Poisson-Randomized Gamma Dynamical Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

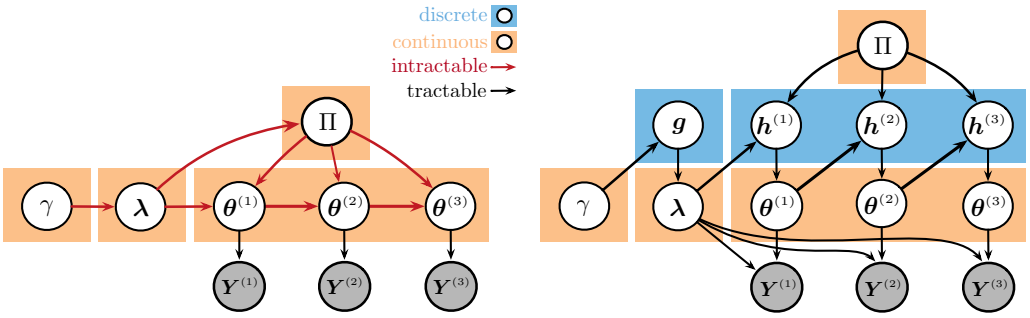
This paper presents the Poisson-randomized gamma dynamical system (PrGDS), a model for sequentially-observed count tensors that encodes a strong inductive bias towards sparsity and burstiness. The PrGDS is based on a new motif in Bayesian latent variable modeling: an alternating series of discrete Poisson and continuous gamma latent states. This motif is widely applicable and analytically tractable, yielding closed-form complete conditionals for all variables by way of the Bessel distribution and a novel distribution that we call the shifted confluent hypergeometric distribution. We draw connections to closely-related models and compare the PrGDS to them in studies of real-world count data of text, international events, and neural spike trains. We find that a sparse variant of the PrGDS—which allows continuous latent states to take values of exactly zero—often obtains the lowest smoothing and forecasting perplexity of all models and is uniquely capable of inferring latent structure that is highly localized in time.

1 Introduction

Political scientists regularly analyze counts of the number of times country i took action a towards country j during time step t [1]. Such data exhibits "complex dependence structures" [2] like coalitions of countries and bursty temporal dynamics. These dependence structures violate the independence assumptions of traditional regression methods that political scientists have traditionally used to test theories of international relations [3, 4, 5]. Political scientists have thus advocated for using latent variable models to infer unobserved structure as a way of controlling for it [6]. The latter approach motivates interpretable yet expressive models, capable of capturing a variety of complex latent structures. Event data sets can be represented as a sequence of count tensors $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ each of which contains the $V \times V \times A$ event counts for that time step for every combination of V sender countries, V receivers, and A action types. Recent work applies tensor decomposition methods to event data sets [7, 8, 9, 10, 11], which infers interpretable coalition structure among countries and topic structure among actions. Like most tensor decomposition methods though, these methods assume that the sequence of tensors is exchangeable and cannot capture the temporal structure in the data.

Sequentially observed count tensors present unique statistical challenges because they tend to be *bursty* [12], *high-dimensional*, and *sparse* [13, 14]. There are few models that are tailored to both the challenging properties count time-series and count tensors. In recent years, Poisson factorization has emerged as a framework for modeling sparse count matrices [15, 16, 17, 18, 19, 20] and tensors [13, 21, 9]. While tensor decomposition methods generally scale with the size the tensor, many Poisson factorization models yield inference algorithms that scale linearly with only the non-zeros. This property allows researchers to efficiently explore latent structure in massive tensors, provided they are sparse. However, this property is unique to a subset of Poisson factorization models that only use non-negative prior distributions, which are difficult to chain in state-space models for time series. Hierarchical compositions of non-negative priors—notably, the gamma and Dirichlet—typically introduce non-conjugate dependencies that require innovative posterior schemes.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.



(a) Poisson–gamma dynamical systems [22] (b) Poisson-randomized gamma dynamical systems

Figure 1: *Left*: The PGDS imposes dependencies directly between the continuous variables that do not yield closed-form conditional distributions. *Right*: The PrGDS (this paper) breaks the intractable dependencies with discrete Poisson variables—doing so yields closed-form conditionals for all variables without any data augmentation.

This paper seeks to fill a gap in the literature between Poisson factorization models that are *tractable*—i.e., yielding closed-form complete conditionals that make approximate inference easy to derive—and those that are *expressive*—i.e., capable of capturing a variety of complex dependence structures. To do so, we introduce alternating chains of discrete Poisson and continuous gamma latent states, a new modeling motif that is analytically convenient and computationally tractable. We rely on this motif to construct the Poisson-randomized gamma dynamical system (PrGDS), a model for sequentially observed count tensors that is tractable, expressive, and efficient. The PrGDS is closely related to the Poisson–gamma dynamical system (PGDS) [22], a recently introduced model for dynamic count matrices, that is based on non-conjugate chains of gamma-distributed states. These chains are intractable—thus, posterior inference in the PGDS relies on sophisticated data augmentation schemes that are cumbersome to derive and impose unnatural restrictions on the priors over other variables. The PrGDS instead introduces intermediate Poisson states that break the intractable dependency between the gamma states (see Fig. 1). While this construction is only *semi*-conjugate, it is tractable, yielding closed-form complete conditionals for the Poisson states by way of the little-known Bessel distribution [23] and a novel discrete distribution that we derive and call the *shifted confluent hypergeometric (SCH) distribution*.

We study the inductive bias of the PrGDS by comparing its smoothing and forecasting ability to the PGDS and two other baselines on a range of real-world count matrices and tensors of text, international events, and neural spike data. We find that the PrGDS often obtains lower smoothing and forecasting perplexity than the PrGDS and related baselines. The PrGDS under a specific hyperparameter settings permits the continuous states to take values of *exactly* zero thus encoding a unique inductive bias tailored to sparsity and burstiness. We find that this variant, in particular, often obtains the lowest perplexity of all models. We also find that the sparse PrGDS is representing of inferring a qualitatively broader range of latent structure—specifically, bursty latent structure that is highly localized in time.

2 Poisson-randomized gamma dynamical systems (PrGDS)

Notation. Consider a data set of of sequentially observed tensors $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$. An entry $y_{\mathbf{i}}^{(t)} \in \{0, 1, 2, \ldots\}$ in the t^{th} tensor is subscripted by a multi-index $\mathbf{i} \equiv (\mathrm{i}_1, \ldots, \mathrm{i}_M)$ which indexes into the M modes of the tensor. As an example, international event counts $y_{i \xrightarrow{a} j}^{(t)}$ collectively form a sequence of 3-mode count tensors where each multi-index corresponds to a unique combination of sender, receiver, and action type—e.g., $\mathbf{i} = (i, j, a)$.

Generative process. The PrGDS is a form of canonical polyadic decomposition [24] that models $y_{\mathbf{i}}$ as

$$y_{\mathbf{i}}^{(t)} \sim \text{Pois}\Big(\rho^{(t)} \sum_{k=1}^{K} \lambda_k \, \theta_k^{(t)} \prod_{m=1}^{M} \phi_{k\mathrm{i}_m}^{(m)}\Big). \tag{1}$$

Here $\theta_k^{(t)}$ represents the activation of the k^{th} *component* at time step t. Each component describes a dependence structure in the data by way of a factor vector $\boldsymbol{\phi}_k^{(m)}$ for each mode m. For international events data, the first factor vector $\boldsymbol{\phi}_k^{(1)} = (\phi_{k1}^{(1)}, \ldots, \phi_{kV}^{(1)})$ would describe the rate at which each of the V countries acts as a sender in the k^{th} component while the second $\boldsymbol{\phi}_k^{(2)}$ would describe the rate

at which each acts as a receiver. The weights λ_k and $\rho^{(t)}$ represent the overall scale of component k and time step t. The PrGDS is called *stationary* if $\rho^{(t)} = \rho$. We posit the following conjugate priors,

$$\rho^{(t)} \sim \text{Gam}\,(a_0, b_0) \quad \text{and} \quad \boldsymbol{\phi}_k^{(m)} \sim \text{Dir}(a_0, \ldots, a_0). \tag{2}$$

The PrGDS is characterized by an alternating series of discrete and continuous latent states. The *continuous latent states* $\theta_k^{(t)}$ evolve via intermediate discrete states $h_k^{(t)}$ from $t = 1, \ldots, T$ as

$$\theta_k^{(t)} \sim \text{Gam}\left(\epsilon_0^{(\theta)} + h_k^{(t)},\, \tau\right) \quad \text{and} \quad h_k^{(t)} \sim \text{Pois}\Big(\tau \sum_{k_2=1}^{K} \pi_{kk_2}\, \theta_{k_2}^{(t-1)}\Big), \tag{3}$$

where for $t = 0$ we define $\theta_k^{(0)} = \lambda_k$ to be the per-component weight that also appears in Eq. (1). The PrGDS assumes $\theta_k^{(t)}$ is conditionally gamma distributed with *rate* τ and shape equal to a latent count $h_k^{(t)}$ plus hyperparameter $\epsilon_0^{(\theta)} \geq 0$. We adopt the convention that a gamma random variable is zero, almost surely, if its shape parameter is zero—thus, setting $\epsilon_0^{(\theta)} = 0$ defines the *sparse PrGDS* wherein the continuous states are exactly zero $\theta_k^{(t)} = 0$ if $h_k^{(t)} = 0$. The *transition weights* π_{kk_2} in the Poisson rate of $h_k^{(t)}$ represent how strongly each component k_2 excites component k at the subsequent time step. We view these weights collectively as a $K \times K$ transition matrix Π and impose Dirichlet priors over its columns. We also place a gamma prior over the *concentration parameter* τ which is conjugate to both the gamma and Poisson distributions it appears in:

$$\tau \sim \text{Gam}\,(\alpha_0, \alpha_0) \quad \text{and} \quad \boldsymbol{\pi}_k \sim \text{Dir}\,(a_0, \ldots, a_0) \;\text{ such that }\; \textstyle\sum_{k_1}^{K} \pi_{k_1 k} = 1. \tag{4}$$

For the per-component weights λ_k, we place a hierarchical prior with a similar flavor to Eq. (3):

$$\lambda_k \sim \text{Gam}\Big(\tfrac{\epsilon_0^{(\lambda)}}{K} + g_k,\, \beta\Big) \quad \text{and} \quad g_k \sim \text{Pois}\left(\tfrac{\gamma}{K}\right), \tag{5}$$

where $\epsilon_0^{(\lambda)}$ is a hyperparameter analogous to $\epsilon_0^{(\theta)}$. The following gamma priors are then both conjugate:

$$\gamma \sim \text{Gam}\,(a_0, b_0) \quad \text{and} \quad \beta \sim \text{Gam}\,(\alpha_0, \alpha_0)\,. \tag{6}$$

Properties. Both $\epsilon_0^{(\lambda)}$ and γ are divided by the number of components K in Eq. (5)—as the number of components grows $K \to \infty$, the expected sum of the weights thus remains finite and fixed:

$$\sum_{k=1}^{\infty} \mathbb{E}\left[\lambda_k\right] = \sum_{k=1}^{\infty} \left(\tfrac{\epsilon_0^{(\lambda)}}{K} + \mathbb{E}\left[g_k\right]\right)\beta^{-1} = \sum_{k=1}^{\infty} \left(\tfrac{\epsilon_0^{(\lambda)}}{K} + \tfrac{\gamma}{K}\right)\beta^{-1} = \left(\epsilon_0^{(\lambda)} + \gamma\right)\beta^{-1}. \tag{7}$$

Thus, this prior encodes an inductive bias towards small values of λ_k and may be interpreted as the finite truncation of a novel Bayesian nonparametric process. A small value of λ_k shrinks the Poisson rates of both the data $y_{\mathbf{i}}^{(t)}$ and the first discrete latent state $h_k^{(0)}$—this prior encourages the model to only infer components that are both predictive of the data and useful for fitting the latent dynamics.

The marginal expectation of the state vector $\boldsymbol{\theta}^{(t)}$ takes the canonical form of linear dynamical systems,

$$\mathbb{E}\left[\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{\theta}^{(t-1)}\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{h}^{(t-1)}\right]\right] = \epsilon_0^{(\theta)} \tau^{-1} + \Pi\, \boldsymbol{\theta}^{(t-1)}, \tag{8}$$

since by iterated expectation $\mathbb{E}\left[\theta_k^{(t)}\right] = \left(\epsilon_0^{(\theta)} + \mathbb{E}\left[h_k^{(t)}\right]\right)\tau^{-1} = \left(\epsilon_0^{(\theta)} + \tau \sum_{k_2=1}^{K} \pi_{kk_2}\, \theta_{k_2}^{(t-1)}\right)\tau^{-1}$. The *concentration parameter* τ appears in both the Poisson and gamma distributions in Eq. (3). It contributes to the variance of the process while canceling out of the expectation, except for the additive term $\epsilon_0^{(\theta)} \tau^{-1}$ which vanishes when $\epsilon_0^{(\theta)} = 0$.

More generally, we can marginalize out all of the discrete latent states $h_k^{(t)}$ to obtain a purely continuous dynamical system in terms of the *randomized gamma distribution of the first type* (RG1) [23, 25],

$$\theta_k^{(t)} \sim \text{RG1}\Big(\epsilon_0^{(\theta)},\, \tau \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)},\, \tau\Big), \tag{9}$$

when $\epsilon_0^{(\theta)} > 0$ and in terms of a limiting form of the RG1 when $\epsilon_0^{(\theta)} = 0$. We describe the RG1 in Fig. 2.

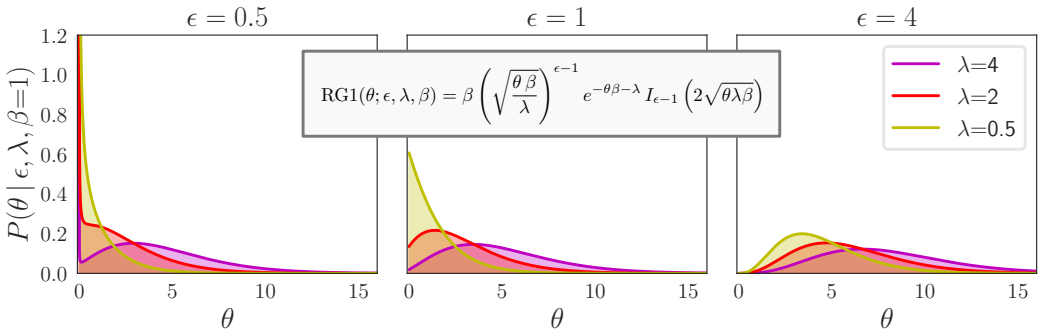


Figure 2: The randomized gamma distribution of the first type (RG1) [23, 25] has support $\theta > 0$ and is defined by three parameters $\epsilon, \lambda, \beta > 0$. Its PDF is given in the plot where $I_\nu(a)$ is the modified Bessel function of the first kind [26]. When $\epsilon < 1$ (*Left*) the RG1 resembles a soft "spike-and-slab" while when $\epsilon \geq 1$ (*Middle and Right*) it resembles a more-dispersed form of the gamma distribution. A limiting case of the RG1 when $\epsilon \to 0$ is the Poisson-randomized gamma distribution [27] which includes zeros in its support $\theta \geq 0$.

3 Related work

The PrGDS closely relates to the Poisson–gamma dynamical system (PGDS) [22] wherein

$$\theta_k^{(t)} \sim \mathrm{Gam}\Big(\tau \sum_{k_2=1}^{K} \pi_{kk_2}\theta_{k_2}^{(t-1)},\, \tau\Big) \;\text{ such that }\; \mathbb{E}\left[\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{\theta}^{(t-1)}\right] = \Pi\,\boldsymbol{\theta}^{(t-1)}. \tag{10}$$

The PGDS posits non-conjugate dependence between gamma-distributed states. The complete conditional $P(\theta_k^{(t)}|-)$ is not available in closed form under the PDGS and posterior inference relies on a sophisticated data augmentation scheme. The PrGDS instead introduces intermediate Poisson states $h_k^{(t)}$ that break the intractable dependencies between the gamma states. The Poisson is not a conjugate prior to the gamma—however, this construction is still tractable, and yields a closed-form conditional $P(h_k^{(t)}|-)$, as we'll show in § 4. The PGDS is limited by the data augmentation scheme it requires for posterior inference—specifically, this augmentation scheme does not allow any per-component weights λ_k to appear in the Poisson rate of $y_{\mathbf{i}}^{(t)}$ in Eq. (1). To encourage parsimony, the PGDS instead draws weights $\lambda_k \sim \mathrm{Gam}(\frac{\gamma}{K}, \beta)$ and uses them to shrink the transition matrix Π. This introduces more intractable dependencies that necessitate a different augmentation scheme for inference. These augmentation schemes additionally impose restrictions that the factors $\boldsymbol{\phi}_k^{(m)}$ and columns $\boldsymbol{\pi}_k$ of the transition matrix are Dirichlet distributed—while we adopt those same assumptions in this paper, the PrGDS is not bound to them. We provide a graphical comparison of PGDS to PrGDS in Fig. 1.

The PGDS and its "deep" variants [28, 29], generalize gamma process dynamic Poisson factor analysis (GP-DPFA) [30] which assumes a simple random walk $\theta_k^{(t)} \sim \mathrm{Gam}\left(\theta_k^{(t-1)},\, c^{(t)}\right)$; see also a related model [31]. These models belong to a line of work exploring the application of the "augment-and-conquer" data augmentation scheme [32] to perform inference in hierarchies of gamma variables chained via their shape and linked to Poisson observations—beyond time-series models, this construction can be used to build belief networks [33]. An alternative approach is to chain gamma variables via their rate—e.g., $\theta^{(t)} \sim \mathrm{Gam}\,(a,\, \theta^{(t-1)})$. This motif is conjugate and tractable and has been applied to time-series models [34, 35, 36] as well as deep belief networks [37]. However, the rate contributes quadratically to the variance of the gamma distribution and these chains may be highly volatile.

More broadly, gamma shape and rate chains are examples of non-negative chains. Such chains are particularly motivated in the context of Poisson factorization, which is particularly efficient when only non-negative prior distributions are used. In general, Poisson factorization assumes that each observed count is drawn $y_{\mathbf{i}} \sim \mathrm{Pois}\,(\mu_{\mathbf{i}})$ with latent rate $\mu_{\mathbf{i}}$ defined to be some function of model parameters. When the rate is linear—i.e., $\mu_{\mathbf{i}} = \sum_k \mu_{\mathbf{i}k}$—Poisson factorization yields a *latent source representation* [16, 18] wherein we define $y_{\mathbf{i}} \triangleq \sum_k y_{\mathbf{i}k}$ to be the sum of latent sources, each of which is drawn $y_{\mathbf{i}k} \sim \mathrm{Pois}\,(\mu_{\mathbf{i}k})$. Conditioning on the latent sources often induces conditional independences between the latent variables and parameters that facilitates closed-form, efficient, and parallelizable posterior inference—thus, the first step in either MCMC or variational inference is to update the latent sources from their complete conditionals, which is multinomial [38] ,

$$\left((y_{\mathbf{i}1}, \ldots, y_{\mathbf{i}K}) \,|\, -\right) \sim \mathrm{Multinom}\,\left(y_{\mathbf{i}},\, (\mu_{\mathbf{i}1}, \ldots, \mu_{\mathbf{i}K})\right), \tag{11}$$

where we leave implicit the normalization of the non-negative rates $\mu_{\mathbf{i}k}$ into a probability vector. When the observed count is zero $y_{\mathbf{i}}=0$, the sources are zero $y_{\mathbf{i}k}=0$, almost surely, and no computation is required to update them. Thus, any Poisson factorization model that admits a latent source representation scales linearly with only the non-zero counts in the data. This property is indispensable when modeling count tensors which typically contain exponentially more entries than non-zeros [39].

Since the latent source representation is only available when the rate $\mu_{\mathbf{i}}$ is a (multi)linear function of parameters and since the rate must be non-negative, by definition of the Poisson distribution, efficient Poisson factorization is only compatible with non-negative priors over parameters. Modeling time series and other complex dependence structures with efficient Poisson factorization often requires developing novel motifs that notably exclude Gaussian priors which researchers have traditionally relied on for their analytic convenience and tractability. The Poisson LDS, for instance, [40] links the widely-used Gaussian linear dynamical system [41, 42] to Poisson observations via an exponential link function $\mu_{\mathbf{i}} = \exp\left(\sum_k \cdots\right)$. This is one instance of the generalized linear model (GLM) [43] approach that relies on a non-linear link function and thus does not yield a latent source representation. Another approach is to use log-normal priors, as in Dynamic Poisson Factorization [44]—while this approach satisfies the non-negative constraint, the log-normal is not conjugate to the Poisson and does not closed-form conditionals. We also note a long tradition of autoregressive models for count time series including VAR models [45] and those based on Hawkes processes [46, 47, 48]. This approach avoids the challenge of constructing tractable state-space models from non-negative priors by modeling temporal correlation directly between the count observations. For high-dimensional data, such as sequentially-observed tensors, an autoregressive approach is often untenable.

4 Posterior inference

The complete conditionals for all latent variables in the PrGDS are immediately available in closed form without any data augmentation. Iteratively re-sampling each variable from its conditional constitutes a Gibbs sampling algorithm. We provide conditionals for the latent variables with non-standard priors here and relegate the rest to the Appendix. The PrGDS is based on a new modeling motif—we first introduce it in its general form, derive its complete conditionals, and then apply these identities to the PrGDS.

4.1 Poisson–gamma–Poisson recursions

Consider the following model of m involving latent variables θ and h and fixed $c_1, c_2, c_3, \epsilon_0^{(\theta)} > 0$:

$$m \sim \text{Pois}\left(\theta c_3\right), \;\; \theta \sim \text{Gam}\left(\epsilon_0^{(\theta)} + h, c_2\right), \;\; \text{and} \;\; h \sim \text{Pois}\left(c_1\right). \tag{12}$$

This model is *semi*-conjugate. The gamma prior of θ is conjugate to the Poisson and its posterior is

$$\left(\theta \mid -\right) \sim \text{Gam}\left(\epsilon_0^{(\theta)} + h + m, \; c_2 + c_3\right). \tag{13}$$

The Poisson prior over h is not conjugate to the gamma— however the conditional posterior of h is still available in closed form by way of the Bessel distribution [23] which we define in Fig. 3(a):

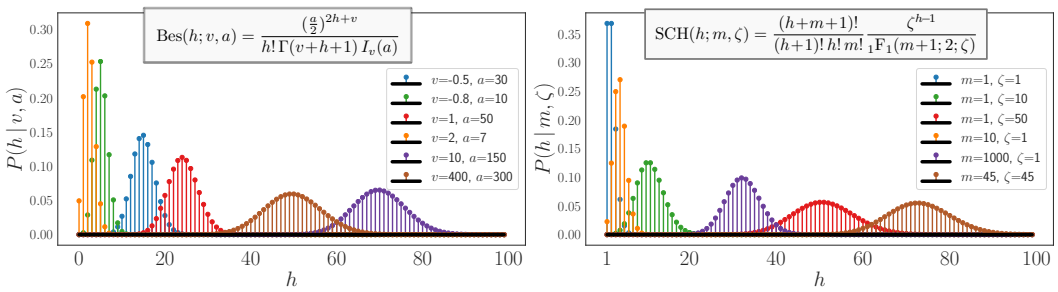
$$\left(h \mid -\right) \sim \text{Bes}\left(\epsilon_0^{(\theta)} - 1, \; 2\sqrt{\theta \, c_2 \, c_1}\right). \tag{14}$$

The Bessel distribution can be sampled efficiently [49]; we will release our Cython implementation. Provided that $\epsilon_0^{(\theta)} > 0$, sampling θ and h iteratively from Eqs. (13) and (14) constitutes a valid Markov chain for posterior inference. When $\epsilon_0^{(\theta)} = 0$ though, $\theta = 0$, almost surely, if $h = 0$, and vice versa—thus, this Markov chain has an absorbing condition at $h = 0$ and violates detailed balance. In this case, we must therefore sample h with θ marginalized out—towards that end, we give Theorem 1.

Theorem 1: *The incomplete conditional* $P(h \mid \epsilon_0^{(\theta)} = 0, -\backslash\theta) \triangleq \int P(h, \theta \mid \epsilon_0^{(\theta)} = 0, -) \, \mathrm{d}\theta$ *is*

$$(h \mid -\backslash\theta) \sim \begin{cases} \text{Pois}\left(\frac{c_1 \, c_2}{c_3 + c_2}\right) & \text{if } m = 0 \\ \text{SCH}\left(m, \; \frac{c_1 \, c_2}{c_3 + c_2}\right) & \text{otherwise} \end{cases} \tag{15}$$

where SCH is a discrete distribution we call the shifted confluent hypergeometric distribution. We describe the SCH in Fig. 3(b) and provide further details about it in the Appendix including the derivation of its PMF, PGF, and mode, along with details of how we sample from it and the proof for Theorem 1.



(a) Bessel distribution. [23] (b) Shifted confluent hypergeometric (SCH) distribution

Figure 3: Two discrete distributions that arise as posteriors in Poisson–gamma–Poisson recursions.

4.2 Closed-form complete conditionals for the PrGDS

The PrGDS yields a latent source representation (see Eq. (11))—posterior inference begins with

$$\left((y_{\mathbf{i}k}^{(t)})_{k=1}^K \mid -\right) \sim \text{Multinom}\left(y_{\mathbf{i}}^{(t)}, \left(\lambda_k \theta_k^{(t)} \textstyle\prod_{m=1}^M \phi_{k\mathbf{i}_m}^{(m)}\right)_{k=1}^K\right). \tag{16}$$

We may similarly represent $h_k^{(t)}$ under its latent source representation as $h_k^{(t)} \equiv h_{k\cdot}^{(t)} = \sum_{k_2=1}^K h_{kk_2}^{(t)}$ where $h_{kk_2}^{(t)} \sim \text{Pois}\left(\tau\, \pi_{kk_2} \theta_{k_2}^{(t-1)}\right)$. When useful, we use dot-notation ("$\cdot$") to denote summing over an axis—in this case $h_{k\cdot}^{(t)}$ denotes the sum of the k^{th} row of the $K \times K$ matrix of latent counts $h_{kk_2}^{(t)}$. The complete conditional of the k^{th} row of counts, when conditioned on their sum $h_{k\cdot}^{(t)}$, is

$$\left((h_{kk_2}^{(t)})_{k_2=1}^K \mid -\right) \sim \text{Multinom}\left(h_{k\cdot}^{(t)}, (\pi_{kk_2} \theta_{k_2}^{(t-1)})_{k_2=1}^K\right). \tag{17}$$

To derive the conditional for $\theta_k^{(t)}$ we aggregate all Poisson variables that depend on it. By Poisson additivity, the column sum $h_{\cdot k}^{(t+1)} = \sum_{k_1=1}^K h_{k_1 k}^{(t+1)}$ is distributed $h_{\cdot k}^{(t+1)} \sim \text{Pois}\left(\theta_k^{(t)}\, \tau\, \pi_{\cdot k}\right)$ and similarly $y_{\cdot k}^{(t)}$ is distributed $y_{\cdot k}^{(t)} \sim \text{Pois}\left(\theta_k^{(t)} \rho^{(t)} \lambda_k \prod_{m=1}^M \phi_{\cdot k}^{(m)}\right)$. The count $m_k^{(t)} \triangleq h_{\cdot k}^{(t+1)} + y_{\cdot k}^{(t)}$ then isolates all dependence on $\theta_k^{(t)}$ and is also Poisson distributed. By gamma–Poisson conjugacy, the conditional of $\theta_k^{(t)}$ is

$$\left(\theta_k^{(t)} \mid -\right) \sim \text{Gam}\left(\epsilon_0^{(\theta)} + h_{k\cdot}^{(t)} + m_k^{(t)},\ \tau + \tau\, \pi_{\cdot k} + \rho^{(t)} \lambda_k \textstyle\prod_{m=1}^M \phi_{\cdot k}^{(m)}\right). \tag{18}$$

When $\epsilon_0^{(\theta)} > 0$, we may apply the identity in Eq. (14) and sample $h_{k\cdot}^{(t)}$ from its complete conditional:

$$\left(h_{k\cdot}^{(t)} \mid -\right) \sim \text{Bessel}\left(\epsilon_0^{(\theta)} - 1,\ 2\sqrt{\theta_k^{(t)}\, \tau^2 \textstyle\sum_{k_2=1}^K \pi_{kk_2} \theta_{k_2}^{(t-1)}}\right). \tag{19}$$

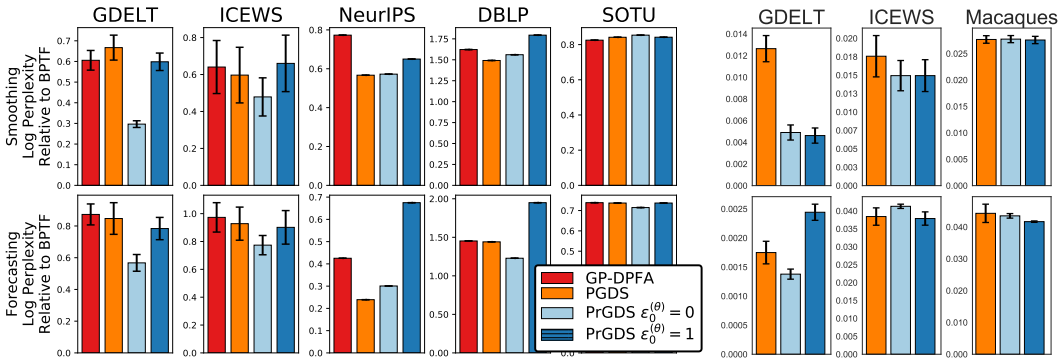
When $\epsilon_0^{(\theta)} = 0$, we instead apply Theorem 1 to sample $h_{k\cdot}^{(t)}$ where $m_k^{(t)}$ is analogous to m in Eq. (15):

$$\left(h_{k\cdot}^{(t)} \mid -\backslash\theta_k^{(t)}\right) \sim \begin{cases} \text{Pois}(\zeta_k^{(t)}) & \text{if } m_k^{(t)} = 0 \\ \text{SCH}(m_k^{(t)},\, \zeta_k^{(t)}) & \text{otherwise} \end{cases} \quad \text{where} \quad \zeta_k^{(t)} \triangleq \frac{\tau^2 \sum_{k_2=1}^K \pi_{kk_2} \theta_{k_2}^{(t-1)}}{\tau + \tau\, \pi_{\cdot k} + \rho^{(t)} \lambda_k \prod_{m=1}^M \phi_{\cdot k}^{(m)}}. \tag{20}$$

The conditionals for λ_k and g_k follow from applying the same Poisson–gamma–Poisson identities while those for γ, β, $\boldsymbol{\phi}_k^{(m)}$, $\boldsymbol{\pi}_k$, and τ all follow from conjugacy. We provide them all in the Appendix.

5 Empirical study

We've seen how the Poisson–gamma–Poisson motif of the PrGDS (see § 4.1) yields a more tractable (see Fig. 1) and more flexible (see § 3) model family than previous work. This motif also encodes a unique inductive bias that we empirically test by comparing the PrGDS to the Poisson–gamma dynamical system (PGDS) [22]. The PGDS is the pure gamma analog to the PrGDS, as we see by comparing Eqs. (9) and (10)—comparing these models thus isolates the impact of the Poisson–gamma–Poisson motif. The PGDS was only previously introduced to model a $T \times V$ *matrix* Y of sequentially observed V-dimensional vectors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)}$. To compare to it, we generalize the PGDS to M-mode tensors. We have provided our Cython implementation of the generalized PGDS (in addition to the PrGDS) and derive its complete conditionals in the Appendix. We also compare the PrGDS variant with $\epsilon_0^{(\theta)} = 1$ to the one with $\epsilon_0^{(\theta)} = 0$, which permits *sparse activations* $\theta_k^{(t)} = 0$.



(a) Matrix empirical studies (originally described by Schein et al. [22]). (b) Tensor empirical studies.

Figure 4: The smoothing (top row) and forecasting (bottom row) performance of each model is measured by log-perplexity—where *lower is better*—divided by the log-perplexity of a non-dynamic baseline, BPTF [9].

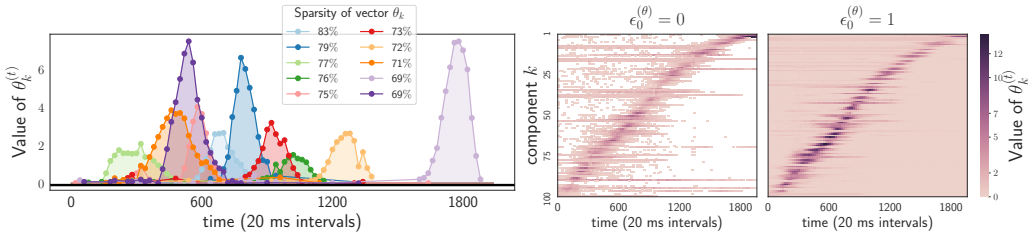
Setup. All empirical studies have the following setup: for some data set $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$, all of the counts $\boldsymbol{Y}^{(t)}$ at randomly selected time steps t are masked. The last two time steps are also always masked. Each model is fit to the masked data using independent chains of MCMC that impute the missing data each at iteration and ultimately return S posterior samples of the latent variables and parameters. The samples are then used to compute the expectations $\mu_{\mathbf{i}}^{(t)}$ (as defined in Eq. (1)) of the heldout counts. We distinguish the task of predicting counts in the final time steps when subsequent observed data is unavailable (*forecasting*) from the task of predicting intermediate time steps (*smoothing*). To assess predictive performance we compute log-perplexity as defined—$\log \text{Perp}(\Delta) = -\frac{1}{|\Delta|} \sum_{(t,\mathbf{i}) \in \Delta} \log \left[\frac{1}{S} \sum_{s=1}^{S} \text{Pois}\big(y_{\mathbf{i}}^{(t)}; \mu_{\mathbf{i},s}^{(t)}\big) \right]$—where Δ is the set of multi-indices of the heldout counts and $\mu_{\mathbf{i},s}^{(t)}$ is the expectation of the heldout count computed from the s^{th} sample. In all studies, we fit a simple non-dynamic baseline—i.e., Bayesian Poisson tensor factorization (BPTF) [9]—that assumes the data at different time slices are i.i.d. $y_{\mathbf{i}}^{(t)} \sim \text{Pois}\,(\mu_{\mathbf{i}})$. This model thus fits $\mu_{\mathbf{i}}$ from training data and predicts it for all heldout time slices. For the dynamic models, we then report their improvement over non-dynamic BPTF by dividing their log-perplexity by BTPF's.

Matrices. We first replicated the empirical studies on $T \times V$ dynamic *matrices* (i.e., 2-mode tensors) reported by Schein et al. (2016) [22]. These studies followed the setup described above and compared the PGDS to GP-DPFA [30], a simple dynamic baseline that we describe in § 3. The matrices in these studies are based on three text corpora—NeurIPS papers [50], DBLP abstracts [51], and State of the Union (SOTU) speeches [52]—where $y_v^{(t)}$ is the number of times word v occurs in time step t, and two international events data sets—GDELT [53] and ICEWS [54]—where $y_v^{(t)}$ is the number of times sender–receiver pair v interacted during time step t. We obtained the matrices and random masks along with the original results for both PGDS and GP-DPFA from the authors and ran the PrGDS with the same MCMC settings they describe (see their paper for details [22]) and BTPF (which uses variational inference) on all matrices and masks. We display the results in Fig. 4(a).

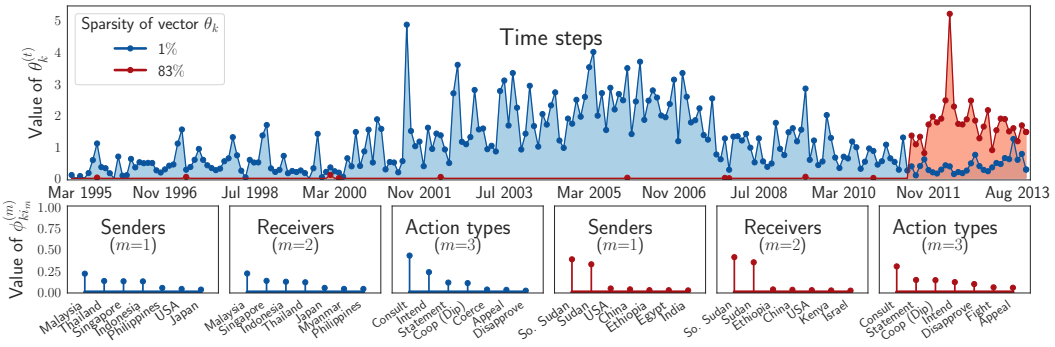
Tensors. We obtained tensor data from two international events data sets—GDELT and ICEWS—wherein a count $y_{i \xrightarrow{a} j}^{(t)}$ is the number of times country i took action a towards country j during time step t. These counts form a sequence of 3-mode tensors of size $V \times V \times A$ for $V = 249$ countries and $A = 20$ actions. For both data sets, we treat months as time steps, where for GDELT we consider the date range 2003–2008 (thus, $T = 72$) and for ICEWS we consider 1995–2013 ($T = 228$).

We also obtained neuroscience data of multielectrode recordings of macaque monkey motor cortices from Williams et al. (2018) [55]. In this data, a count $y_{ij}^{(t)}$ is the number of times neuron i fired in trial j during time step t. These counts form a sequence of matrices each of size $N \times S$ where $N = 100$ is the number of neurons and $S = 1,716$ is the number of trials. We consider each time step to be a 20 millisecond interval which yields $T = 162$. See Fig. 4(b).

For each of the three tensors, we randomly generated two masks that each holdout three pairs of adjacent time steps in the range $t \in [2, T-2]$ as well as the last two time steps $T-1, T$. For each dynamic model, mask, and tensor, we run two independent chains of $4,000$ MCMC iterations, saving every 100^{th} sample after the first 1,000 to compute heldout log-perplexity; we then also fit BPTF and report the relative improvement over it for the dynamic models in Fig. 4(b).



(a) Latent activation structure inferred from macaque cortex data by the PrGDS. *Right:* Comparison of the $K \times T$ state matrix Θ inferred by the two PrGDS variants: $\epsilon_0^{(\theta)} = 0$ (sparse) vs. $\epsilon_0^{(\theta)} = 1$. The components k are sorted by which time step t had the largest $\theta_k^{(t)}$. The banded structure suggests both infer components that activate within specific short durations. White cells correspond to zeros $\theta_k^{(t)} = 0$—the PrGDS can only represent sparse activation structure when $\epsilon_0^{(\theta)} = 0$. *Left:* Alternative visualization of the ten sparsest rows of the matrix $\boldsymbol{\theta}_k$. Each is a component's T-length activation vector—they collectively depict localized and "bursty" neuronal spiking.



(b) Two components inferred by the sparse PrGDS from ICEWS data—the blue component was found by other models, the red component was not. The red component is specific to South Sudan, as can be seen by visualizing largest values of the factor vectors for the sender and receiver modes (*bottom row, second and third rightmost*). South Sudan was not a country until July 2011. The activation vector (*top*) is thus sparse—$\theta_k^{(t)} = 0$ at 94% of time steps (months) prior to July 2011 (83% overall). By contrast, the blue component representss Southeast Asian relations, which are persistently active. The sparse PrGDS can infer both temporally-persistent latent structure that other models infer (blue), as well as temporally-localized structure that other models do not (red).

Figure 5: Sparse representations of macaque cortex activity (Fig. 5(a)) and ICEWS events (Fig. 5(b)).

Discussion. A PrGDS variant obtained the lowest perplexity of all models in ten out of the 14 studies (i.e., subplots in Fig. 4). The sparse PrGDS obtained lower perplexity than the non-sparse PrGDS in nine studies, sometimes dramatically lower. We conjecture that the better performance of the sparse PrGDS can be explained by the expectation of future states $\boldsymbol{\theta}^{(t)}$ given in Eq. (8)—when $\epsilon_0^{(\theta)} > 0$ this expectation includes an additive bias term which grows as we forecast further time steps. When $\epsilon_0^{(\theta)} = 0$ the bias term vanishes and the expectation matches the analogous one for the PGDS, given in Eq. (10).

We also find that the sparse variant is capable of inferring a qualitatively broader range of latent structure that includes bursty and sparse latent structure. In Fig. 5(a) and Fig. 5(b) we explore some of the sparse latent structure inferred by the PrGDS from the macaque cortex data and the ICEWS event data. In the Appendix, we provide examples of how we aligned components across models and provide examples of some well-aligned ones. We found that all models inferred qualitatively similar components—however, a small number of components were unique to the sparse PrGDS. We give one such example in Fig. 5(b).

6 Conclusion

A novel modeling motif—Poisson–gamma–Poisson recursions—allows us to construct the Poisson-randomized gamma dynamical system, a tractable and flexible model family for sequentially observed count tensors. A variant of the PrGDS permits a truly sparse latent representation that is both qualitatively appealing and provides a natural inductive bias for sparse and "bursty" count time series.

References

[1] Philip A Schrodt. Event data in foreign policy analysis. *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, pages 145–166, 1995.

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

[3] Donald P Green, Soo Yeon Kim, and David H Yoon. Dirty pool. *International Organization*, 55(2):441–468, 2001.

[4] Paul Poast. (Mis)using dyadic data to analyze multilateral events. *Political Analysis*, 18(4):403–425, 2010.

[5] Robert S Erikson, Pablo M Pinto, and Kelly T Rader. Dyadic analysis in international relations: A cautionary tale. *Political Analysis*, 22(4):457–463, 2014.

[6] Brandon Stewart. Latent factor regressions for the social sciences. *Harvard University: Department of Government Job Market Paper*, 2014.

[7] Peter D Hoff and Michael D Ward. Modeling dependencies in international relations networks. *Political Analysis*, 12(2):160–175, 2004.

[8] Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

[9] A. Schein, J. Paisley, D. M. Blei, and H. Wallach. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the Twenty-First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054, 2015.

[10] Peter D Hoff et al. Equivariant and scale-free Tucker decomposition models. *Bayesian Analysis*, 11(3):627–648, 2016.

[11] Aaron Schein, Mingyuan Zhou, David M. 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*, 2016.

[12] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[13] E. C. Chi and T. G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

[14] Tsuyoshi Kunihama and David B Dunson. Bayesian modeling of temporal dependence in large sparse contingency tables. *Journal of the American Statistical Association*, 108(504):1324–1338, 2013.

[15] J. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129, 2004.

[16] D. B. Dunson and A. H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25, 2005.

[17] Michalis K. Titsias. The infinite gamma–Poisson feature model. In *Advances in Neural Information Processing Systems 21*, pages 1513–1520, 2008.

[18] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[19] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 2012.

[20] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[21] Beyza Ermis and A Taylan Cemgil. A bayesian tensor factorization model via variational inference for link prediction. *arXiv preprint arXiv:1409.8276*, 2014.

[22] Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson-gamma dynamical systems. In *Advances in Neural Information Processing Systems*, pages 5005–5013, 2016.

[23] Lin Yuan and John D Kalbfleisch. On the Bessel distribution and related problems. *Annals of the Institute of Statistical Mathematics*, 52(3):438–447, 2000.

[24] R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[25] Roman N Makarov and Devin Glew. Exact simulation of Bessel diffusions. *Monte Carlo Methods and Applications*, 16(3-4):283–306, 2010.

[26] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

[27] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *The Journal of Machine Learning Research*, 17(1):5656–5699, 2016.

[28] Chengyue Gong et al. Deep dynamic Poisson factorization model. In *Advances in Neural Information Processing Systems*, pages 1666–1674, 2017.

[29] Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. Deep Poisson gamma dynamical systems. In *Advances in Neural Information Processing Systems*, pages 8442–8452, 2018.

[30] A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian factor analysis for dynamic count matrices. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

[31] Sikun Yang and Heinz Koeppl. Dependent relational gamma process models for longitudinal networks. In *International Conference on Machine Learning*, pages 5547–5556, 2018.

[32] M. Zhou and L. Carin. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems Twenty-Five*, pages 2546–2554, 2012.

[33] Mingyuan Zhou, Yulai Cong, and Bo Chen. The Poisson gamma belief network. In *Advances in Neural Information Processing Systems*, pages 3043–3051, 2015.

[34] A Taylan Cemgil and Onur Dikmen. Conjugate gamma Markov random fields for modelling nonstationary sources. In *International Conference on Independent Component Analysis and Signal Separation*, pages 697–705. Springer, 2007.

[35] Cédric Févotte, Jonathan Le Roux, and John R Hershey. Non-negative dynamical system with application to speech and audio. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3158–3162. IEEE, 2013.

[36] Ghassen Jerfel, Mehmet E Basbug, and Barbara E Engelhardt. Dynamic collaborative filtering with compound Poisson factorization. *arXiv preprint arXiv:1608.04839*, 2016.

[37] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

[38] Robert GD Steel. Relation between Poisson and multinomial distributions. 1953.

[39] Anirban Bhattacharya and David B Dunson. Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377, 2012.

[40] Jakob H Macke, Lars Buesing, John P Cunningham, M Yu Byron, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. In *Advances in neural information processing systems*, pages 1350–1358, 2011.

[41] Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108, 1961.

[42] Zoubin Ghahramani and Sam T Roweis. Learning nonlinear dynamical systems using an EM algorithm. In *Advances in neural information processing systems*, pages 431–437, 1999.

[43] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[44] Laurent Charlin, Rajesh Ranganath, James McInerney, and David M Blei. Dynamic Poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 155–162. ACM, 2015.

[45] Patrick T Brandt and Todd Sandler. A Bayesian Poisson vector autoregression model. *Political Analysis*, 20(3):292–315, 2012.

[46] Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

[47] Aleksandr Simma and Michael I Jordan. Modeling events with cascades of Poisson processes. *arXiv preprint arXiv:1203.3516*, 2012.

[48] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

[49] Luc Devroye. Simulating Bessel random variables. *Statistics & Probability Letters*, 57(3):249–257, 2002.

[50] NeurIPS corpus. UCI Machine Learning Repository.

[51] dblp computer science bibliography. http://dblp.uni-trier.de/.

[52] State of the Union Addresses (1790-2006) by United States Presidents. https://www.gutenberg.org/ebooks/5050?msg=welcome_stranger.

[53] K. Leetaru and P. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. Working paper, 2013.

[54] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. ICEWS coded event data. Harvard Dataverse, 2015. V10.

[55] Alex H Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I Ryu, Krishna V Shenoy, Mark Schnitzer, Tamara G Kolda, and Surya Ganguli. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115, 2018.