, we have $T = 365$ time steps and $V = 6197$ dimensions, and we set $M = 30$. 2) **NIPS**: NIPS dataset contains the papers published in the NeurIPS conference from 1987 to 2015. We have $T = 28$ time steps and $V = 2000$ dimensions for NIPS dataset and we set $M = 5$. 3) **U.S. Earthquake Intensity (USEI)**: USEI contains a collection of damage and felt reports for U.S. (and a few other countries) earthquakes. We use the monthly reports from 1957-1986 and have $T = 348, V = 64$ and set $M = 34$. 4) **COVID-19**: This dataset contains daily death cases data for states in the United States, spanning from March 2020 to June 2020. For this dataset, we have $V = 51$ dimensions and $T = 90$ time steps and set $M = 20$.

6.1 Predictive Analysis

To compare the predictive performance of the proposed model with the baselines, we considered two standard tasks: data smoothing and forecasting. For data smoothing task, our objective is to predict $\boldsymbol{y}^{(t)}$ given the remaining data observation $\boldsymbol{Y}\backslash\boldsymbol{y}^{(t)}$. To this end, we randomly masked 10 percents of the observed data over non-adjacent time steps, and predicted the masked values. For forecasting task, we held out data of the last S time steps, and predicted $\boldsymbol{y}^{(T+1)}, \cdots, \boldsymbol{y}^{(T+S)}$ given $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(T)}$. In this experiment we set $S = 2$. We ran the baseline models including GP-DPFA, PGDS, GMC-RATE, GMC-HIER, BGAR, using their default settings as provided in [15, 13, 33]. For the NS-PGDS, we set $K = 100$ for ICEWS, $K = 10$ for other datasets, and set $\tau_0 = 1, \gamma_0 = 50, \epsilon_0 = 0.1$. We performed 4000 Gibbs sampling iterations. In the experiments, we found that the Gibbs sampler started to converge after 1000 iterations, and thus we set the burn-in time be 2000 iterations. We retained every hundredth sample, and averaged the predictions over the samples. Mean relative error (MRE) and mean absolute error (MAE) are adopted to evaluate the model's predictive capability, which are defined as $\mathrm{MRE} = \frac{1}{TV}\sum_t\sum_v \frac{|y_v^{(t)} - \hat{y}_v^{(t)}|}{1+y_v^{(t)}}$ and $\mathrm{MAE} = \frac{1}{TV}\sum_t\sum_v \mid y_v^{(t)} - \hat{y}_v^{(t)} \mid$ respectively, where $y_v^{(t)}$ indicates the true count and $\hat{y}_v^{(t)}$ is the prediction.

As the experiment results shown in Table 1, the NS-PGDS exhibits improved performance in both data smoothing and forecasting tasks. We attribute this enhanced capability to the time-varying transition kernels, which effectively adapt to the non-stationary environment, and thus achieve improved predictive performance. For some datasets (e.g. ICEWS) and tasks, the effectiveness of the Dir-Gam-Dir and Pr-Gam-Dir constructions does not be exhibited in the numerical results. However, these two constructions indeed induce more informative patterns compared with Dir-Dir construction, as shown in the exploratory analysis.

6.2 Exploratory Analysis

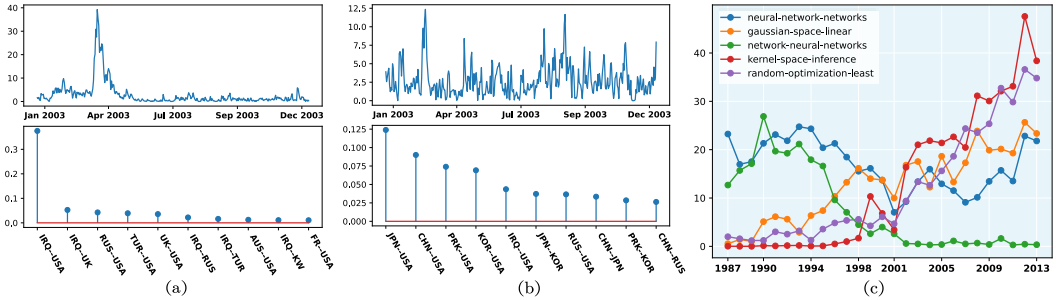

Figure 4: The latent factors inferred by the NS-PGDS. (a) and (b) illustrate the top 2 latent factors inferred from ICEWS dataset, (a) corresponds to Iraq war and (b) corresponds to the Six-Party Talks. (c) illustrates the evolving trends of the top 5 latent factors inferred from NIPS dataset.

We used ICEWS and NIPS datasets for exploratory analysis, and chose the NS-PGDS with Dirichlet-Dirichlet Markov chains for illustration. Figure 4(a) and Figure 4(b) demonstrate the top 2 latent factors inferred by NS-PGDS from ICEWS dataset. From Figure 4(a) we can see that the main labels are "Iraq (IRQ)–United States (USA)", "Iraq (IRQ)–United Kingdom (UK)", "Russia (RUS)–United States (USA)", and so on. This latent factor probably corresponds to the topic about Iraq war. Besides, in Figure 4(a), there is a peak around March, 2003, and we know that the Iraq war broke out exactly on 20 March, 2003. In addition, the most dominant labels shown in Figure 4(b) are "Japan (JPN)–United States (USA)", "China (CHN)–United States (USA)", "North Korea (PRK)–United States (USA)", "South Korea (KOR)–United States (USA)", and so on. We can infer that this latent factor corresponds to "Six-Party Talks" and other accidents about it.

Figure 4(c) demonstrates the evolving trends of the top 5 latent factors inferred by the NS-PGDS from NIPS dataset, and the legend indicates the representative words of the corresponding latent factors. Clearly, the green and blue lines correspond to the latent factors of neural network research which started to decline from the 1990s. From the 1990s we see that the latent factors about statistical and probabilistic methods began to dominate the NeurIPS conference. In addition, the NS-PGDS also captured the revival of neural networks (blue line) from the 2010s. The above observations from the latent structure inferred by the NS-PGDS match our prior knowledge.

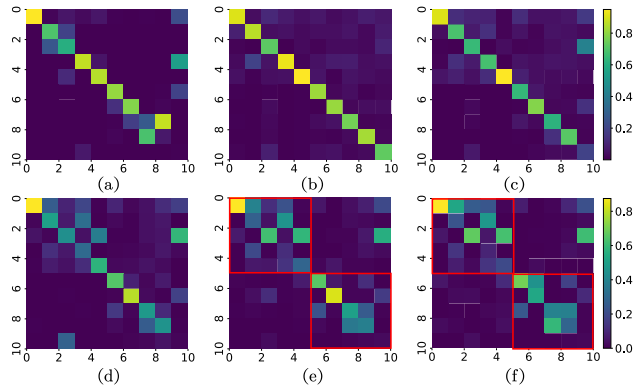

Figure 5: Transition matrices inferred from NIPS dataset. (a) illustrates the transition matrix inferred by the PGDS. (b)-(f) illustrate the time-varying transition matrices inferred by the NS-PGDS.

Next, we explored the time-varying transition matrices inferred by the NS-PGDS. We chose NIPS dataset for illustratiuon, and set $K = 10$ and the interval length M to be 5. The time-varying transition matrices are shown from Figure 5(b) to Figure 5(f). At the beginning, matrices shown in Figure 5(b) and Figure 5(c) are close to identity matrices. Then the transition matrices tend to become block diagonal matrices with 2 blocks, as shown in Figure 5(d)-5(f). The representative words for latent factors in the first block are "state-linear-classification", "network-neural-networks", "kernel-image-space", "network-neural-networks", "neural-networks-state". The representative words for latent factors in the second block are "image-sparse-matrix", "kernel-supervised-random", "matrix-sample-random", "inference-prior-latent", "state-policy-gamma". The first block primarily captured the correlations among the research topics about neural networks. The second block reflects that, from the 1990s, statistical learning and Bayesian methods began to dominate, and these topics are highly correlated. Figure 5(a) illustrates the transition matrix inferred

by the PGDS, which is averaged over all time steps. Compared with the NS-PGDS, the PGDS can not capture the informative time-varying transition dynamics. We also analyzed the features of the proposed Dirichlet Markov chains. The left column of Figure 6 demonstrates transition matrices of the first four sub-intervals of ICEWS dataset inferred by the NS-PGDS (Dir-Dir). Because of the Dir-Dir construction, the consecutive transition matrices smoothly change over time and thus the NS-PGDS may lack sufficient flexibility to capture rapid dynamics. The middle column of Figure 6 illustrates the transition matrices inferred by the NS-PGDS (Dir-Gam-Dir), which takes mutations among latent components into account and captured more complicated patterns. Transition matrices inferred by the PR-Gam-Dir construction are shown in the right column of Figure 6, these matrices not only exhibited sufficient flexibility but also captured sparser patterns compared with the Dir-Gam-Dir construction.

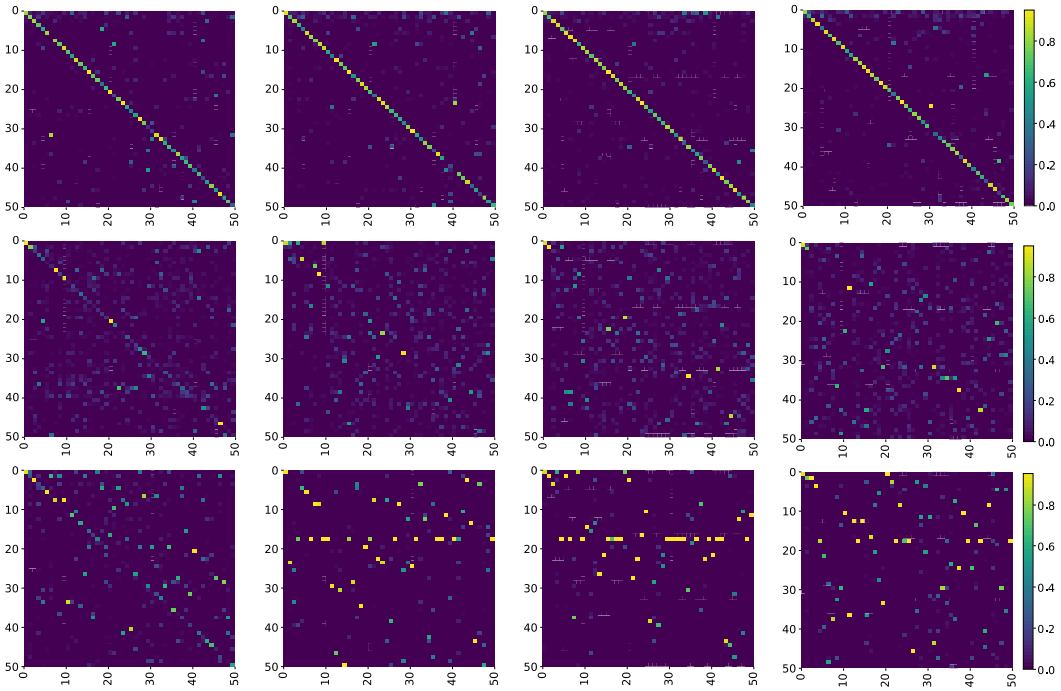

Figure 6: From top to bottom are the first four transition matrices inferred by different Dirichlet Markov chains from ICEWS dataset. Top row: Matrices inferred by the Dir-Dir construction. Middle row: Matrices inferred by the Dir-Gam-Dir construction. Bottom row: Matrices inferred by the PR-Gam-Dir construction.

7 Conclusion

The Poisson-gamma dynamical systems with time-varying transition matrices, have been proposed to capture complicated dynamics observed in *non-stationary* count sequences. In particular, Dirichlet Markov chains are constructed to allow the underlying transition matrices to evolve over time. Although the Dirichlet Markov processes lack conjugacy, we have developed tractable-but-efficient Gibbs sampling algorithms to perform posterior simulation. The experiment results demonstrate the improved performance of the proposed NS-PGDS in data smoothing and forecasting tasks, compared with the PGDS with a stationary transition kernel. Moreover, the experimental results on several real-world data sets show the explainable structures inferred by the proposed NS-PGDS. For the future work, we plan to design a method that can find the point of change and thus the length of each sub-interval can be determined automatically instead of a constant. We also consider to generalize Dirichlet belief networks by incorporating the proposed Dirichlet Markov chain constructions, which allow the hierarchical topics to mutate across layers, and thus can generate more rich text information. And we also consider to capture non-stationary interaction dynamics among individuals over online social networks in the future research.

References

[1] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[2] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[3] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. Scalable generalized dynamic topic models. In *International Conference on Artificial Intelligence and Statistics*, pages 1427–1435, 2018.

[4] Daniel Sheldon and Thomas G Dietterich. Collective graphical models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1161–1169, 2011.

[5] James Raymer, Arkadiusz Wiśniowski, Jonathan J Forster, Peter WF Smith, and Jakub Bijak. Integrated modeling of european migration. *Journal of the American Statistical Association*, 108(503):801–819, 2013.

[6] Tom Wilson. Methods for estimating sub-state international migration: The case of australia. *Spatial Demography*, 5(3):171–192, 2017.

[7] Philippe Wanner. How well can we estimate immigration trends using google data? *Quality & Quantity*, 55(4):1181–1202, 2021.

[8] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[9] Zoubin Ghahramani and Sam T Roweis. Learning nonlinear dynamical systems using an em algorithm. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, pages 431–437, 1998.

[10] M Zhou and L Carin. Augment-and-conquer negative binomial processes. *Advances in Neural Information Processing Systems*, 4:2546–2554, 2012.

[11] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(02):307–320, 2015.

[12] Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International conference on knowledge discovery and data mining*, pages 1045–1054, 2015.

[13] Aaron Schein, Mingyuan Zhou, and Hanna Wallach. Poisson-gamma dynamical systems. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 5012–5020, 2016.

[14] Aaron Schein, Mingyuan Zhou, David Blei, and Hanna Wallach. Bayesian poisson tucker decomposition for learning the structure of international relations. In *International Conference on Machine Learning*, pages 2810–2819, 2016.

[15] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. Nonparametric bayesian factor analysis for dynamic count matrices. In *Artificial Intelligence and Statistics*, pages 1–9, 2015.

[16] Aaron Schein, Scott W Linderman, Mingyuan Zhou, David M Blei, and Hanna Wallach. Poisson-randomized gamma dynamical systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 782–793, 2019.

[17] Jonathan Chang and David Blei. Relational topic models for document networks. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pages 81–88, 2009.

[18] H Juliette T Unwin, Swapnil Mishra, Valerie C Bradley, Axel Gandy, Thomas A Mellan, Helen Coupland, Jonathan Ish-Horowicz, Michaela AC Vollmer, Charles Whittaker, Sarah L Filippi, et al. State-level tracking of covid-19 in the united states. *Nature Communications*, 11(1):1–9, 2020.

[19] Rainer Winkelmann. *Econometric Analysis of Count Data*. Springer Publishing Company, Incorporated, 5th edition, 2008.

[20] Guy Grossman, Soojong Kim, Jonah M Rexer, and Harsha Thirumurthy. Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences*, 117(39):24144–24153, 2020.

[21] IHME COVID-19 Forecasting Team. Modeling covid-19 scenarios for the united states. *Nature medicine*, 27(1):94–105, 2021.

[22] Luzhao Feng, Ting Zhang, Qing Wang, Yiran Xie, Zhibin Peng, Jiandong Zheng, Ying Qin, Muli Zhang, Shengjie Lai, Dayan Wang, et al. Impact of covid-19 outbreaks and interventions on influenza in china and the united states. *Nature communications*, 12(1):3249, 2021.

[23] Shaobo Han, Lin Du, Esther Salazar, and Lawrence Carin. Dynamic rank factor model for text streams. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 2663–2671, 2014.

[24] Rahi Kalantari and Mingyuan Zhou. Graph gamma process generalized linear dynamical systems. *arXiv preprint arXiv:2007.12852*, 2020.

[25] Lin Yuan and John D Kalbfleisch. On the bessel distribution and related problems. *Annals of the Institute of Statistical Mathematics*, 52:438–447, 2000.

[26] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[27] Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

[28] Mingyuan Zhou. Nonparametric bayesian negative binomial factor analysis. *Bayesian Analysis*, 13(4):1065–1093, 2018.

[29] Chengyue Gong and Win-bin Huang. Deep dynamic poisson factorization model. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1665–1673, 2017.

[30] Sikun Yang and Heinz Koeppl. Dependent relational gamma process models for longitudinal networks. In *International Conference on Machine Learning*, pages 5551–5560, 2018.

[31] Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. Deep poisson gamma dynamical systems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8451–8461, 2018.

[32] Wenchao Chen, Bo Chen, Yicheng Liu, Qianru Zhao, and Mingyuan Zhou. Switching poisson gamma dynamical systems. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2029–2036, 2021.

[33] Louis Filstroff, Olivier Gouvert, Cédric Févotte, and Olivier Cappé. A comparative study of gamma markov chains for temporal non-negative matrix factorization. *IEEE Transactions on Signal Processing*, 69:1614–1626, 2021.

[34] Peter AW Lewis, Edward McKenzie, and David Kennedy Hugus. Gamma processes. *Stochastic Models*, 5(1):1–30, 1989.

[35] John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

A MCMC Inference

Notation. When expressing the full conditionals for Gibbs sampling, we use the shorthand "–" to denote all other variables. We use "$\cdot$" as an index summation shorthand, e.g., $x_{\cdot j} = \sum_i x_{ij}$.

In this section, we present a fully-conjugate and efficient Gibbs sampler for the proposed NS-PGDS. The sampling algorithms depend on several key technical results, which we will repeatedly exploit, thus we list them below.

Negative-binomial Distribution. Let $y \sim \mathrm{Pois}(c\lambda)$, and $\lambda \sim \mathrm{Gam}(a, b)$. If we marginalize over λ, then $y \sim \mathrm{NB}\left(a, \frac{c}{b+c}\right)$ is a negative-binomial distributed random variable. We can further parameterize it as $y \sim \mathrm{NB}\left(a, g\left(\zeta\right)\right)$, where $g\left(z\right) = 1 - \exp\left(-z\right)$ and $\zeta = \ln\left(1 + \frac{c}{b}\right)$.

Lemma 1. If $y \sim \mathrm{NB}\left(a, g\left(\zeta\right)\right)$ and $l \sim \mathrm{CRT}\left(y, a\right)$, where $\mathrm{CRT}\left(\cdot\right)$ represents Chinese restaurant table distribution [26], then the joint distribution of y and l can be equivalently distributed as $y \sim \mathrm{SumLog}\left(l, g\left(\zeta\right)\right)$ and $l \sim \mathrm{Pois}\left(a\zeta\right)$ [11], i.e.

$$\mathrm{NB}\left(y; a, g\left(\zeta\right)\right)\mathrm{CRT}\left(l; y, a\right) = \mathrm{SumLog}\left(y; l, g\left(\zeta\right)\right)\mathrm{Pois}\left(l; a\zeta\right),$$

where $\mathrm{SumLog}\left(l, g\left(\zeta\right)\right) = \sum_{i=1}^{l} x_i$ and $x_i \sim \mathrm{Log}\left(g\left(\zeta\right)\right)$ are independently and identically logarithmic distributed random variables [27].

Lemma 2. Suppose $\mathbf{n} = (n_1, \cdots, n_K)$ and $\mathbf{n} \mid n \sim \mathrm{DirMult}\left(n, r_1, \cdots, r_K\right)$, where $\mathrm{DirMult}\left(\cdot\right)$ refers to Dirichlet-multimonial distribution. We sample the augmented variable $q \mid n \sim \mathrm{Beta}\left(n, r_\cdot\right)$, where $r_\cdot = \sum_{k=1}^{K} r_k$. According to [28], conditioning on q, we have $n_k \sim \mathrm{NB}\left(r_k, q\right)$.

Lemma 3. If $y_\cdot = \sum_{s=1}^{S} y_s$, and $y_s \overset{\text{i.i.d}}{\sim} \mathrm{Pois}(\lambda_s), s = 1, \cdots, S$. Then $y_\cdot \sim \mathrm{Pois}(\sum_{s=1}^{S} \lambda_s)$ and $(y_1, \cdots, y_S) \sim \mathrm{Mult}(y_\cdot, (\frac{\lambda_1}{\sum_{s=1}^{S}\lambda_s}, \cdots, \frac{\lambda_S}{\sum_{s=1}^{S}\lambda_s}))$, where $\mathrm{Mult}\left(\cdot\right)$ represents multinomial distribution [35].

Sampling $y_{vk}^{(t)}$: Use the relationship between Poisson and multinomial distributions as described by Lemma 3, given observed counts and latent parameters, we sample

$$\left(\left(y_{vk}^{(t)}\right)_{k=1}^{K} \mid -\right) \sim \mathrm{Mult}\left(y_v^{(t)}, \left(\frac{\phi_{vk}\theta_k^{(t)}}{\sum_{k=1}^{K}\phi_{vk}\theta_k^{(t)}}\right)_{k=1}^{K}\right). \tag{6}$$

Then the distribution of $y_{vk}^{(t)}$ is $y_{vk}^{(t)} \sim \mathrm{Pois}(\delta^{(t)}\phi_{vk}\theta_k^{(t)})$.

Sampling $\boldsymbol{\phi_k}$: Via Dirichlet-multinomial conjugacy, the posterior of $\boldsymbol{\phi_k}$ is

$$\left(\boldsymbol{\phi_k} \mid -\right) \sim \mathrm{Dir}\left(\epsilon_0 + \sum_{t=1}^{T} y_{1k}^{(t)}, \cdots, \epsilon_0 + \sum_{t=1}^{T} y_{Vk}^{(t)}\right). \tag{7}$$

Marginalizing over $\theta_k^{(t)}$: Note that $y_v^{(t)} = y_{v\cdot}^{(t)} = \sum_{k=1}^{K} y_{vk}^{(t)}$ and $y_{vk}^{(t)} \sim \mathrm{Pois}(\delta^{(t)}\phi_{vk}\theta_k^{(t)})$. Then we define $y_{\cdot k}^{(t)} = \sum_{v=1}^{V} y_{vk}^{(t)}$. Because $\sum_{v=1}^{V}\phi_{vk} = 1$, we obtain $y_{\cdot k}^{(t)} \sim \mathrm{Pois}(\delta^{(t)}\theta_k^{(t)})$.

We start by marginalizing over $\theta_k^{(T)}$, using the definition of negative-binomial distribution, we obtain

$$y_{\cdot k}^{(T)} \sim \mathrm{NB}\left(\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{i(T-1)}\theta_{k_2}^{(T-1)}, g\left(\zeta^{(T)}\right)\right),$$

where $\zeta^{(T)} = \ln(1 + \frac{\delta^{(T)}}{\tau_0})$. Next, we further marginalize over $\theta_k^{(T-1)}$. To this end, we first sample auxiliary variables

$$l_k^{(T)} \sim \mathrm{CRT}\left(y_{\cdot k}^{(T)}, \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{i(T-1)}\theta_{k_2}^{(T-1)}\right).$$

By Lemma 1, the joint distribution of $y_{\cdot k}^{(T)}$ and $l_k^{(T)}$ can be expressed as

$$y_{\cdot k}^{(T)} \sim \mathrm{SumLog}\left(l_k^{(T)}, g\left(\zeta^{(T)}\right)\right) \text{ and } l_k^{(T)} \sim \mathrm{Pois}\left(\zeta^{(T)}\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}^{i(T-1)}\theta_{k_2}^{(T-1)}\right).$$

Via Lemma 3, we re-express the auxiliary variables as

$$l_k^{(T)} = l_{k\cdot}^{(T)} = \sum_{k_2=1}^{K} l_{kk_2}^{(T)}, \text{ and obtain } l_{kk_2}^{(T)} \sim \text{Pois}\left(\zeta^{(T)}\tau_0\pi_{kk_2}^{i(T-1)}\theta_{k_2}^{(T-1)}\right).$$

Then we define $l_{\cdot k}^{(T)} = \sum_{k_1=1}^{K} l_{k_1k}^{(T)}$. Leveraging Lemma 3 and $\sum_{k_1=1}^{K}\pi_{k_1k}^{i(T-1)} = 1$, we obtain

$$l_{\cdot k}^{(T)} \sim \text{Pois}\left(\zeta^{(T)}\tau_0\theta_k^{(T-1)}\right) \text{ and } \left(l_{1k}^{(T)}, \cdots, l_{Kk}^{(T)}\right) \sim \text{Mult}\left(l_{\cdot k}^{(T)}, \left(\pi_{1k}^{i(T-1)}, \cdots, \pi_{Kk}^{i(T-1)}\right)\right).$$

Next, note that $y_{\cdot k}^{(T-1)} \sim \text{Pois}(\delta^{(T-1)}\theta_k^{(T-1)})$, if we introduce $m_k^{(T-1)} = y_{\cdot k}^{(T-1)} + l_{\cdot k}^{(T)}$, then we have

$$m_k^{(T-1)} \sim \text{Pois}\left(\theta_k^{(T-1)}\left(\delta^{(T-1)} + \zeta^{(T)}\tau_0\right)\right).$$

Because the prior of $\theta_k^{(T-1)}$ is gamma distributed, by the definition of negative-binomial distribution, we can again marginalize over $\theta_k^{(T-1)}$ to obtain

$$m_k^{(T-1)} \sim \text{NB}\left(\tau_0\sum_{k_2=1}^{K}\pi_{kk_2}^{i(T-2)}\theta_{k_2}^{(T-2)}, g\left(\zeta^{(T-1)}\right)\right),$$

where $\zeta^{(T-1)} = \ln(1 + \frac{\delta^{(T-1)}}{\tau_0} + \zeta^{(T)})$. Then we introduce auxiliary variables

$$l_k^{(T-1)} \sim \text{CRT}\left(m_k^{(T-1)}, \tau_0\sum_{k_2=1}^{K}\pi_{kk_2}^{i(T-2)}\theta_{k_2}^{(T-2)}\right).$$

And similar to the case for $t = T$, we can obtain

$$l_{\cdot k}^{(T-1)} \sim \text{Pois}\left(\zeta^{(T-1)}\tau_0\theta_k^{(T-2)}\right) \text{ and } m_k^{(T-2)} \sim \text{NB}\left(\tau_0\sum_{k_2=1}^{K}\pi_{kk_2}^{i(T-3)}\theta_{k_2}^{(T-3)}, g\left(\zeta^{(T-2)}\right)\right).$$

Thus we have marginalized over $\theta_k^{(T-2)}$. Note that we can repeat this marginalization process recursively until $t = 1$ with $\zeta^{(t)} = \ln(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)})$ and $m_k^{(T)} = y_{\cdot k}^{(T)}$ to maginalize over all the $\theta_k^{(t)}$.

Sampling $\theta_k^{(t)}$: Via the above marginalization process, to sample from the posterior of $\theta_k^{(t)}$, we first sample the auxiliary variables. Setting $l_{\cdot k}^{(T+1)} = 0$ and $\zeta^{(T+1)} = 0$, we sample the augmented variables backwards from $t = T, \cdots, 2$,

$$\left(l_{k\cdot}^{(t)} \mid -\right) \sim \text{CRT}\left(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)}, \tau_0\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}\right), \tag{8}$$

$$\left(l_{k1}^{(t)}, \cdots, l_{kK}^{(t)} \mid -\right) \sim \text{Mult}\left(l_{k\cdot}^{(t)}, \left(\frac{\pi_{k1}^{i(t-1)}\theta_1^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}}, \cdots, \frac{\pi_{kK}^{i(t-1)}\theta_K^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}}\right)\right). \tag{9}$$

And via Lemma 3, we obtain

$$\left(l_{1k}^{(t)}, \cdots, l_{Kk}^{(t)}\right) \sim \text{Mult}\left(l_{\cdot k}^{(t)}, \pi_{1k}^{i(t-1)}, \cdots, \pi_{Kk}^{i(t-1)}\right) \tag{10}$$

We compute $\zeta^{(t)}$ recursively via

$$\zeta^{(t)} = \ln\left(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)}\right). \tag{11}$$

After sampling the auxiliary variables, then for $t = 1, \cdots, T$, by Poisson-gamma conjugacy, we obtain

$$\left(\theta_k^{(1)} \mid -\right) \sim \text{Gam}\left(y_{\cdot k}^{(1)} + l_{\cdot k}^{(2)} + \tau_0\nu_k, \tau_0 + \delta^{(1)} + \zeta^{(2)}\tau_0\right), \tag{12}$$

$$\left(\theta_k^{(t)} \mid -\right) \sim \text{Gam}\left(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)} + \tau_0\sum_{k_2=1}^{K}\pi_{kk_2}^{i(t-1)}\theta_{k_2}^{(t-1)}, \tau_0 + \delta^{(t)} + \zeta^{(t+1)}\tau_0\right). \tag{13}$$

Sampling $\mathbf{\Pi}^{(i)}$: We define M as the length of each sub-interval, and I as the number of intervals. For $i = I$, by Eq.(10), $(l_{1k}^{(I)}, \cdots, l_{Kk}^{(I)})$ is multinomial distributed. Thus by multinomial-Dirichlet conjugacy, we obtain

$$\left(\boldsymbol{\pi}_k^{(I)} \mid -\right) \sim \mathrm{Dir}\left(\alpha_{1k}^{(I)} + l_{1k}^{(I)}, \cdots, \alpha_{Kk}^{(I)} + l_{Kk}^{(I)}\right), \tag{14}$$

where $l_{k_1k}^{(I)}$ indicates the summation of $l_{k_1k}^{(t)}$ over I-th sub-interval, i.e. $l_{k_1k}^{(I)} = \sum_{t=(I-1)M+1}^{T} l_{k_1k}^{(t)}$.

Inference for Dirichlet-Dirichlet Markov chains. For Dirichlet-Dirichlet Markov chains, $\alpha_{k_1k}^{(i)} = \eta K \pi_{k_1k}^{(i-1)}$. By Eq.(10), $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ is multinomial distributed. If we marginalize $(\pi_{1k}^{(i)}, \cdots, \pi_{Kk}^{(i)})$, $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ will be Dirichlet-multinomial distributed. Thus by Lemma 2, for $i = I$, we first sample the auxiliary variables as

$$\left(q_k^{(I)} \mid -\right) \sim \mathrm{Beta}\left(l_{\cdot k}^{(I)}, \eta K\right) \text{ and } \left(h_{k_1k}^{(I)} \mid -\right) \sim \mathrm{CRT}\left(l_{k_1k}^{(I)}, \eta K \pi_{k_1k}^{(I-1)}\right). \tag{15}$$

Similarly, by Eq.(18), $(h_{1k}^{(i)}, \cdots, h_{Kk}^{(i)})$ is also Dirichlet-multinomial distributed. Thus for $i = I-1, \cdots, 2$, we sample the auxiliary variables as

$$\left(q_k^{(i)} \mid -\right) \sim \mathrm{Beta}\left(l_{\cdot k}^{(i)} + h_{\cdot k}^{(i+1)}, \eta K\right) \text{ and } \left(h_{k_1k}^{(i)} \mid -\right) \sim \mathrm{CRT}\left(l_{k_1k}^{(i)} + h_{k_1k}^{(i+1)}, \eta K \pi_{k_1k}^{(i-1)}\right), \tag{16}$$

where $l_{k_1k}^{(i)} = \sum_{(i-1)M+1}^{iM} l_{k_1k}^{(t)}$ refers to the summation of $l_{k_1k}^{(t)}$ over i-th interval. Via Lemma 2, conditioning on $q_k^{(i)}$, we have

$$\left(l_{k_1k}^{(i)} + h_{k_1k}^{(i+1)}\right) \sim \mathrm{NB}\left(\eta K \pi_{k_1k}^{(i-1)}, q_k^{(i)}\right).$$

Then via Lemma 1, we obtain

$$h_{k_1k}^{(i)} \sim \mathrm{Pois}\left(-\eta K \pi_{k_1k}^{(i-1)} \ln\left(1 - q_k^{(i)}\right)\right). \tag{17}$$

Note that by Eq.(17), $h_{k_1k}^{(i)}$ is Poisson distributed and by Lemma 3, we obtain

$$\left(h_{1k}^{(i)}, \cdots, h_{Kk}^{(i)}\right) \sim \mathrm{Mult}\left(h_{\cdot k}^{(i)}, \left(\pi_{1k}^{(i-1)}, \cdots, \pi_{Kk}^{(i-1)}\right)\right). \tag{18}$$

In addition, note that

$$\left(l_{1k}^{(i-1)}, \cdots, l_{Kk}^{(i-1)}\right) \sim \mathrm{Mult}\left(l_{\cdot k}^{(i-1)}, \left(\pi_{1k}^{(i-1)}, \cdots, \pi_{Kk}^{(i-1)}\right)\right),$$

via Dirichlet-multinomial conjugacy, for $i = I-1, \cdots, 2$, we obtain

$$\left(\boldsymbol{\pi}_k^{(i)} \mid -\right) \sim \mathrm{Dir}\left(\eta K \pi_{1k}^{(i-1)} + l_{1k}^{(i)} + h_{1k}^{(i+1)}, \cdots, \eta K \pi_{Kk}^{(i-1)} + l_{Kk}^{(i)} + h_{Kk}^{(i+1)}\right). \tag{19}$$

Specifically, for $i = 1$, we have

$$\left(\boldsymbol{\pi}_k^{(1)} \mid -\right) \sim \mathrm{Dir}\left(\nu_1\nu_k + l_{1k}^{(1)} + h_{1k}^{(2)}, \cdots, \xi\nu_k + l_{kk}^{(1)} + h_{kk}^{(2)}, \cdots, \nu_K\nu_k + l_{Kk}^{(1)} + h_{Kk}^{(2)}\right). \tag{20}$$

For sampling η, note that $(h_{k_1k}^{(i)} \mid -) \sim \mathrm{Pois}(-\eta K \pi_{k_1k}^{(i-1)} \ln\left(1 - q_k^{(i)}\right))$, $i = I, \cdots, 2$. Given the prior $\eta \sim \mathrm{Gam}\left(e_0, f_0\right)$, via Poisson-gamma conjugacy, we obtain

$$(\eta \mid -) \sim \mathrm{Gam}\left(e_0 + \sum_{i=2}^{I}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} h_{k_1k_2}^{(i)}, f_0 - K\sum_{i=2}^{I}\sum_{k=1}^{K} \ln\left(1 - q_k^{(i)}\right)\right). \tag{21}$$

Inference for Dirichlet-Gamma-Dirichlet Markov chains. For Dirichlet-Gamma-Dirichlet Markov chains

$$\alpha_{k_1k}^{(i)} \sim \mathrm{Gam}\left(\gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}, c_k^{(i)}\right).$$

By Eq.(10), $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ is multinomial distributed. If we marginalize $(\pi_{1k}^{(i)}, \cdots, \pi_{Kk}^{(i)})$, $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ will be Dirichlet-multinomial distributed. Thus by Lemma 2, for $i = I$, we first sample the auxiliary variables as

$$\left(q_k^{(I)} \mid -\right) \sim \text{Beta}\left(l_{\cdot k}^{(I)}, \alpha_{\cdot k}^{(I)}\right) \text{ and } \left(h_{k_1 k}^{(I)} \mid -\right) \sim \text{CRT}\left(l_{k_1 k}^{(I)}, \alpha_{k_1 k}^{(I)}\right). \tag{22}$$

Similarly, by Eq.(27), $(g_{\cdot 1k}^{(i)}, \cdots, g_{\cdot Kk}^{(i)})$ is also Dirichlet-multinomial distributed. Thus for $i = I-1, \cdots, 2$, we sample the auxiliary variables as

$$\left(q_k^{(i)} \mid -\right) \sim \text{Beta}\left(l_{\cdot k}^{(i)} + g_{\cdot k}^{(i+1)}, \alpha_{\cdot k}^{(i)}\right) \text{ and } \left(h_{k_1 k}^{(i)} \mid -\right) \sim \text{CRT}\left(l_{k_1 k}^{(i)} + g_{\cdot k_1 k}^{(i+1)}, \alpha_{k_1 k}^{(i)}\right). \tag{23}$$

Via Lemma 2, conditioning on $q_k^{(i)}$, we have

$$\left(l_{k_1 k}^{(i)} + g_{\cdot k_1 k}^{(i+1)}\right) \sim \text{NB}\left(\alpha_{k_1 k}^{(i)}, q_k^{(i)}\right).$$

Then via Lemma 1, we obtain

$$h_{k_1 k}^{(i)} \sim \text{Pois}\left(-\alpha_{k_1 k}^{(i)} \ln\left(1 - q_k^{(i)}\right)\right).$$

Thus via Poisson-gamma conjugacy, we obtain

$$\left(\alpha_{k_1 k}^{(i)} \mid -\right) \sim \text{Gam}\left(\gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)} + h_{k_1 k}^{(i)}, c_k^{(i)} - \ln\left(1 - q_k^{(i)}\right)\right). \tag{24}$$

Marginalizing over $\alpha_{k_1 k}^{(i)}$, and via the definition of negative-binomial distribution, we have

$$h_{k_1 k}^{(i)} \sim \text{NB}\left(\gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)}, \frac{-\ln\left(1 - q_k^{(i)}\right)}{c_k^{(i)} - \ln\left(1 - q_k^{(i)}\right)}\right).$$

Then using Lemma 1, we sample

$$\left(g_{k_1 k}^{(i)} \mid -\right) \sim \text{CRT}\left(h_{k_1 k}^{(i)}, \gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)}\right), \tag{25}$$

and obtain

$$g_{k_1 k}^{(i)} \sim \text{Pois}\left(\gamma_k^{(i-1)} \sum_{k2=1}^{K} \psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)} \ln\left(1 - \ln\left(1 - q_k^{(i)}\right) / c_k^{(i)}\right)\right).$$

If we define $g_{k_1 k}^{(i)} = g_{k_1 \cdot k}^{(i)} = \sum_{k2=1}^{K} g_{k_1 k_2 k}^{(i)}$, and augment

$$\left(g_{k_1 1k}^{(i)}, \cdots, g_{k_1 Kk}^{(i)}\right) \sim \text{Mult}\left(g_{k_1 k}^{(i)}, \left(\psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)}\right)_{k_2=1}^{K}\right). \tag{26}$$

By Lemma 3, we have

$$g_{k_1 k_2 k}^{(i)} \sim \text{Pois}\left(\gamma^{(i-1)} \psi_{k k_1 k_2}^{(i-1)} \pi_{k_2 k}^{(i-1)} \ln\left(1 - \ln\left(1 - q_k^{(i)}\right) / c_k^{(i)}\right)\right).$$

Using Lemma 3 and $\sum_{k_1}^{K} \psi_{k k_1 k_2}^{(i-1)} = 1$, we have,

$$\left(g_{\cdot 1k}^{(i)}, \cdots, g_{\cdot Kk}^{(i)}\right) \sim \text{Mult}\left(g_{\cdot k}^{(i)}, \left(\pi_{k_1 k}^{(i-1)}\right)_{k_1=1}^{K}\right), \tag{27}$$

$$\left(g_{1k_2 k}^{(i)}, \cdots, g_{Kk_2 k}^{(i)}\right) \sim \text{Mult}\left(g_{\cdot k_2 k}^{(i)}, \left(\psi_{k k_1 k_2}^{(i-1)}\right)_{k1=1}^{K}\right).$$

Thus by Dirichlet-multinomial conjugacy, for $i = I, \cdots, 2$, we can obtain

$$\left(\left(\psi_{k1k_2}^{(i-1)}, \cdots, \psi_{kKk_2}^{(i-1)}\right) \mid -\right) \sim \text{Dir}\left(\epsilon_0 + g_{1k_2 k}^{(i)}, \cdots, \epsilon_0 + g_{Kk_2 k}^{(i)}\right), \tag{28}$$

$$\left(\boldsymbol{\pi}_k^{(i-1)} \mid -\right) \sim \text{Dir}\left(\alpha_{1k}^{(i-1)} + l_{1k}^{(i-1)} + g_{\cdot 1k}^{(i)}, \cdots, \alpha_{Kk}^{(i-1)} + l_{Kk}^{(i-1)} + g_{\cdot Kk}^{(i)}\right). \tag{29}$$

For sampling $\gamma_k^{(i-1)}$, note that by Eq.(26) and $\sum_{k_1}^K \psi_{kk_1k_2}^{(i-1)} = 1$, we have

$$g_{\cdot k}^{(i)} = \sum_{k_1=1}^K g_{k_1k}^{(i)} \text{ and } g_{\cdot k}^{(i)} \sim \text{Pois}\left(\gamma_k^{(i-1)} \ln\left(1 - \ln\left(1 - q_k^{(i)}\right)/c_k^{(i)}\right)\right). \tag{30}$$

Thus via Poisson-gamma conjugacy, we obtain

$$\left(\gamma_k^{(i-1)} \mid -\right) \sim \text{Gam}\left(\epsilon_0 + g_{\cdot k}^{(i)}, \epsilon_0 + \ln\left(1 - \ln\left(1 - q_k^{(i)}\right)\right)\right). \tag{31}$$

By gamma-gamma conjugacy, we have

$$\left(c_k^{(i)} \mid -\right) \sim \text{Gam}\left(\epsilon_0 + \gamma_k^{(i-1)}, \epsilon_0 + \sum_{k_1=1}^K \alpha_{k_1k}^{(i)}\right). \tag{32}$$

Inference for Dirichlet-Randomized-Gamma-Dirichlet Markov chains. For Dirichlet-Randomized-Gamma-Dirichlet Markov chains,

$$\alpha_{k_1k}^{(i)} \sim \text{RG1}\left(\epsilon^\alpha, \gamma^{(i-1)} \sum_{k2=1}^K \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}, c_k^{(i)}\right),$$

which can be equivalently represented as

$$\alpha_{k_1k}^{(i)} \sim \text{Gam}\left(g_{k_1k}^{(i)} + \epsilon^\alpha, c_k^{(i)}\right), \text{ and } g_{k_1k}^{(i)} = \text{Pois}\left(\gamma^{(i-1)} \sum_{k2=1}^K \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}\right).$$

By Eq.(10), $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ is multinomial distributed. If we marginalize $(\pi_{1k}^{(i)}, \cdots, \pi_{Kk}^{(i)})$, $(l_{1k}^{(i)}, \cdots, l_{Kk}^{(i)})$ will be Dirichlet-multinomial distributed. Thus by Lemma 2, for $i = I$, we first sample the auxiliary variables as

$$\left(q_k^{(I)} \mid -\right) \sim \text{Beta}\left(l_{\cdot k}^{(I)}, \alpha_{\cdot k}^{(I)}\right) \text{ and } \left(h_{k_1k}^{(I)} \mid -\right) \sim \text{CRT}\left(l_{k_1k}^{(I)}, \alpha_{k_1k}^{(I)}\right). \tag{33}$$

Similarly, by Eq.(39), $(g_{\cdot 1k}^{(i)}, \cdots, g_{\cdot Kk}^{(i)})$ is also Dirichlet-multinomial distributed. Thus for $i = I-1, \cdots, 2$, we sample the auxiliary variables as

$$\left(q_k^{(i)} \mid -\right) \sim \text{Beta}\left(l_{\cdot k}^{(i)} + g_{\cdot k}^{(i+1)}, \alpha_{\cdot k}^{(i)}\right) \text{ and } \left(h_{k_1k}^{(i)} \mid -\right) \sim \text{CRT}\left(l_{k_1k}^{(i)} + + g_{\cdot k_1k}^{(i+1)}, \alpha_{k_1k}^{(i)}\right). \tag{34}$$

Via Lemma 2, conditioning on $q_k^{(i)}$, we have

$$\left(l_{k_1k}^{(i)} + g_{\cdot k_1k}^{(i+1)}\right) \sim \text{NB}\left(\alpha_{k_1k}^{(i)}, q_k^{(i)}\right).$$

Then via Lemma 1, we obtain

$$h_{k_1k}^{(i)} \sim \text{Pois}\left(-\alpha_{k_1k}^{(i)} \ln\left(1 - q_k^{(i)}\right)\right).$$

Via Poisson-gamma conjugacy, we first sample

$$\left(\alpha_{k_1k}^{(i)} \mid -\right) \sim \text{Gam}\left(g_{k_1k}^{(i)} + \epsilon^\alpha + h_{k_1k}^{(i)}, c_k^{(i)} - \ln\left(1 - q_k^{(i)}\right)\right). \tag{35}$$

If $\epsilon^\alpha > 0$, we can sample the posterior of $g_{k_1k}^{(i)}$ via

$$\left(g_{k_1k}^{(i)} \mid -\right) \sim \text{Bessel}\left(\epsilon^\alpha - 1, 2\sqrt{\alpha_{k_1k}^{(i)} c_k^{(i)} \gamma_k^{(i-1)} \sum_{k2=1}^K \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}}\right), \tag{36}$$

where $\text{Bessel}(\cdot)$ denotes Bessel distribution. If $\epsilon^\alpha = 0$, we sample $g_{k_1k}^{(i)}$ via

$$\left(g_{k_1k}^{(i)} \mid -\right) \sim \begin{cases} \text{Pois}\left(\frac{c_k^{(i)} \gamma_k^{(i-1)} \sum_{k2=1}^K \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}}{c_k^{(i)} - \ln\left(1 - q_k^{(i)}\right)}\right) & \text{if } h_{k_1k}^{(i)} = 0 \\ \text{SCH}\left(h_{k_1k}^{(i)}, \frac{c_k^{(i)} \gamma_k^{(i-1)} \sum_{k2=1}^K \psi_{kk_1k_2}^{(i-1)} \pi_{k_2k}^{(i-1)}}{c_k^{(i)} - \ln\left(1 - q_k^{(i)}\right)}\right) & \text{otherwise,} \end{cases} \tag{37}$$

where $\mathrm{SCH}(\cdot)$ denotes the shifted confluent hypergeometric distribution [16].

Defining $g^{(i)}_{k_1k}=g^{(i)}_{k_1\cdot k}=\sum_{k2=1}^{K}g^{(i)}_{k_1k_2k}$, we first augment

$$\left(g^{(i)}_{k_11k},\cdots,g^{(i)}_{k_1Kk}\right)\sim\mathrm{Mult}\left(g^{(i)}_{k_1k},\left(\psi^{(i-1)}_{kk_1k_2}\pi^{(i-1)}_{k_2k}\right)_{k_2=1}^{K}\right). \tag{38}$$

By Lemma 3, we have

$$g^{(i)}_{k_1k_2k}\sim\mathrm{Pois}\left(\gamma^{(i-1)}\psi^{(i-1)}_{kk_1k_2}\pi^{(i-1)}_{k_2k}\right),$$

and because $\sum_{k_1}^{K}\psi^{(i-1)}_{kk_1k_2}=1$, we have

$$\begin{aligned}\left(g^{(i)}_{\cdot 1k},\cdots,g^{(i)}_{\cdot Kk}\right)&\sim\mathrm{Mult}\left(g^{(i)}_{\cdot k},\left(\pi^{(i-1)}_{k_1k}\right)_{k_1=1}^{K}\right),\text{ and}\\ \left(g^{(i)}_{1k_2k},\cdots,g^{(i)}_{Kk_2k}\right)&\sim\mathrm{Mult}\left(g^{(i)}_{\cdot k_2k},\left(\psi^{(i-1)}_{kk_1k_2}\right)_{k1=1}^{K}\right).\end{aligned} \tag{39}$$

Thus by Dirichlet-multinomial conjugacy, for $i=I,\cdots,2$, we have

$$\left(\left(\psi^{(i-1)}_{k1k_2},\cdots,\psi^{(i-1)}_{kKk_2}\right)\mid-\right)\sim\mathrm{Dir}\left(\epsilon_0+g^{(i)}_{1k_2k},\cdots,\epsilon_0+g^{(i)}_{Kk_2k}\right), \tag{40}$$

$$\left(\boldsymbol{\pi}^{(i-1)}_{k}\mid-\right)\sim\mathrm{Dir}\left(\alpha^{(i-1)}_{1k}+l^{(i-1)}_{1k}+g^{(i)}_{\cdot 1k},\cdots,\alpha^{(i-1)}_{Kk}+l^{(i-1)}_{Kk}+g^{(i)}_{\cdot Kk}\right). \tag{41}$$

Via Poisson-gamma conjugacy, we obtain

$$\left(\gamma^{(i-1)}_{k}\mid-\right)\sim\mathrm{Gam}\left(\epsilon_0+g^{(i)}_{\cdot k},\epsilon_0+1\right). \tag{42}$$

By gamma-gamma conjugacy, we have

$$\left(c^{(i)}_{k}\mid-\right)\sim\mathrm{Gam}\left(\epsilon_0+\gamma^{(i-1)}_{k},\epsilon_0+\sum_{k_1=1}^{K}\alpha^{(i)}_{k_1k}\right). \tag{43}$$

Specifically, for $i=1$, we have $\alpha^{(1)}_{k_1k}=\nu_{k_1}\nu_k$, if $k_1\neq k$. And $\alpha^{(1)}_{k_1k}=\xi\nu_k$, if $k_1=k$.

Sampling ν_k and ξ : As we sample $\boldsymbol{\Pi}^{(i)}$, by the definition of Dirichlet-multinomial distribution, we obtain

$$\left(l^{(1)}_{1k}+g^{(2)}_{\cdot 1k},\cdots,l^{(1)}_{Kk}+g^{(2)}_{\cdot Kk}\right)\sim\mathrm{DirMult}\left(\nu_1\nu_K,\cdots,\xi\nu_k,\cdots,\nu_K\nu_k\right),$$

where $l^{(1)}_{k_1k}=\sum_{t=1}^{M}l^{(t)}_{k_1k}$. In particular, with a little abuse of notation here, for Dir-Dir construction, we take $g^{(2)}_{\cdot k_1k}=h^{(2)}_{k_1k}$. We first sample

$$\left(h^{(1)}_{k_1k}\mid-\right)\sim\begin{cases}\mathrm{CRT}\left(l^{(1)}_{k_1k}+g^{(2)}_{\cdot k_1k},\nu_{k_1}\nu_k\right) & k_1\neq k\\ \mathrm{CRT}\left(l^{(1)}_{k_1k}+g^{(2)}_{\cdot k_1k},\xi\nu_k\right) & k_1=k.\end{cases} \tag{44}$$

Then we sample

$$q^{(1)}_{k}\sim\mathrm{Beta}\left(l^{(1)}_{\cdot k}+g^{(2)}_{\cdot k},\nu_k\left(\sum_{k_1\neq k}\nu_{k1}+\xi\right)\right). \tag{45}$$

We further introduce

$$\begin{aligned}n_k=&h^{(1)}_{kk}+\sum_{k_1\neq k}h^{(1)}_{k_1k}+\sum_{k_2\neq k}h^{(1)}_{kk_2}+l^{(1)}_{k\cdot},\text{ and}\\ \rho_k=&\tau_0\zeta^{(1)}-\ln\left(1-q^{(1)}_{k}\right)\left(\xi+\sum_{k_1\neq k}\nu_{k_1}\right)-\sum_{k_2\neq k}\ln\left(1-q^{(1)}_{k_2}\right)\nu_{k_2}.\end{aligned}$$

Via Poisson-gamma conjugacy, we have

$$(\xi \mid -) \sim \mathrm{Gam}\left(\frac{\gamma_0}{K} + \sum_k h_{kk}^{(1)}, \beta - \sum_k \nu_k \ln\left(1 - q_k^{(1)}\right)\right), \tag{46}$$

$$(\nu_k \mid -) \sim \mathrm{Gam}\left(\frac{\gamma_0}{K} + n_k, \beta + \rho_k\right). \tag{47}$$

Sampling $\delta^{(t)}$ and β : Via Poisson-gamma conjugacy

$$\left(\delta^{(t)} \mid -\right) \sim \mathrm{Gam}\left(\epsilon_0 + \sum_{v=1}^{V} y_v^{(t)}, \epsilon_0 + \sum_{k=1}^{K} \theta_k^{(t)}\right). \tag{48}$$

And by gamma-gamma conjugacy, we obtain

$$(\beta \mid -) \sim \mathrm{Gam}\left(\epsilon_0 + \gamma_0, \epsilon_0 + \sum_{k=1}^{K} \nu_k\right). \tag{49}$$

The full procedure of our Gibbs sampling algorithms are summarized in Algorithm 1, Algorithm 2 and Algorithm 3.

Algorithm 1 Gibbs sampling algorithm for NS-PGDS (Dir-Dir Markov construction)

Input: observed count sequence $\{\boldsymbol{y}^{(t)}\}_{t=1}^{T}$, iterations $\mathcal{J}$.
Initialize the model's rank K, hyperparameters $\gamma_0, \epsilon_0, e_0, f_0$.
for $iter = 1$ to $\mathcal{J}$ **do**
  Sample $\{y_{vk}^{(t)}\}_{v,k}$ via Eq.(6).
  Sample $\{\boldsymbol{\phi_k}\}_k$ via Eq.(7).
  Sample $\{\delta^{(t)}\}_t$ via Eq.(48). Update $\zeta^{(t)}$ as
    $\zeta^{(T+1)} = 0, \quad \zeta^{(t)} = \ln\left(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)}\right),\ t = T, \cdots, 1.$
  Set $l_{\cdot k}^{(T+1)} = 0$.
  for $t = T$ to 2 **do**
    Sample $\{l_{k\cdot}^{(t)}\}_k$ and $\{l_{kk_2}^{(t)}\}_{k,k_2}$ via Eq.(8) and Eq.(9) respectively.
  end for
  for $t = 1$ to T **do**
    Sample $\{\theta_k^{(t)}\}_k$ via Eq.(12) and Eq.(13).
  end for
  for $i = 1$ to I **do**
    Sample $\{q_k^{(i)}\}_k$ and $\{h_{k_1k}^{(i)}\}_{k_1,k}$ via Eq.(16), Eq.(44) and Eq.(45).
    Sample $\{\boldsymbol{\pi}_k^{(i)}\}_k$ via Eq.(14) and Eq.(19).
    Sample η via Eq.(21).
  end for
  Sample ξ, $\{\nu_k\}_k$, β via Eq.(46), Eq.(47) and Eq.(49) respectively.
end for
Output posterior means: $\{\theta_k^{(1:T)}\}_k$, $\{\boldsymbol{\phi}_k\}_k$, $\{\boldsymbol{\pi}_k^{(i)}\}_k$, $\delta^{(1:T)}$, ξ, $\{\nu_k\}_k$, β.

Algorithm 2 Gibbs sampling algorithm for NS-PGDS (Dir-Gam-Dir Markov construction)

Input: observed count sequence $\{\boldsymbol{y}^{(t)}\}_{t=1}^{T}$, iterations $\mathcal{J}$.
Initialize the model's rank K, hyperparameters $\gamma_0, \epsilon_0, e_0, f_0$.
for $iter = 1$ to $\mathcal{J}$ **do**
  Sample $\{y_{vk}^{(t)}\}_{v,k}$ via Eq.(6).
  Sample $\{\boldsymbol{\phi_k}\}_k$ via Eq.(7).
  Sample $\{\delta^{(t)}\}_t$ via Eq.(48). Update $\zeta^{(t)}$ as
    $\zeta^{(T+1)} = 0, \quad \zeta^{(t)} = \ln\left(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)}\right),\ t = T, \cdots, 1.$
  Set $l_{\cdot k}^{(T+1)} = 0$.
  for $t = T$ to 2 **do**
    Sample $\{l_{k\cdot}^{(t)}\}_k$ and $\{l_{kk_2}^{(t)}\}_{k,k_2}$ via Eq.(8) and Eq.(9) respectively.
  end for
  for $t = 1$ to T **do**
    Sample $\{\theta_k^{(t)}\}_k$ via Eq.(12) and Eq.(13).
  end for
  for $i = 1$ to I **do**
    Sample $\{\alpha_{k_1k}^{(i)}\}_{k_1,k}$ and $\{c_k^{(i)}\}_k$ via Eq.(24) and Eq.(32).
    Sample $\{q_k^{(i)}\}_k$ and $\{h_{k_1k}^{(i)}\}_{k_1,k}$ via Eq.(22), Eq.(23), Eq.(44) and Eq.(45).
    Sample $\{g_{k_1k}\}_{k_1,k}$ and $\{g_{k_1k_2k}\}_{k_1,k_2,k}$ via Eq.(25) and Eq.(26) respectively.
    Sample $\{\psi_{kk_1k_2}\}_{k,k_1,k_2}$ via Eq.(28).
    Sample $\{\gamma_k^{(i)}\}_k$ via Eq.(31).
    Sample $\{\boldsymbol{\pi}_k^{(i)}\}_k$ via Eq.(14) and Eq.(29).
  end for
  Sample ξ, $\{\nu_k\}_k$, β via Eq.(46), Eq.(47) and Eq.(49) respectively.
end for
Output posterior means: $\{\theta_k^{(1:T)}\}_k$, $\{\boldsymbol{\phi}_k\}_k$, $\{\boldsymbol{\pi}_k^{(i)}\}_k$, $\delta^{(1:T)}$, ξ, $\{\nu_k\}_k$, β.

Algorithm 3 Gibbs sampling algorithm for NS-PGDS (PR-Gam-Dir Markov construction)

Input: observed count sequence $\{\boldsymbol{y}^{(t)}\}_{t=1}^{T}$, iterations $\mathcal{J}$.
Initialize the model's rank K, hyperparameters $\gamma_0, \epsilon_0, e_0, f_0$.
for $iter = 1$ to $\mathcal{J}$ **do**
  Sample $\{y_{vk}^{(t)}\}_{v,k}$ via Eq.(6).
  Sample $\{\boldsymbol{\phi_k}\}_k$ via Eq.(7).
  Sample $\{\delta^{(t)}\}_t$ via Eq.(48). Update $\zeta^{(t)}$ as
    $\zeta^{(T+1)} = 0, \quad \zeta^{(t)} = \ln\left(1 + \frac{\delta^{(t)}}{\tau_0} + \zeta^{(t+1)}\right), \ t = T, \cdots, 1.$
  Set $l_{\cdot k}^{(T+1)} = 0$.
  for $t = T$ to 2 **do**
    Sample $\{l_{k\cdot}^{(t)}\}_k$ and $\{l_{kk_2}^{(t)}\}_{k,k_2}$ via Eq.(8) and Eq.(9) respectively.
  end for
  for $t = 1$ to T **do**
    Sample $\{\theta_k^{(t)}\}_k$ via Eq.(12) and Eq.(13).
  end for
  for $i = 1$ to I **do**
    Sample $\{\alpha_{k_1k}^{(i)}\}_{k_1,k}$ and $\{c_k^{(i)}\}_k$ via Eq.(33) and Eq.(43).
    Sample $\{q_k^{(i)}\}_k$ and $\{h_{k_1k}^{(i)}\}_{k_1,k}$ via Eq.(33), Eq.(34), Eq.(44) and Eq.(45).
    Sample $\{g_{k_1k}\}_{k_1,k}$ via Eq.(36) and Eq.(37).
    Sample $\{g_{k_1k_2k}\}_{k_1,k_2,k}$ via Eq.(38).
    Sample $\{\gamma_k^{(i)}\}_k$ via Eq.(42).
    Sample $\{\psi_{kk_1k_2}\}_{k,k_1,k_2}$ via Eq.(40).
    Sample $\{\boldsymbol{\pi}_k^{(i)}\}_k$ via Eq.(14), and Eq.(41).
  end for
  Sample ξ, $\{\nu_k\}_k$, β via Eq.(46), Eq.(47) and Eq.(49) respectively.
end for
Output posterior means: $\{\theta_k^{(1:T)}\}_k$, $\{\boldsymbol{\phi}_k\}_k$, $\{\boldsymbol{\pi}_k^{(i)}\}_k$, $\delta^{(1:T)}$, ξ, $\{\nu_k\}_k$, β.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions of this paper are the constructions of Poisson-Gamma dynamical systems with non-stationary transition dynamics and the corresponding Gibbs sampler. The constructions can be found in sec.2 and sec.3 and the derivation of Gibbs sampler can be found in sec.4 and the appendix. The experiments have demonstrated the effectiveness and features of the proposed model.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: As we discussed in the conclusion part, the length of each sub-interval is a constant and is treated as a hyper-parameter of the model. In the future work, we plan to design a method that can find the point of change and thus the length of each sub-interval can be determined automatically.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The derivation of the Gibbs sampler is the main theoretical part of this paper which can be found in sec.4 and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We carefully described the proposed model in sec.2 and sec.3 and the experiment details can be found in sec.6.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The authors will release the data and code as soon as possible if this paper could be accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment details can be found in sec.6.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports the predictive performance of the proposed model and the corresponding standard deviation. The results are computed by running the Gibbs sampling several times from different initialization.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: The experiments are conducted on a server with an Intel(R) Xeon(R) CPU E5-2699Cv4 @ 2.20GHz and 64G RAM.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have checked the NeurIPS Code of Ethics and make sure this work is with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: For positive societal impacts, we have discussed in conclusion section for the potential application for textual analysis and social networks. And the authors think this work does not have potential negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.