Negative-Binomial Randomized Gamma Dynamical Systems for Heterogeneous Overdispersed Count Time Sequences

Rui Huang[1,2,3,4] , **Sikun Yang**[1,2,3,4*] , **Heinz Koeppl**[5]
[1]School of Computing and Information Technology, Great Bay University, 523000 Dongguan, China
[2]Great Bay Institute for Advanced Study, Great Bay University
[3]Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems, Great Bay University
[4]Dongguan Key Laboratory for Data Science and Intelligent Medicine, Great Bay University
[5]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt
huangrui@pku.edu.cn, sikunyang@gbu.edu.cn, heinz.koeppl@bcs.tu-darmstadt.de

Abstract

Modeling count-valued time sequences has been receiving growing interests because count time sequences naturally arise in physical and social domains. Poisson gamma dynamical systems (PGDSs) are newly-developed methods, which can well capture the expressive latent transition structure and *bursty* dynamics behind count sequences. In particular, PGDSs demonstrate superior performance in terms of data imputation and prediction, compared with canonical linear dynamical system (LDS) based methods. Despite these advantages, PGDS cannot capture the *heterogeneous* overdispersed behaviours of the underlying dynamic processes. To mitigate this defect, we propose a negative-binomial-randomized gamma Markov process, which not only significantly improves the predictive performance of the proposed dynamical system, but also facilitates the fast convergence of the inference algorithm. Moreover, we develop methods to estimate both factor-structured and graph-structured transition dynamics, which enable us to infer more explainable latent structure, compared with PGDSs. Finally, we demonstrate the explainable latent structure learned by the proposed method, and show its superior performance in imputing missing data and forecasting future observations, compared with the related models.

1 Introduction

Count time sequences, naturally arise in many domains such as text mining [Blei and Lafferty, 2006; Wang *et al.*, 2008; Rudolph and Blei, 2018; Acharya *et al.*, 2018; Dieng *et al.*, 2019], cell genomic analysis [Levitin *et al.*, 2019; Tong *et al.*, 2020; Jones *et al.*, 2023], population movement forecasting [Sheldon *et al.*, 2013; Stuart and Wolfram, 2020; Roy and Dunson, 2020], and etc. Modeling count sequences has been drawing increasing research attention because these real-world count data usually exhibit *bursty* and *overdispersed* behaviours, which cannot be well-captured by canonical linear dynamical systems (LDSs) [Ghahramani and Roweis, 1998]. In addition, some previous works use extended rank likelihood functions [Han *et al.*, 2014] which link the count observations to latent continuous dynamics to model count time sequences. Nonetheless, the extended rank likelihood functions cannot faithfully capture *bursty* dynamics underlying real-world count sequences. Meanwhile, the extended rank likelihood functions often require an approximate inference scheme, and thus scale poorly with high-dimensional count sequences, such as single-cell RNA sequencing data [Chandra *et al.*, 2023]. Notably, some recent works [Acharya *et al.*, 2015; Schein *et al.*, 2016a; Schein *et al.*, 2016b; Schein *et al.*, 2019] model sequential count observations using gamma Poisson family distributions. More specifically, [Acharya *et al.*, 2015] develops a gamma Markov process to capture continuous dynamics underlying count-valued sequences. In particular, the number of latent factors behind high-dimensional count data, can be appropriately determined by the gamma process prior, in a Bayesian non-parametric manner. Following the success of [Acharya *et al.*, 2015], Schein *et al.* [2016a] study a Poisson gamma dynamical system, in which a transition kernel is designed to capture how the latent dimensions interact with each other to model complicated observed dynamics. Another appealing aspect of the Poisson gamma dynamic model is that the posterior simulation can be performed using a tractable-yet-efficient Gibbs sampling algorithm via Poisson-Logarithm data augmentation strategy [Zhou and Carin, 2012; Zhou and Carin, 2015]. Hence, the Poisson gamma dynamic models [Acharya *et al.*, 2015; Schein *et al.*, 2016a; Schein *et al.*, 2019] are in particular well-fitted to impute missing entries, to predict future unseen observations and to estimate uncertainties.

Despite these advantages, these models still cannot well capture the *heterogeneous* overdispersion effects of the latent dynamic processes behind count observations. For instance, international event data, usually consists of multiple latent dynamic processes, which often change rapidly with the different magnitudes [King, 2001; Stewart, 2014]. To capture such *heterogeneous* overdispersed behaviours, we de-

[*] Corresponding Author.

velop a negative-binomial-randomized gamma Markov chain structure, which not only greatly enhances the model flexibility, but also facilitates the fast convergence of the derived Gibbs sampling algorithms. Moreover, the transition dynamics behind real-world high-dimensional count data, are often *sparse*, and exhibit a certain amount of graph structure. Hence, we propose to learn the graph-structured transition dynamics using relational gamma processes [Zhou, 2015]. To the best of our knowledge, this is the first attempt to learn the latent graph-structured transition dynamics under the Poisson gamma dynamical system.

The main contributions of the paper include: 1) A negative-binomial-randomized gamma Markov process (NBRGMP) is proposed to estimate the *heterogeneous* overdispersion effects of the latent dimensions underlying sequential count observations; 2) Relational gamma processes are thoroughly studied to learn both factor-structured and graph-structured transition dynamics, which renders the estimated latent structure more explainable, compared with transition structure inferred using non-informative priors; 3) Although the proposed NBRGMP and its factor-structured and graph-structured extensions are intractable, simple-yet-efficient Gibbs sampling algorithms are developed via Negative-binomial data augmentation strategies to perform inference; 4) Extensive experiments are conducted to illustrate the explainable transition structure learned by the proposed model. We demonstrate the superior performance of the proposed method in missing data imputation and future snapshot forecasting, with several related works.

2 Preliminary

Suppose we have a sequentially-observed count data over time interval $[0, T]$ specified by $\mathbf{N} = (\mathbf{n}_1, \dots, \mathbf{n}_V)^{\mathrm{T}}$ of V dimensions, where $\mathbf{n}_v = (n_v^{(1)}, \dots, n_v^{(T)})^{\mathrm{T}}$ with $n_v^{(t)}$ denoting the v-th observation at time t. The Poisson gamma dynamical system [Schein *et al.*, 2016a] models the count $n_v^{(t)}$ as

$$n_v^{(t)} \sim \mathrm{Pois}(\delta^{(t)} \sum\nolimits_{k=1}^{K} \phi_{vk} \theta_k^{(t)}), \tag{1}$$

where $\theta_k^{(t)}$ captures the strength of latent component k at time t, and ϕ_{vk} represents the involvement degree of dimension v in latent component k. To model the underlying dynamics, the PGDS assumes that the latent components evolve over time according to a gamma Markov chain structure as

$$\theta_k^{(t)} \sim \mathrm{Gam}(\tau_0 \sum\nolimits_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}, \tau_0), \tag{2}$$

where the latent components $\boldsymbol{\theta}^{(t-1)} = (\theta_1^{(t-1)}, \dots, \theta_K^{(t-1)})^{\mathrm{T}}$ evolve over time through the transition matrix $\mathbf{\Pi}$. The $\theta_k^{(t-1)}$ captures how strongly the k-th latent component activates at time $t-1$, and π_{kk_2} models how strongly the k_2-th component $\theta_{k_2}^{(t-1)}$ at time $t-1$ affect the k-th component $\theta_k^{(t)}$ at time t. Eq. 2 naturally defines a gamma Markov chain structure. The expectation and variance of the gamma Markov chain can be calculated respectively as $\mathsf{E}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathbf{\Pi}] = \mathbf{\Pi}\boldsymbol{\theta}^{(t-1)}$ and $\mathsf{Var}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}, \mathbf{\Pi}] = (\mathbf{\Pi}\boldsymbol{\theta}^{(t-1)})\tau_0^{-1}$, where τ_0 controls the variance of $\boldsymbol{\theta}^{(t)}$.

Schein *et al.* [2019] further develop a Poisson-randomized gamma Markov chain (PRGMC) structure specified by

$$\theta_k^{(t)} \sim \mathrm{Gam}(\epsilon_0^{(\theta)} + h_k^{(t)}, \tau),\ h_k^{(t)} \sim \mathrm{Pois}(\tau \sum\nolimits_{k_2} \pi_{kk_2} \theta_{k_2}^{(t-1)}).$$

By marginalizing out the Poisson latent states $h_k^{(t)}$, we have a continuous-valued dynamical system given by

$$\theta_k^{(t)} \sim \mathrm{RG1}(\epsilon_0^{(\theta)}, \tau \sum\nolimits_{k_2} \pi_{kk_2} \theta_{k_2}^{(t-1)}, \tau),$$

where RG1 is the randomized gamma distribution of the first type. The marginal expectation and variance of the PRGMC is $\mathsf{E}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathbf{\Pi}] = \mathbf{\Pi}\boldsymbol{\theta}^{(t-1)} + \epsilon_0^{(\theta)}\tau^{-1}$ and $\mathsf{Var}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \mathbf{\Pi}] = 2\mathbf{\Pi}\boldsymbol{\theta}^{(t-1)}\tau^{-1} + \epsilon_0^{(\theta)}\tau^{-2}$, respectively.

3 The Proposed Model

In this section we will introduce the novel negative-binomial-randomized gamma Markov chain structure to capture the *heterogeneous* overdispersion effects of the latent dimensions behind count data. Then we shall describe how to learn explainable latent transition structure with relational gamma processes. The proposed negative-binomial-randomized gamma dynamical system is defined by

$$n_v^{(t)} \sim \mathrm{Pois}(\delta^{(t)} \sum\nolimits_{k=1}^{K} \lambda_k \phi_{vk} \theta_k^{(t)}), \tag{3}$$

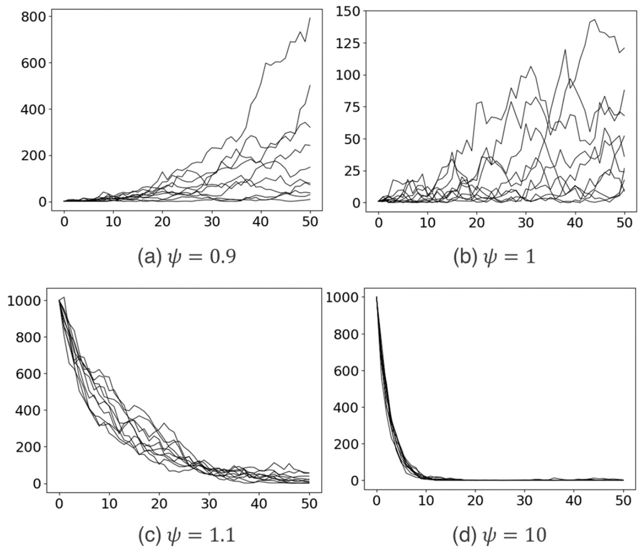
where $\delta^{(t)}$ is a nonnegative multiplicative term capturing time-dependent bursty dynamics. We place a gamma prior on $\delta^{(t)}$ as $\delta^{(t)} \sim \mathrm{Gam}(\epsilon_0, \epsilon_0)$, and let $\delta^{(t)} = \delta$ if the generative process (Eq. 3) is stationary over time. Here $\boldsymbol{\phi}_k = (\phi_{1k}, \phi_{2k}, \dots, \phi_{Vk})^{\mathrm{T}}$ denotes the loading coefficient of k-th latent component, and λ_k denotes the weight of k-th latent component. To ensure model identifiability, we require $\sum_v \phi_{vk} = 1$ and thus have a Dirichlet prior over $\boldsymbol{\phi}_k$ given by $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\epsilon_0, ..., \epsilon_0)$. More specifically, we draw λ_k from a hierarchical prior as $\lambda_k \sim \mathrm{Gam}(\frac{\epsilon_0^{(\lambda)}}{K} + g_k, \beta)$, in which $g_k \sim \mathrm{Pois}(\frac{\gamma}{K})$. We specify gamma priors over γ and β as $\gamma \sim \mathrm{Gam}(\epsilon_0, \epsilon_0), \beta \sim \mathrm{Gam}(\epsilon_0, \epsilon_0)$. Note that as $K \to \infty$, the summation of the weight expectation remains finite and fixed, i.e., $\sum_{k=1}^{\infty} \mathsf{E}[\lambda_k] = \beta^{-1}(\epsilon_0^{(\lambda)} + \gamma)$. Hence, this hierarchical prior enables us to effectively estimate a finite number of latent factors that are representative to capture the temporal dynamics.

3.1 Negative-Binomial Randomized Gamma Markov Processes

To capture the *heterogeneous* overdispersed behaviors of the latent dimensions behind count sequences, we introduce a negative-binomial randomized gamma Markov process (NBRGMP) specified by

$$\begin{aligned} &\theta_k^{(t)} \sim \mathrm{Gam}(\epsilon_0^{(\theta)} + h_k^{(t)}, \tau), \\ &h_k^{(t)} \sim \mathrm{NB}(\tau \sum\nolimits_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}, \frac{\psi}{1+\psi}), \end{aligned} \tag{4}$$

where we set $\theta_k^{(0)} = \lambda_k$, and $\theta_k^{(t)}$ is gamma distributed with shape parameter $\epsilon_0^{(\theta)} + h_k^{(t)}$ where $\epsilon_0^{(\theta)} \geq 0$, and the rate



(a) $\psi = 0.9$

(b) $\psi = 1$

(c) $\psi = 1.1$

(d) $\psi = 10$

Figure 1: The realizations of the negative-binomial-randomized gamma Markov processes defined in Eq. 4. Here $\epsilon_0^{(\theta)}$ and τ were set to 0 and 1, respectively. The initial values of the NBRGMPs in (a) and (b), were set to 1, (c) and (d) were set to 1000, and the chains were simulated until $t = 50$. Each subplot contains ten independent realizations.

parameter τ. Here, instead of specifying a Poisson prior as did in [Schein *et al.*, 2019], we draw the intermediate latent state $h_k^{(t)}$ from a negative-binomial distribution to enhance the flexibility of $\theta_k^{(t)}$, which enable us to estimate the heterogeneous overdispersed behaviours of the latent dynamic processes. More specifically, the marginal expectation and variance of the NBRGMP can be calculated by iteration as

$$\mathsf{E}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}] = \epsilon_0^{(\theta)}\tau^{-1} + \frac{\boldsymbol{\Pi}\boldsymbol{\theta}^{(t-1)}}{\psi},$$

$$\mathsf{Var}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}] = \epsilon_0^{(\theta)}\tau^{-2} + \frac{(1+2\psi)\boldsymbol{\Pi}\boldsymbol{\theta}^{(t-1)}}{\psi^2\tau},$$

respectively. We note that both the concentration parameter τ and hyperparameter $\epsilon_0^{(\theta)}$ appear in the additive term of the expectation and variance, which can be concealed by letting $\epsilon_0^{(\theta)} = 0$. The hyperparameter ψ plays a crucial role in controlling the variance of the NBRGMP. More specifically, when $\psi \in (0, 1)$ the values of $\boldsymbol{\theta}^{(t)}$ will fluctuate dramatically because of its large expectation and variance. When $\psi = 1$, the expectation of the NBRGMP will be the same with the PRGMC (as discussed in Sec. 2), while the variance of the NBRGMP will be three times of the variance of the PRGMC, which thus allows the proposed dynamical system to capture reasonable heterogeneous overdispersed behaviors. When $\psi \in (1, 1 + \sqrt{2})$, the expectation of the NBRGMP will tend to be smaller compared with the expectation of the PRGMC, which will be more suitable to capture sparse counts. Meanwhile, the variance of the NBRGMP still allows us to capture a limited range of overdispersion effects. If $\psi \geq 1+\sqrt{2}$, both expectation and variance converge to zeros as ψ goes to infinity. Fig. 1 plots the realizations of the NBRGMP by varying the parameter ψ. Note that the negative-binomial distributed latent state $h_k^{(t)}$ can be equivalently drawn from a gamma-Poisson mixture as

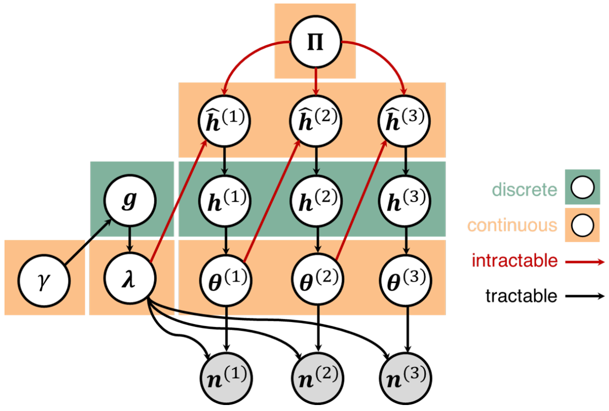


Figure 2: The hierarchical structure of the NBRGMP. The red arrows indicate intractable dependencies that require data augmentation schemes for posterior inference.

$$h_k^{(t)} \sim \mathrm{Pois}(\hat{h}_k^{(t)}), \ \ \hat{h}_k^{(t)} \sim \mathrm{Gam}(\tau \sum\nolimits_{k_2=1}^{K} \pi_{kk_2}\theta_{k_2}^{(t-1)}, \psi).$$

Fig. 2 shows the graphical representation of the developed NBRGMP. When $\tau \sum_{k_2} \pi_{kk_2}\theta_{k_2}^{(t-1)} \to \infty$, the $h_k^{(t)}$ is approximately characterized by $\mathrm{Pois}(\frac{\tau}{\psi} \sum_{k_2} \pi_{kk_2}\theta_{k_2}^{(t-1)})$. Hence, by marginalizing the Poisson distributed latent states $h_k^{(t)}$ from Eq. 4, the negative-binomial randomized gamma dynamical system can be equivalently represented by randomized gamma distribution of the first type as

$$\theta_k^{(t)} \sim \mathrm{RG1}(\epsilon_0^{(\theta)}, \frac{\tau}{\psi} \sum\nolimits_{k_2=1}^{K} \pi_{kk_2}\theta_{k_2}^{(t-1)}, \tau).$$

3.2 Factor-structured Transition Dynamics

We first propose to learn the latent factor structure behind transition dynamics. To that end, we specify a hierarchical Dirichlet prior over $\boldsymbol{\pi}_k$ as $\boldsymbol{\pi}_k \sim \mathrm{Dir}(a_{1k}, \ldots, a_{Kk})$,where $\mathbf{a}_k = (a_{1k}, \ldots, a_{Kk})^{\mathrm{T}}$ is the hyper-parameter. Our goal here is to capture the correlation structure between the latent dimensions of the transition kernel. Thus, we model the hyperparameter $\mathbf{A} = [a_{k_1k_2}]_{k_1,k_2}^{K}$ using a Poisson factor model as

$$a_{k_1k_2} \sim \mathrm{Pois}(\sum\nolimits_{c=1}^{C} m_{k_1c}r_c m_{k_2c}),$$

where r_c is the weight of c-th latent factor, and m_{kc} captures how strongly k-th component associates with c-th factor. Naturally, k_1-th component interact with k_2-th component through the weight $\sum_{c=1}^{C} m_{k_1c}r_c m_{k_2c}$. To ensure the latent factor to be nonnegative, we draw the factor r_c, and factor loading m_{kc} from the priors specified by

$$m_{kc} \sim \mathrm{Gam}(\hat{a}_k, \hat{b}_k), \ \ r_c \sim \mathrm{Gam}(\frac{r_0}{C}, c_0),$$

respectively. Here, C is the maximum number of latent factors. As $C \to \infty$, the weights of the latent factors $\{r_c\}_c^C$

and the factor loading $\{\mathbf{m}_c\}_c^C$ can be considered as a draw $G = \sum_{c=1}^{\infty} r_c \delta_{\mathbf{m}_c}$ from a gamma process $\mathrm{GaP}(G_0, c_0)$, where G_0 denotes the base measure over the metric space Ω, and c_0 the concentration parameter [Ferguson, 1973].

3.3 Graph-Structured Transition Dynamics

For high-dimensional count sequences, the underlying transition dynamics are often *sparse* and exhibit a certain amount of graph structure. Hence, we further study to learn the latent graph-structured transition kernel behind count time series, using relational gamma process prior. In particular, we sample the transition parameter $\boldsymbol{\pi}_k$ from a hierarchical Dirichlet prior, as $\boldsymbol{\pi}_k \sim \mathrm{Dir}(a_{1k}, \ldots, a_{Kk})$. To introduce a sparse graph-structured transition kernel, we model the matrix of the hyper-parameter $\mathbf{A} = [a_{k_1 k_2}]_{k_1,k_2}^K$ as $\mathbf{A} = \mathbf{D} \odot \mathbf{Z}$, where $\mathbf{D} = [d_{k_1 k_2}]_{k_1,k_2}^K$ denotes the matrix of the nonnegative hyper-parameters, and $\mathbf{Z} = [z_{k_1 k_2}]_{k_1,k_2}^K$ is a binary mask. More specifically, we consider the dimensions of the transition kernel as vertices, and the non-zero transition behaviours as graph edges. Naturally, we can capture the sparse structure of the transition kernel $\mathbf{\Pi}$ using a graph. As shown in Fig. 3, for each pair of two vertices i and j, $z_{ij} = 1$ means that the transition probability from i-th component to j-th component is non-zero, and vice versa. In particular, we model the binary mask $\mathbf{Z}$ using relational gamma processes as

$$z_{k_1 k_2} \sim \mathrm{Ber}[1 - \exp(\sum\nolimits_{c=1}^{C} m_{k_1 c} r_c m_{k_2 c})],$$

where r_c can be considered as the weight of latent community c, and m_{kc} measures how strongly k-th vertex (the dimension of the transition kernel) relate to c-th latent community, as illustrated in Fig. 3. Note that the binary mask $\mathbf{Z}$ can be equivalently drawn via the Bernoulli-Poisson link function as

$$z_{k_1 k_2} \sim \delta(w_{k_1 k_2} \geq 1), \;\; w_{k_1 k_2} \sim \mathrm{Pois}(\sum\nolimits_{c=1}^{C} m_{k_1 c} r_c m_{k_2 c}).$$

To ensure the model explainability, we restrict r_c and m_{kc} to be nonnegative, and thus place gamma priors over these two parameters as $m_{kc} \sim \mathrm{Gam}(\hat{a}_k, \hat{b}_k)$, $r_c \sim \mathrm{Gam}(\frac{r_0}{C}, c_0)$, respectively. As we discussed in Sec. 3.2, this hierarchical gamma prior can be considered as a draw $G = \sum_{c=1}^{\infty} r_c \delta_{\mathbf{m}_c}$ from a gamma process $\mathrm{GaP}(G_0, c_0)$. In particular, we call this Bayesian non-parametric prior the relational gamma process, as a graph-structure can be naturally induced. The nonnegative hyper-parameters $\mathbf{D} = [d_{k_1 k_2}]_{k_1,k_2}^K$ are drawn from a gamma distribution as $d_{k_1 k_2} \sim \mathrm{Gam}(\epsilon_0, \epsilon_0)$.

The proposed gamma dynamical systems are not fully conjugate. Nonetheless, tractable-yet-efficient Gibbs sampling algorithms are developed to perform posterior simulation via negative-binomial data augmentation strategies [Zhou, 2016a]. The full derivation of the inference procedure is presented in the supplementary material.

4 Related Work

Modeling sequentially observed count sequences has been receiving growing interests in recent years. Here we discuss several types of methods closely related to our studies. [Acharya *et al.*, 2015] first studies the gamma Markov process on sequentially count sequences, in which the latent states evolve independently over time. [Schein *et al.*, 2016a] tries to capture the excitations among the latent gamma Markov processes using a transition structure. [Schein *et al.*, 2019] investigates a Poisson-randomized gamma Markov process which can capture a certain amount of bursty dynamics, and thus demonstrates advantages over gamma Markov processes. [Virtanen and Girolami, 2020] studies a second type of gamma Markov chain structure via the scale parameter of the latent gamma states, which demonstrate better stationary property over the gamma Markov chain proposed by [Acharya *et al.*, 2015]. [Filstroff *et al.*, 2021] recently provides a thorough survey on the studies of the developed gamma Markov processes, and evaluates these models through standard tasks including data smoothing and forecasting. [Han *et al.*, 2014] first tries to capture sequential count observations using linear dynamical systems, via the extend rank likelihood function. [Linderman *et al.*, 2017] proposes to learn switching behaviors of sequential data using recurrent linear dynamical systems (rLDS). [Nassar *et al.*, 2018] further develops a tree-structured extension of rLDS, with multi scale resolution. [Chen *et al.*, 2020] extends the Poisson gamma dynamical systems to learn non-stationary transition dynamics behind count time series. Some efforts are also dedicated to developing Bayesian deep models to capture count sequences. [Gan *et al.*, 2015] develops a temporal sigmoid belief network for count time series.

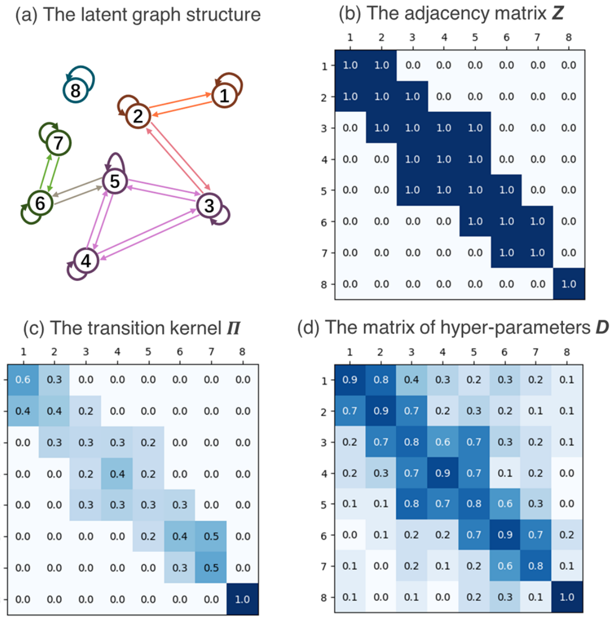


Figure 3: The graph structure of the latent dimensions of the transition kernel behind sequential count observations.

5 Experiments

We evaluate the proposed relational gamma process dynamical systems, and compare it with closely-related methods, using both synthetic and real-world count data.

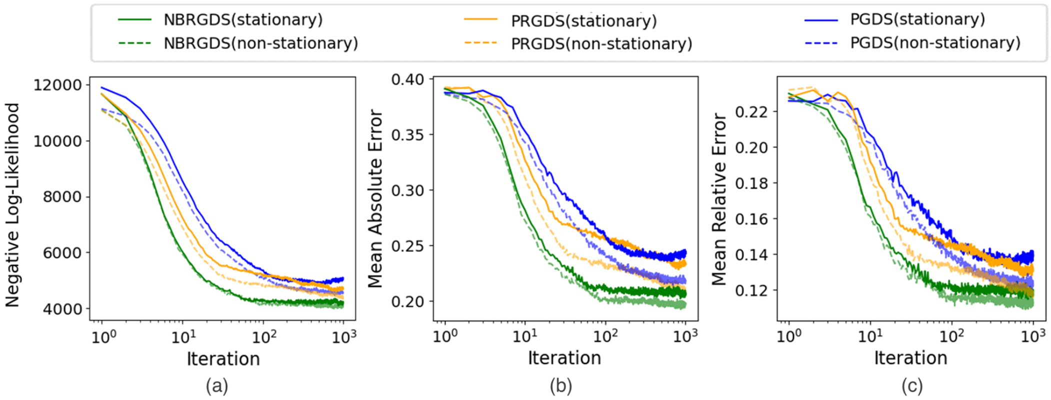


Figure 4: Negative-binomial-randomized gamma dynamical systems (NBRGDSs) demonstrate strong ability in capturing *heterogeneous* overdispersion effects, and thus achieves faster convergence (a), lowest mean absolute error (b) and mean relative error (c), compared with the other related baselines. The stationary and non-stationary generative process i.e. $\delta^{(t)} = \delta$ and $\delta^{(t)}$, denoted as solid line and dotted line, respectively.

Real-world data. We conducted the experiments with the following real-world datasets: **(1) Integrated Crisis Early Warning System (ICEWS)** dataset contains the count number of $6,000$ pairwise interactions between 233 countries over 365 days. By screening out $4,800$ dimensions where the sample sparsity exceeds 99%, we used a subset of ICEW data which contains $V = 1,200$ dimensions, and $T = 365$ time steps; **(2) Last.fm** contains the listening information of $7,071$ music artists over 51 months, where we have $T = 51$ time steps, and $V = 7,071$ dimensions; **(3) Earthquake Reports Database (EQDB):** records more than $120,000$ earthquake reports over $15,000$ earthquakes whose epicenters in the United States and nearby U.S. territories from 1928 to 1985. We created a count matrix where each column represents a month and each row represents a state. The EQDB used in the experiments, contains $T = 696$ time steps, and $V = 64$ dimensions. **(4) COVID-19** contains the daily death toll in the fifty states and Washington DC of the United States, from March 2020 to March 2021. We have $T = 365$ time steps, and $V = 51$ dimensions.

Baselines. In the experiments we compared the predictive of the proposed models with (1) the gamma process dynamic Poisson factor analysis (GaP-DPFA) [Acharya *et al.*, 2015], in which the gamma Markov chain evolves independently over time; (2) the Poisson-gamma dynamical system (PGDS) [Schein *et al.*, 2016a], in which a transition kernel is used to capture the excitations among latent gamma Markov chains; (3) the Poisson-randomized gamma dynamical system (PRGDS) [Schein *et al.*, 2019] where the Poisson-randomized gamma Markov chain structure can capture a certain amount of bursty dynamics.

We denote the proposed negative-binomial-randomized gamma dynamical system as NBRGDS. The proposed NBRGDS with factor-structured prior imposed over the transition kernel, is denoted as FS-NBRGDS. The proposed NBRGDS with graph-structured prior placed over the transition structure, is denoted by GS-NBRGDS.

To evaluate the performance of the compared models in capturing heterogeneous overdispersed behaviours of latent dynamic processes behind count sequences, we considered a subset of ICEWS data that consists of heterogeneous overdispersed counts. More specifically, we sorted the oberserved dimensions according to their variance/expectation ratio, and selected the first L dimensions in descending order, i.e., those dimensions with larger variation/expectation ratio. We present the result for the setting $L = 300$. We treated the 80 percent of the data as the training set, and the remaining 20 percent as the test set. Then, we trained all the compared models with the training set, and evaluated the model performance using the test set. In the experiments, we used mean absolute error (MAE) and mean relative error (MRE) to evaluate the model performance in fitting count sequences:

$$\mathrm{MAE} = \frac{1}{VT}\sum\nolimits_{v=1}^{V}\sum\nolimits_{t=1}^{T}|n_v^{(t)} - \hat{n}_v^{(t)}|,$$

$$\mathrm{MRE} = \frac{1}{VT}\sum\nolimits_{v=1}^{V}\sum\nolimits_{t=1}^{T}\frac{|n_v^{(t)} - \hat{n}_v^{(t)}|}{1 + n_v^{(t)}},$$

where $n_v^{(t)}$ and $\hat{n}_v^{(t)}$ denotes the ground true value and estimated value of dimension v at time t, respectively. They differ because MRE considers the relative magnitude of the errors in relation to the actual values, taking into account the scale of the data, while MAE simply measures the absolute magnitude of the errors without considering the data scale. Fig. 4 shows the results of the compared models averaged over ten random training-testing repeats.

As shown in Fig. 4 (a), NBRGDS has started to converge to its steady states after almost 10^2 iterations, while both PGDS and PRGDS start to converge until 10^3 iterations. Fig. 4 (b) and (c) compares the mean absolute errors and mean relative errors of the compared models, respectively. Overall, NBRGDS achieves the lowest MAE and MRE. PRGDS performs better than PGDS as PRGDS can capture a certain amount of overdispersion effects via its Poisson-randomized chain structure. We also note that NBRGDS with a time-varing scaling factor $\delta^{(t)}$, performs better than stationary NBRGDS because this scaling factor $\delta^{(t)}$ can also capture bursty dynamics. Nonetheless, stationary NBRGDS still out-

			GaP-DPFA	PGDS	PRGDS	NBRGDS	FS-NBRGDS	GS-NBRGPDS
ICEWS	MAE	S	1.29±0.01	1.04±0.02	1.06±0.01	**1.04±0.00**	1.04±0.01	1.04±0.01
		F	0.94±0.01	0.95±0.03	0.98±0.04	1.12±0.02	0.91±0.02	**0.90±0.01**
	MRE	S	0.61±0.00	**0.41±0.01**	0.42±0.00	0.43±0.01	0.42±0.00	**0.41±0.02**
		F	0.53±0.01	0.51±0.04	0.59±0.03	0.58±0.02	0.54±0.01	**0.51±0.01**
Last.fm	MAE	S	1.71±0.03	1.38±0.01	1.39±0.01	1.37±0.01	**1.36±0.02**	1.38±0.02
		F	8.04±0.07	1.41±0.02	5.47±0.22	1.41±0.02	**1.12±0.01**	1.13±0.02
	MRE	S	0.52±0.01	0.34±0.01	0.34±0.00	0.34±0.01	**0.33±0.00**	0.34±0.00
		F	4.02±0.04	0.86±0.02	2.59±0.09	0.85±0.03	0.53±0.04	**0.53±0.01**
EQDB	MAE	S	3.37±0.09	3.37±0.20	3.41±0.33	**3.26±0.12**	3.26±0.20	3.34±0.27
		F	10.17±2.21	10.89±0.94	7.08±0.90	5.65±0.63	3.55±0.04	**3.33±0.06**
	MRE	S	0.89±0.17	0.89±0.05	0.84±0.08	0.83±0.09	0.83±0.07	**0.82±0.05**
		F	8.12±2.14	6.54±0.93	2.45±0.46	1.49±0.11	1.52±0.15	**1.40±0.10**
COVID-19	MAE	S	12.09±0.26	11.42±0.62	11.08±0.25	**10.99±0.43**	11.57±0.07	11.37±0.62
		F	23.35±1.07	27.67±0.18	21.95±0.59	**20.87±0.29**	23.55±0.08	23.30±0.05
	MRE	S	1.47±0.19	1.54±0.11	1.23±0.11	**1.19±0.13**	1.41±0.22	1.32±0.11
		F	6.30±0.58	7.87±0.32	6.32±0.32	1.94±0.15	**1.36±0.03**	1.40±0.03

Table 1: Results for the data smoothing ("S") and future data forecasting ("F") tasks. For both mean absolute error (MAE) and mean relative error (MRE), lower values are better.

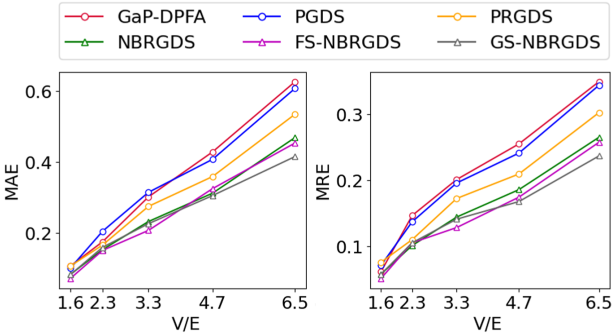


Figure 5: The proposed NBRGDS consistently achieves lower mean absolute and relative errors, when we vary the overdispersed magnitude (the ratio of variance to expection) of the synthetic count sequences, compared with the other closely-related models.

performs both the PRGDS and PGDS with time-varing $\delta^{(t)}$. We conjecture that this improved prediction accuracy is because the time-varying scaling factor $\delta^{(t)}$ fail to capture the underlying overdispersed behaviours, although it still can model a certain amount of bursty dynamics in observed dimensions. This observation further demonstrates the strong ability of the NBRGDS in capturing hetergeneous overdispersion effects of the latent dimensions behind count sequences.

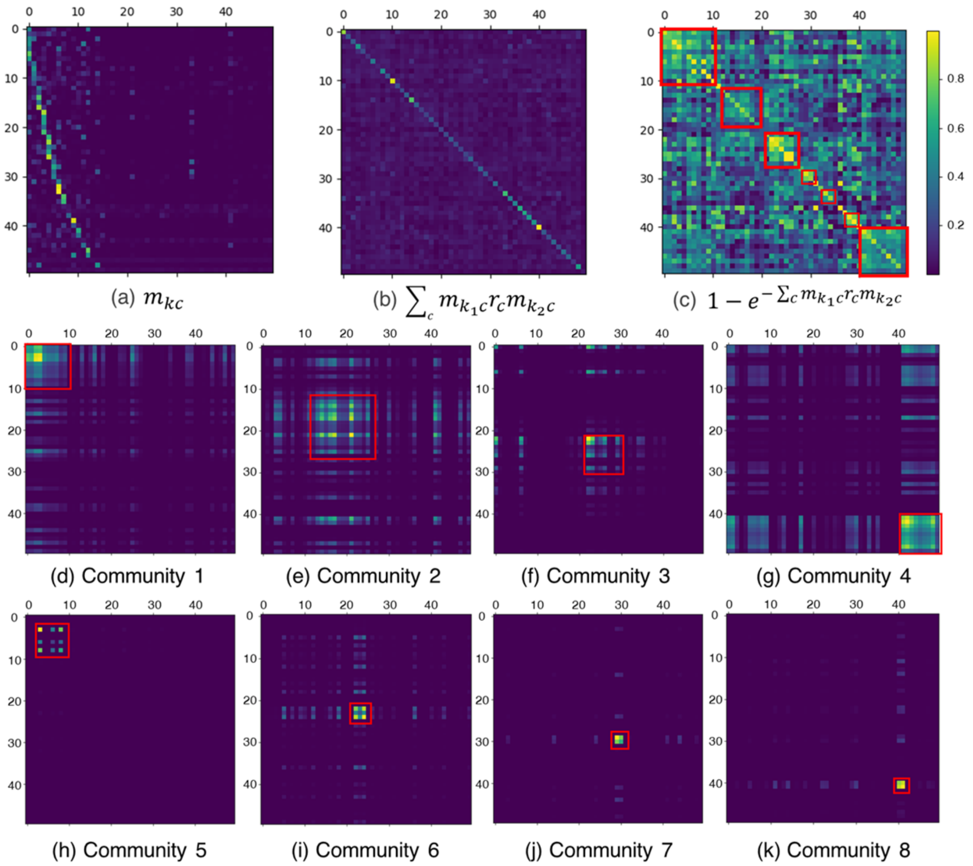
Synthetic data. To further evaluate the performance of the compared models in capturing overdispersion effects, we also considered generating synthetic data with heterogeneous overdispersed dynamics. To that end, we considered to simulate synthetic data using zero-inflated negative-binomial (ZINB) models given by

$$f_{\text{ZINB}}(n \mid p_0, r, p) = p_0 I_0(n) + (1 - p_0) f_{\text{NB}}(n \mid r, p),$$

where f_{ZINB} and f_{NB} represents the probability mass function (PMF) of the zero-inflated negative-binomial distribution and negative-binomial distribution, respectively. Here, $I_0(n)$ is an indicator function that takes 1 when $n = 0$, otherwise 0. The parameter $p_0 \in [0, 1]$ controls the ratio of zero counts, while r and p are the two parameters of the negative-binomial distribution. Hence, we can effectively control the sparsity and overdispersion magnitude of the dimensions by tuning the values of p_0 and p, respectively[1]. More specifically, we generated five groups of synthetic data, in which each group contains $V = 10$ dimensions and $T = 365$ time steps, using the following configurations: **(1)** $p_0 = 0.9$, $r = 5$, $p = 0.9$, $\mathsf{V/E} = 1.6$; **(2)** $p_0 = 0.9$, $r = 5$, $p = 0.8$, $\mathsf{V/E} = 2.3$; **(3)** $p_0 = 0.9$, $r = 5$, $p = 0.7$, $\mathsf{V/E} = 3.3$; **(4)** $p_0 = 0.9$, $r = 5$, $p = 0.6$, $\mathsf{V/E} = 4.7$; **(5)** $p_0 = 0.9$, $r = 5$, $p = 0.5$, $\mathsf{V/E} = 6.5$, where V and E represents variance and expectation of each group data, respectively. Then $\mathsf{V/E}$ denotes the ratio of variance to expectation, and thus measures overdispersion effects. Fig. 5 plots the model performance of the compared methods by varying the ratio of variance to expectation. NBRGDS models including its factor-structured and graph-structured versions, consistently outperforms the other methods. In particular, NBRGDS achieves a significant improvement compared with PRGDS, although PRGDS still can capture a certain amount of overdispersion effects. Additional experiments on synthetic data under different configurations are available in the supplementary material.

Data Smoothing and Forecasting. To quantatively evaluate the predictive performance of the compared methods, we considered two standard tasks: data smoothing and future data forecasting. The data smoothing is to predict $y_v^{(t)}$ given the remaining observations $y_v^{(1:T)} \setminus y_v^{(t)}$, while the future data prediction is to predict future observations at next S time steps $y_v^{(T+1):(T+S)}$ given the history up to T, $y_v^{(1:T)}$. Here we considered to predict next two time steps ($S = 2$). We used the default settings of GaP-DPFA, and PRGDS as

[1] Assume an random variable $x \sim \text{ZINB}(p_0, r, p)$. The expectation and variance of x are $\mathsf{E}[x] = r(1 - p_0)(1 - p)/p$, and $\mathsf{Var}[x] = (1 - p_0)r(1 - p)/p^2 + p_0(1 - p_0)r^2(1 - p)^2/p^2$, respectively. Thus, the ratio of variance to expectation of x is $(1 + rp_0(1 - p))/p$.



(a) m_{kc} (b) $\sum_c m_{k_1c}r_cm_{k_2c}$ (c) $1-e^{-\sum_c m_{k_1c}r_cm_{k_2c}}$

(d) Community 1 (e) Community 2 (f) Community 3 (g) Community 4

(h) Community 5 (i) Community 6 (j) Community 7 (k) Community 8

Figure 6: The latent graph structure inferred by the proposed method on **ICEWS** data

provided in the corresponding releases [Acharya *et al.*, 2015; Schein *et al.*, 2019]. For NBRGDS, FS-NBRGDS and GS-NBRGDS, we choose $K = 100$ when $V \geq 1000$, while the dimensions of EQDB and COVID-19 datasets are smaller than 100, thus we choos $K = 25$. We set $C = K$ for FS-NBRGDS and GS-NBRGDS. We ran 5000 iterations of the Gibbs sampler, which have started to converge after 1000 iterations. We discarded the first 3000 samples which were treated as burn-in time and collected a posterior sample every tenth sample thereafter. Tab. 1 shows the results of the compared methods in these two tasks. Overall, NBRGDS outperforms the GaP-DPFA, PGDS and PRGDS on almost all the datasets. In particular, we found that FS-NBRGDS and GS-NBRGDS show superior performance in future data forecasting. We conjecture this improved prediction accuracy is due to that FS-NBRGDS and GS-NBRGDS can effectively leverage the structure information underlying dynamic count data, and thus yields better predictive accuracy. We provide more comparative results on data smoothing and forecasting over different models in appendix Sec.D.

Graph-Structured Transition Dynamics. Fig. 6 shows the latent graph structure underlying the transition kernel, estimated by the proposed model. Although the model was initialized with $C = 50$ latent communities, the latent graph only consists of approximately ten communities with non-zero weights, as shown in Fig. 6(a). Fig. 6(d-k) plots the eight evident latent communities in which the vertices are densely connected as the corresponding dimensions are more likely to interact with each other. Fig. 6(c) demonstrates that the most estimated latent dynamic processes are almost independent to the other dynamic processes, but only interact with a few other dimensions. We provide the latent graphs inferred for the other real-world data in the supplement.

6 Conclusion

Novel negative-binomial-randomized gamma dynamical systems, have been proposed to capture the *heterogeneous* overdispersed behaviors of latent dynamics behind count time sequences. The new framework demonstrates more explainable latent structure, by learning the factor structure and sparse graph structure of the transition kernels, compared with transition kernel by non-informative priors. Although the prior specification of the proposed framework lacks conjugacy, tractable-yet-efficient sampling algorithms are developed to perform posterior inference. In the future, we plan to capture time-varying graph-structured transition dynamics, which will enable to better understand non-stationary count sequences. We are also considering to enhance the modeling capacities of gamma belief networks [Zhou *et al.*, 2016; Zhou, 2018] and convex polytopes [Zhou, 2016b] using the negative-binomial-randomized gamma Markov processes.

Acknowledgments

This work was partly supported by the Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems.

References

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

[Acharya *et al.*, 2018] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. A dual markov chain topic model for dynamic environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1099–1108, 2018.

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

[Chandra *et al.*, 2023] Noirrit Kiran Chandra, Antonio Canale, and David B. Dunson. Escaping the curse of dimensionality in bayesian model-based clustering. *Journal of Machine Learning Research*, 24(144):1–42, 2023.

[Chen *et al.*, 2020] Wenchao Chen, Bo Chen, Yicheng Liu, Qianru Zhao, and Mingyuan Zhou. Switching poisson gamma dynamical systems. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2029–2036. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[Dieng *et al.*, 2019] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *ArXiv*, 2019.

[Ferguson, 1973] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230, 1973.

[Filstroff *et al.*, 2021] Louis Filstroff, Olivier Gouvert, Cedric Fevotte, and Olivier Cappe. A comparative study of gamma markov chains for temporal non-negative matrix factorization. *IEEE Transactions on Signal Processing*, 69:1614 – 1626, 2021.

[Gan *et al.*, 2015] Zhe Gan, Chunyuan Li, Ricardo Henao, David Carlson, and Lawrence Carin. Deep temporal sigmoid belief networks for sequence modeling. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2467–2475, Cambridge, MA, USA, 2015. MIT Press.

[Ghahramani and Roweis, 1998] Zoubin Ghahramani and Sam Roweis. Learning nonlinear dynamical systems using an em algorithm. *Advances in neural information processing systems*, 11:431–437, 1998.

[Han *et al.*, 2014] Shaobo Han, Lin Du, Esther Salazar, and Lawrence Carin. Dynamic rank factor model for text streams. *Advances in Neural Information Processing Systems*, 27:2663–2671, 2014.

[Jones *et al.*, 2023] Andrew Jones, F. William Townes, Didong Li, and Barbara E. Engelhardt. Alignment of spatial genomics data using deep Gaussian processes. *Nature Methods*, 20(9):1379–1387, 2023.

[King, 2001] Gary King. Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization*, 55(2):497–507, 2001.

[Levitin *et al.*, 2019] Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, et al. De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology*, 15(2):e8557, 2019.

[Linderman *et al.*, 2017] Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 914–922, 2017.

[Nassar *et al.*, 2018] Josue Nassar, Scott Linderman, Monica Bugallo, and Il Memming Park. Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. 2018.

[Roy and Dunson, 2020] Arkaprava Roy and David B Dunson. Nonparametric graphical model for counts. *The Journal of Machine Learning Research*, 21(1):9353–9373, 2020.

[Rudolph and Blei, 2018] Maja Rudolph and David Blei. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003—-1011, 2018.

[Schein *et al.*, 2016a] Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson-gamma dynamical systems. *Advances in Neural Information Processing Systems*, 29:5012–5020, 2016.

[Schein *et al.*, 2016b] Aaron Schein, Mingyuan Zhou, David Blei, and Hanna Wallach. Bayesian poisson tucker decomposition for learning the structure of international relations. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2810–2819, 2016.

[Schein *et al.*, 2019] Aaron Schein, Scott Linderman, Mingyuan Zhou, David Blei, and Hanna Wallach. Poisson-randomized gamma dynamical systems. *Advances in Neural Information Processing Systems*, 32:782–793, 2019.

[Sheldon *et al.*, 2013] Daniel Sheldon, Tao Sun, Akshat Kumar, and Tom Dietterich. Approximate inference in collective graphical models. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1004–1012, 2013.

[Stewart, 2014] Brandon M Stewart. Latent factor regressions for the social sciences. Technical report, Harvard University, 2014.

[Stuart and Wolfram, 2020] Andrew M. Stuart and Marie-Therese Wolfram. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 2020.

[Tong *et al.*, 2020] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[Virtanen and Girolami, 2020] Seppo Virtanen and Mark Girolami. Dynamic content based ranking. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2315–2324, 2020.

[Wang *et al.*, 2008] Chong Wang, David M. Blei, and David E. Heckerman. Continuous time dynamic topic models. In *Conference on Uncertainty in Artificial Intelligence*, 2008.

[Zhou and Carin, 2012] Mingyuan Zhou and Lawrence Carin. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems*, volume 25, pages 2546–2554. Curran Associates, Inc., 2012.

[Zhou and Carin, 2015] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):307–320, 2015.

[Zhou *et al.*, 2016] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *Journal of Machine Learning Research*, 17(163):1–44, 2016.

[Zhou, 2015] Mingyuan Zhou. Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 1135–1143, 2015.

[Zhou, 2016a] Mingyuan Zhou. Nonparametric bayesian negative binomial factor analysis. *Bayesian Analysis*, 13(4):1065–1093, 2016.

[Zhou, 2016b] Mingyuan Zhou. Softplus regressions and convex polytopes, 2016.

[Zhou, 2018] Mingyuan Zhou. Parsimonious bayesian deep networks. In *Advances in Neural Information Processing Systems*, 2018.