Dynamic Poisson Factor Analysis

Yizhe Zhang*, Yue Zhao†, Lawrence David*, Ricardo Henao* and Lawrence Carin*
*Duke University, Durham, NC 27708, USA Email:{yizhe.zhang,lawrence.david,r.henao,lcarin}@duke.edu
†Boston University, Boston, MA 02215, USA Email:yuezh@bu.edu

***Abstract*—We introduce a novel dynamic model for discrete time-series data, in which the temporal sampling may be nonuniform. The model is specified by constructing a hierarchy of Poisson factor analysis blocks, one for the *transitions* between latent states and the other for the *emissions* between latent states and observations. Latent variables are binary and linked to Poisson factor analysis via Bernoulli-Poisson specifications. The model is derived for count data but can be readily modified for binary observations. We derive efficient inference via Markov chain Monte Carlo, that scales with the number of non-zeros in the data and latent binary states, yielding significant acceleration compared to related models. Experimental results on benchmark data show the proposed model achieves state-of-the-art predictive performance. Additional experiments on microbiome data demonstrate applicability of the proposed model to interesting problems in computational biology where interpretability is of utmost importance.**

I. Introduction

Probabilistic models for high-dimensional time series have long been an area of significant interest in machine learning. The applicability of these models spans data from different domains, such as music sequences, text streams and time-series of molecular data in computational biology. Among this prior work, Hidden Markov Models (HMMs) [1] and the Linear Dynamical System (LDS) [2] are particularly well understood. However, in some cases these models are limited by the type of dynamic structures they can capture. In fact, real-world time-series data are usually described by complex nonlinear temporal dependencies, while traditional LDS models are restricted to latent representations described by linear dynamics. Models with discrete latent spaces, such as the HMM, are often specified as mixture models that represent the history of a time-series using multinomial distributions.

A newer class of time-series models, better suited to model complex (nonlinear) probability distributions over high-dimensional sequences, rely either on Recurrent Neural Networks (RNNs) [3]–[6], Restricted Boltzmann Machines (RBMs) [7]–[11] or Sigmoid Belief Networks (SBNs) [12]. The Temporal Restricted Boltzmann Machine (TRBM) [8] and the Temporal Sigmoid Belief Network (TSBN) [12], for instance, consist of a sequence of RBMs or SBNs, respectively, where the state of the current RBM or SBN is stochastically determined by previous RBMs or SBNs. Inference approaches for these models are non-trivial. Approximate procedures based mainly on Variational Bayes principles have been proposed and scale well to large datasets [8], [11], [12].

In the context of count data, multi-layer directed models are becoming increasingly popular [12]–[15]. In these models, the likelihood function connecting latent variables to observations is often specified in terms of Poisson distributions or softmax link functions, whereas latent (often deep) layers of the model are described by binary variables, capturing nonlinear dependencies and often modeled via SBNs [13], [14]. For time-series data, a similar idea has been leveraged, where a discrete transition model akin to HMMs connects current binary latent variables to previous ones (at earlier time points), but deviates from HMMs by specifying transitions via SBNs [12], not multinomial distributions.

In the work presented here, we propose a model for time-series data where both emission and transition models are specified in terms of Poisson distributions, borrowing ideas from deep Poisson factor models [15]. In particular, the Poisson-based transition model treats transition *strengths* as latent counts that are transformed to binary variables via the Bernoulli-Poisson (BP) link [16], a recently proposed alternative to the sigmoid link function. The BP link yields efficient learning and inference algorithms [15], [16]. In particular, the key advantage of modeling transitions with the BP link is that learning and inference scales with the number of non-zero latent variables, as opposed to the number of states (like is the case of TRBM or TSBN models), where the sigmoid link function is employed [8], [12].

The main contributions of this work are: 1) We develop a dynamic model for times-series data with count observations based on Poisson factor analysis. The Bernoulli-Poisson link formulation can be readily accommodated to time-series with binary observations. 2) Unlike most previously proposed models, our formulation easily allows for modeling data sampled *nonuniformly* in time. 3) An efficient sampling inference procedure is developed, scaling with the number of non-zeros in the data and binary latent variables, where latent counts and binary variables in the model can be updated in block. This allows our implementation to benefit from significant parallelization using GPUs (demonstrated in our experiments). 4) Results on benchmark and real data highlight the benefits of our modeling strategy from the standpoints of performance and interpretability; this is demonstrated via analysis of a new dynamic microbiome dataset.

II. Dynamic Poisson factor analysis

Assume observed counts for N time-series, where the vector of counts at each time point is of dimension M. The vector of counts at time t for the nth time series is denoted as $\mathbf{x}_{nt} \in \mathbb{N}^M$. Akin to HMMs, we model the dynamics of the time-series by imposing a *transition* model on latent variables, $\mathbf{h}_{nt} \in \{0,1\}^K$, where the state of the current latent variable (at time t) depends on the previous state (at time $t-1$), and K is the number of latent variables. The model allows 2^K different realizations of $\mathbf{h}_{nt}$, yielding 2^K states. However, the proposed model characterizes transition dynamics by more than just these discrete states (moving beyond an HMM), yielding improved modeling flexibility and results. The joint probability of the nth observation at time t is

$$p(\mathbf{X}_n, \mathbf{H}_n | \mathbf{\Xi}, \mathbf{\Omega}) = p(\mathbf{h}_{n0}) \prod_{t=1}^{T_n} p_{\mathbf{\Xi}}(\mathbf{x}_{nt}|\mathbf{h}_{nt}) \; p_{\mathbf{\Omega}}(\mathbf{h}_{nt}|\mathbf{h}_{nt-1}) \,, \quad (1)$$

where $\mathbf{X}_n = [\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nT_n}]$, $\mathbf{H}_n = [\mathbf{h}_{n1}, \ldots, \mathbf{h}_{nT_n}]$ and T_n is the length of the nth time-series. Further, $p(\mathbf{h}_{n0})$ is the prior for the initial value of the latent variables, $p_{\mathbf{\Xi}}(\mathbf{x}_{nt}|\mathbf{h}_{nt})$ is the *emission* model with parameters $\mathbf{\Xi}$ and $p_{\mathbf{\Omega}}(\mathbf{h}_{nt}|\mathbf{h}_{nt-1})$ is the transition model with parameters $\mathbf{\Omega}$.

a) Emission model: For the observed vector $\mathbf{x}_{nt}$, containing counts of M entities (*e.g.*, a vocabulary of M words), we impose the following *emission* model

$$\mathbf{x}_{nt} \sim \mathrm{Poisson}\left(\mathbf{\Psi}(\boldsymbol{\theta}_{nt} \circ \mathbf{h}_{nt})\right) \,, \quad (2)$$

where $\mathbf{\Psi} \in \mathbb{R}_+^{M \times K}$ is the *global* factor loadings matrix with K factors, shared by all time-series and time points; $\boldsymbol{\theta}_{nt} \in \mathbb{R}_+^K$ and $\mathbf{h}_{nt} \in \{0,1\}^K$ are *local* variables representing factor intensities and activations, respectively; and symbol $\circ$ represents the element-wise (Hadamard) product. Note that entries of $\mathbf{h}_{nt}$ indicate which factors are active for observation n at time t, *i.e.*, they define the *state* to which $\mathbf{x}_{nt}$ belongs.

Note that the Poisson emission parameters for data n are not just dependent on the state at time t, $\mathbf{h}_{nt}$. Binary activations, $\mathbf{h}_{nt}$, impose which columns of $\mathbf{\Psi}$ are employed to represent the Poisson rate, *and* the nonnegative intensities, $\boldsymbol{\theta}_{nt}$, provide temporal- and data-dependent scaling; in our experiments we have found this modeling flexibility to be important, particularly on real data, *e.g.*, the motivating microbiome data.

The model in (2) may be rewritten as

$$x_{mnt} = \textstyle\sum_{k=1}^{K} x_{mknt} \,, \quad x_{mknt} \sim \mathrm{Poisson}(\lambda_{mknt}) \,, \quad (3)$$

where $\lambda_{mknt} = \psi_{mk}\theta_{knt}h_{knt}$, x_{mnt} is component m of vector $\mathbf{x}_{nt}$, ψ_{mk} is component m of $\boldsymbol{\psi}_k$, $\boldsymbol{\psi}_k$ is column k of $\mathbf{\Psi}$, θ_{knt} is component k of vector $\boldsymbol{\theta}_n$, and h_{knt} is component k of vector $\mathbf{h}_n$. In (3) we have used the additive property of the Poisson distribution to decompose the mth observed count of $\mathbf{x}_{nt}$ as K latent counts, $\{x_{mknt}\}_{k=1}^K$. This decomposition allows derivation of efficient inference for the entire model, as discussed in Section III.

We specify prior distributions for the model in (2) as previously described [17], *i.e.*,

$$\begin{aligned} &\boldsymbol{\psi}_k \sim \mathrm{Dirichlet}(\eta_\psi \mathbf{1}_M) \,, \quad \theta_{knt} \sim \mathrm{Gamma}(r_k, b_\theta) \,, \\ &h_{knt} \sim \mathrm{Bernoulli}(\pi_{knt}) \,, \end{aligned} \quad (4)$$

where $\mathbf{1}_M$ is an M-dimensional vector of all-ones. Favoring simplicity, we let $\eta_\psi = 1/K$, $b_\theta = 0.5$ and $r_k \sim \mathrm{Gamma}(1,1)$. Prior distributions for η_ψ and b_θ that result in closed form conditionals exist, and can be used if desired; see for instance [18] for η_ψ, and [19] for b_θ.

The hierarchical model defined by (2) and (4), known as Poisson Factor Analysis (PFA), corresponds to the emission model, $p_{\mathbf{\Xi}}(\mathbf{x}_{nt}|\mathbf{h}_{nt})$, succinctly expressed in (1), with parameters $\mathbf{\Xi} = \{\mathbf{\Psi}, \boldsymbol{\theta}_{nt}, r_k\}$. These parameters can be interpreted in the context of topic modeling as follows: $\mathbf{\Psi}$ is a loadings matrix whose columns encode K topics (distributions over M words), that capture the correlation structure of observed variables; binary latent activations $\mathbf{h}_{nt}$ select which topics are used in time-series n at time t; and $\boldsymbol{\theta}_{nt}$, encode the intensities with which each topic is manifested in observation $\mathbf{x}_{nt}$. Interestingly, the PFA is closely related to other well-known topic model approaches, such as latent Dirichlet allocation, hierarchical Dirichlet processes and focused topic models [19].

b) Transition model: The most unique aspect of the proposed model is how state transitions are modeled, and how this allows nonuniform temporal sampling. In order to specify our model for latent variable transitions with respect to time, we first introduce the Bernoulli-Poisson link [16], a recently proposed probabilistic link function particularly useful at relating binary and count variables. For a binary vector $\mathbf{h}_{nt}$ with elements h_{knt},

$$h_{knt} = 1\left(z_{knt} > 0\right) \,, \quad z_{knt} \sim \mathrm{Poisson}\left(\tilde{\lambda}_{knt}\right) \,, \quad (5)$$

where z_{knt} is a latent count associated with binary variable h_{knt}, parameterized by a Poisson distribution with rate $\tilde{\lambda}_{knt}$. The function $1(\cdot)$ is defined as $1(\cdot) = 1$ if the argument holds, and otherwise $1(\cdot) = 0$. The model in (5), denoted here for short as $\mathbf{h}_{nt} \sim \mathrm{BPL}(\tilde{\boldsymbol{\lambda}}_{nt})$, for $\tilde{\boldsymbol{\lambda}}_{nt} \in \mathbb{R}_+^K$ with elements $\tilde{\lambda}_{knt}$, has the interesting property that $p(h_{knt} = 1) = \mathrm{Bernoulli}(\pi_{knt})$, where $\pi_{knt} = 1 - \exp\left(-\tilde{\lambda}_{knt}\right)$. This result can be shown by marginalizing out latent counts, z_{knt}, in (5). In fact, to sample h_{knt} we do not need to instantiate z_{knt}, but the rate of its underlying Poisson distribution, $\tilde{\lambda}_{knt}$.

The distribution implied by (5) is reminiscent of the complementary log-log link function [20], [21], where $\tilde{\lambda} = \exp(-u)$ and $u \in \mathbb{R}$. The logistic link function used in RBMs and SBNs is symmetric around the origin, $u = 0$, with $p(h = 1) = \mathrm{Bernoulli}(\pi)$, where $\pi = 1/(1 + \exp(-u))$; by contrast, the proposed BP link is *asymmetric*, which makes it appropriate for very sparse settings, where the proportion of zeros is large. In our case, this is particularly useful, because we can increase the number of latent variables without forcing the model to increase the number of states *a priori*.

Having defined the BP link, it becomes clear how to specify a transition model using the same Poisson factor analysis framework used for the emission model. Specifically, we write

$$\begin{aligned} &\mathbf{h}_{nt} \sim \mathrm{BPL}\left(\tau_{nt}^{-1}\mathbf{\Phi}(\mathbf{w}_{nt-1} \circ \mathbf{h}_{nt-1}) + \tilde{\boldsymbol{\lambda}}_0\right) \,, \\ &\boldsymbol{\phi}_k \sim \mathrm{Dirichlet}\left(\eta_\phi \mathbf{1}_K\right) \,, w_{knt-1} \sim \mathrm{Gamma}(s_k, b_w) \,, \end{aligned} \quad (6)$$

where $\tilde{\boldsymbol{\lambda}}_{nt} = \tau_{nt}^{-1}\mathbf{\Phi}(\mathbf{w}_{nt-1} \circ \mathbf{h}_{nt-1}) + \tilde{\boldsymbol{\lambda}}_0$ as in (5), $\boldsymbol{\phi}_k$ is a column of $\mathbf{\Phi}$ (transition factor matrix), w_{nkt-1} is an element of $\mathbf{w}_{nt-1}$ (*local* variable representing transition factor intensity), and η_ϕ, b_w and s_k are specified in a similar fashion to the emission model in (4). Bias term, $\tilde{\boldsymbol{\lambda}}_0$, controls the base rate of the Poisson distribution in (5) and is specified below. Parameter τ_{nt} is the time difference between observations t and $t-1$, for time-series n; τ_{nt} can vary with n and t, allowing *nonuniform* temporal sampling.

The specification in (6) corresponds to the transition model, $p_{\boldsymbol{\Omega}}(\mathbf{h}_{nt}|\mathbf{h}_{nt-1})$, in (1), with parameters $\boldsymbol{\Omega} = \{\boldsymbol{\Phi}, \mathbf{w}_{nt}, s_k, \tilde{\boldsymbol{\lambda}}_0\}$, for $n = 1, \ldots, N$, $t = 1, \ldots, T_n$ and $k = 1, \ldots, K$. These parameters have clear interpretations. For instance, $\boldsymbol{\Phi}$ is a transition matrix whose columns, $\boldsymbol{\phi}_k$, encode distinct transition *templates*. These templates are K-dimensional probability vectors and can be viewed as distributions over latent binary variable activations, $\mathbf{h}_{nt}$. Each template defines a particular activation pattern; some latent variables are co-active with high probability, while others are jointly absent, *i.e.*, they define correlation structure among elements of $\mathbf{h}_{nt}$. Interestingly, templates at time t are selected by previous activations $\mathbf{h}_{nt-1}$, and regulated (weighted) by previous intensities $\mathbf{w}_{nt-1}$. This implies that at time t the transition statistics are not only dependent on the 2^K discrete states $\mathbf{h}_{nt}$, but scale these states with weights $\mathbf{w}_{nt-1}$, adding modeling flexibility analogous to that in the emission model discussed above. In addition, the correlation between two adjacent time-points decays inversely with proximity in time, *i.e.*, as τ_{nt} increases, the dependency between latent variables $\mathbf{h}_{nt}$ and $\mathbf{h}_{nt-1}$ decreases. When τ_{nt} is large enough, binary activations loose their time dependency and effectively become a stochastic function of $\tilde{\boldsymbol{\lambda}}_0$, in fact, from (5), $p(h_{knt} = 1) = \text{Bernoulli}(1 - \exp(\tilde{\lambda}_{k0}))$, where $\tilde{\lambda}_{k0}$ is an element of $\tilde{\boldsymbol{\lambda}}_0$.

Note that the fact that intensities are time-dependent adds flexibility to the model compared to a simplified specification where these intensities are global parameters, *i.e.*, w_{kn} instead of w_{knt}. We observed empirically that the simplified model (with time-independent weights) does not perform as well as the specification in (6), however it is worth mentioning that time-dependent intensities may undermine the contribution of τ_{nt} to the transition model.

The emission model is completed by specifying a prior distribution for the initial state of the latent variables of the dynamic model, $p(\mathbf{h}_{n0})$, in (1). We let $h_{kn0} \sim \text{BPL}(\tilde{\lambda}_{k0})$ and $\tilde{\lambda}_{k0} \sim \text{Gamma}(a, b)$. For simplicity, we let $a = b = 1$, so that elements of $\mathbf{h}_{n0}$ are approximately uniform. Note that $p(\mathbf{h}_{n0})$ is a special case of (6) because since we do not have past information about $\mathbf{h}_{n0}$, conceptually, $\tau_{n0} \to \infty$.

c) Binary data: We can easily extend the dynamic Poisson factor model described above to model binary time-series data, by leveraging the same type of construction used for the transition model, *i.e.*, for $\mathbf{x}_{nt} \in \{0,1\}^M$, we can let $\mathbf{x}_{nt} \sim \text{BPL}\left(\boldsymbol{\Psi}(\boldsymbol{\theta}_{nt} \circ \mathbf{h}_{nt})\right)$ as in (6) but with prior distributions defined as in (4).

III. Learning and Inference

The dynamic Poisson factor model defined above has the convenient property of having all its conditional posteriors available in closed form, due to local conjugacy. In this paper, we focus on Markov Chain Monte Carlo (MCMC) via Gibbs sampling for learning and inference. Stochastic variational inference may be also implemented using ideas from [15].

During the *learning* phase, Gibbs sampling for the model in (2), (4), (5) and (6) involves sampling in sequence from the conditional posterior of all the global parameters of the model, namely, $\{\boldsymbol{\Psi}, r_k, \boldsymbol{\Phi}, s_k, \tilde{\lambda}_{k0}\}$, for $k = 1, \ldots, K$. During the *inference* phase, we sample from the conditional posterior of all local parameters, while also sampling from the *fixed* conditional posterior of the global parameters given the training data (obtained during learning). The local parameters include all latent variable activations, $\{\mathbf{h}_{nt}\}$, intensities for the emission model, $\boldsymbol{\theta}_{nt}$, and intensities for the transition model, $\mathbf{w}_{nt}$, where $n = 1, \ldots, N$, $t = 1, \ldots, T_n$. For prediction tasks, we perform learning on a training set, then perform inference on the test set, while sampling from the learned conditional posterior of the global parameters conditioned only on training data. In this way, we benefit from model averaging at test time.

The hyperparameters of the model are set to fixed values: $\eta_\psi = \eta_\phi = 1/K$, $b_\theta = b_w = 0.5$ and $a_0 = b_0 = 1$. Note that priors for η, b, a_0 and b_0 exist that result in Gibbs-style updates, and can be readily incorporated into the model if desired; however, our priority is keeping the model simple without sacrificing flexibility.

An important property from our dynamic PFA model is that inference does not scale with the size of the data, $\{M, N, T_n\}$, for $n = 1, \ldots, N$, and the number of factors, K, but as a function of their non-zero elements, which is tremendously advantageous in cases where the data is sparse, which is often the case. In order to show that this scaling behavior holds, it is enough to see that by construction, from (3), if $x_{mnt} = \sum_{k=1}^{K} x_{mknt} = 0$ (or z_{mnt}), thus $x_{mkn} = 0$, $\forall k$ with probability 1. From (5) we see that if $h_{knt} = 0$ then $z_{knt} = 0$ with probability 1. As a result, update equations for all parameters of the model except for binary activations $\mathbf{h}_{nt}$, depend only on non-zero elements of $\mathbf{x}_{nt}$ and $\mathbf{z}_{nt}$. Besides, updates for the binary variables can be cheaply obtained in block from $h_{knt} \sim \text{Bernoulli}(\pi_{knt})$ via λ_{knt}.

The most expensive operation in our algorithm is sampling from the multinomial distribution, to obtain latent counts for x_{mknt} (and z_{knt}), $\forall m, k, n, t$. Fortunately, conditioned on data $\mathbf{x}_{nt}$ and emission rates $\boldsymbol{\lambda}_{nt}$ (and $\tilde{\boldsymbol{\lambda}}_{nt}$ for transitions), $\forall n, t$, these can be sampled in block using a heavily parallelized implementation of the multinomial sampler via GPUs.

IV. Related work

Most of the existing work on directed models is based on the Sigmoid Belief Network (SBN) [22], which only recently has shown potential for building large multi-layer (deep) and dynamic models [12]–[14], [23]. Our model is most related with the Temporal SBN (TSBN) of [12], in which the emission model is an SBN or a softmax belief network for binary and count time-series, respectively, and the transition model is an SBN. The TSBN delivers fast inference via the Neural Variational Inference and Learning (NVIL) algorithm [13]. Our model is different from TSBN in three ways: (i) We use Bernoulli-Poisson links, not sigmoid links, for fast inference.

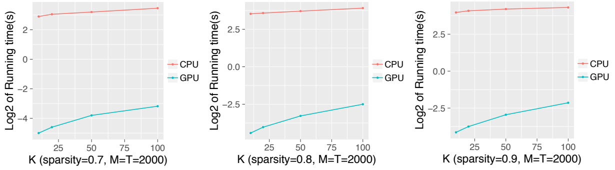


Fig. 1. Computational complexity of dynamic PFA on artificial data. Log$_2$-transformed runtime per full Gibbs iteration (in seconds) vs. number of binary latent variables, for datasets of fixed size, but different sparsity levels.

(ii) We model observed counts directly using Poisson distributions, as opposed to observed count proportions like in TSBN, via softmax link. (iii) Our inference approach scales with the number of non-zeros in the data and binary latent variables, which is not possible for models based on SBNs (or RBMs).

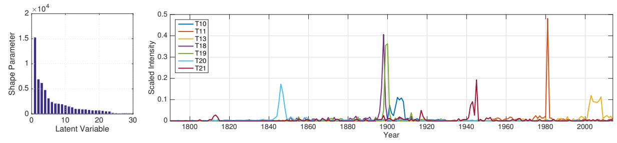
The work presented here is closely related to the deep Poisson factor model [15], a multi-layer (deep) topic model in which adjacent layers are connected via BP links similar to our model. The main differentiator between the deep PFA and our dynamic PFA is that in the former, BP links are introduced as a way to model correlation across latent variables whereas in ours, BP links allow us to model correlation across observations, by coupling adjacent time-points (transition model).

To the best of our knowledge there is only one existing approach for time-series modeling based on Poisson factor analysis [24]. In their work, dynamics are imposed using a linear model on the intensities, θ_{nt}, in terms on previous intensities θ_{nt-1} via a gamma distribution specification, $w_{knt} \sim \mathrm{Gamma}(w_{knt-1}, b_\theta)$. Our model is different in that we learn correlations across binary latent variables using $\mathbf{\Phi}$, whereas in [24] latent variables are *i.i.d. a priori* and restricted to linear dynamics. Unlike ours their model does not model latent variable activations, *i.e*, their model is *dense*.

All prior work on dynamic models using either RBMs, SBNs and PFA assumed uniform temporal sampling. The approach to nonuniform sampling introduced here specifies the dynamics of the transition model in terms of Poisson rates $\tilde{\boldsymbol{\lambda}}_{nt} = \tau_{nt}^{-1}\mathbf{\Phi}(\mathbf{w}_{nt-1} \circ \mathbf{h}_{nt-1}) + \tilde{\boldsymbol{\lambda}}_0$; the explicit dependence on sampling delay τ_{nt} and scaling $\mathbf{w}_{nt-1}$ moves well beyond only depending on the 2^K variants of the states $\mathbf{h}_{nt-1}$.

V. Experiments

a) Artificial data: We wish to evaluate quantitatively the parallelized implementation of our model using GPUs, to sample from the latent counts in (3), in terms of runtime. We compare this *efficient* implementation with our own Matlab and C++ implementation, which only differs from the GPU version in the multinomial sampler routine, where most of the runtime is spent. For the experiment we used a standard 4 core workstation with 24Mb RAM. The GPU is a Geforce GTX 750i with 640 cores and 2Mb RAM. We consider datasets of size $M = T = 2000$, $N = 1$ and different observed sparsity levels, namely fractional levels of sparsity in the set $\{0.7, 0.8, 0.9\}$. Figure 1 shows average runtime in seconds for a full Gibbs iteration, as a function of the number of binary latent variables, K, in dynamic PFA. We observe speedups ranging from 85 to 250x (120x average), which constitutes a substantial acceleration, considering the relatively outdated GPU used for the experiments. It is worth noting that sampling the latent counts represents about 90% of the total runtime.



T10	T11	T13	T18	T19	T20	T21
law	energy	iraq	cuba	islands	mexico	war
business	administration	terrorists	spanish	island	texas	enemy
conditions	percent	america	spain	phillipines	war	forces
labor	nuclear	terror	island	products	mexican	production
matter	programs	women	claim	convention	army	fighting
national	development	afghanistan	international	imperial	peace	japanese
industrial	policy	al	april	military	territory	german

Fig. 2. Selected topics learned from the State of the Union dataset. Top-left: Posterior mean of the shape parameter, r_k, for latent intensities, θ_{kn}. We show the top-30 largest values sorted in decreasing order. Top-right: Scaled intensities, $\theta_{knt}/\sum_t \theta_{knt}$, for selected topics. Scaling was done only to improve visualization. Bottom: Top words from selected topics. Each column represents the top-10 largest weights from 7 columns, $\boldsymbol{\psi}_k$, of loadings matrix, $\mathbf{\Psi}$, with corresponding intensity traces shown above.

b) State of the Union: This dataset contains transcripts of $T = 225$ US presidential State of the Union addresses, from 1790 to 2014. We consider the dataset as a single time-series ($N = 1$) with 225 time-points, where each transcript is a document, thus one document per year ($\tau_{nt} = 1$, $\forall t$ and $n = 1$). We preprocess the data lightly by removing stop words and terms that occur less than 7 times in one document or less that 20 times overall, which results in a vocabulary of size $M = 2375$ terms. This preprocessing scheme was previously used by [12] in their experiments with TSBNs.

For prediction tasks, we exclude the last year (2014) from the learning phase of the model. For all the other documents (years), 1790 to 2013, we randomly partition words from each document into a 80%/20% split. Learning is performed on the 80% subset, while the remaining 20% observations are used during inference to make predictions at each year. Predictions from both held-out sets are ranked according to their average predicted counts using the emission model. For our model, we run 900 iterations of the Gibbs sampler during learning and average predictions over 100 collection samples during inference. We verified empirically that increasing the number of Gibbs iterations does not significantly change results.

To evaluate the prediction performance, we calculate the precision @top-L as in [31], which is defined as the fraction of the top-L words, predicted by the model, that match the true ranking of the observed word counts. We use $L = 50$ as in [12]. We compare our dynamic PFA against three recently proposed models: Gamma Process Dynamic Factor Analysis (GP-DPFA) [24], Dynamic Rank Factor Model (DRFM) [25] and Temporal Sigmoid Belief Network (TSBN) [12]. The parameters of each of these models were selected to maximize performance. For DRFM and TSBN we used 25 latent variables and for GP-DPFA and our model, 100 latent variables. Note that unlike GP-DPFA, our model has latent binary activations that allow for model size selection, thus $K = 100$ can be seen as an upper bound on the total number of active latent variables. For this dataset we observed that the

TABLE I
AVERAGE PREDICTIVE PRECISION FOR STU DATASET.

Model	Mean Precision	Predictive Precision
Dynamic PFA	0.382	0.520
TSBN	0.327	0.353
GP-DPFA	0.223	0.189
DRFM	0.217	0.177

model estimated non-trivial activations for an average of 24 latent variables.

Predictive performance results are summarized in Table I. We report Mean Precision over all documents (years) in the dataset, restricted to the 20% held-out sets, one per year. We also report Predictive Precision for the final year, 2014. We see that our model significantly outperforms all the other methods in terms of mean precision and is comparable to TSBN in terms of predictive precision.

We now examine some of the parameters learned by the model to highlight the interpretability of our model. By looking at the conditional posterior of r_k, the global shape of intensities, θ_{kn}, we verified that the model only uses 24 latent variables from the $K = 100$ set *a priori*, see Figure 2(Top-left). This means that the reminding 76 latent variables are not active, $h_{knt} = 0$, or their intensities are close to zero, $\theta_{knt} \approx 0$, at any given time point. In Figure 2(Top-right) and 2(Bottom) we show intensity traces (θ_{knt}, $\forall t$) and top words (largest weights of ψ_k), respectively, from 7 of the 24 active topics estimated by the model. We observe very relevant topics, tightly localized in time. We see for example that Topic 10 is related to the organized labor movement, Topic 11 with the National Energy Program, Topics 13, 18, 19 and 20 focus on the Middle East, Spanish, Phillipines and Mexican wars, respectively. Finally, Topic 21 is related to the world wars. Topics not shown are less localized in time and range from foreign policy to internal economic affairs.

c) NIPS Abstracts: This dataset contains distributions of words (including authors) in all NIPS papers from years 1988 to 2003. Again, we consider the dataset as one time-series ($N = 1$) with 17 time-points, where each year is a document, thus $\tau_{nt} = 1$, $\forall t$ and $n = 1$. We preprocess the dataset using the same criteria used for the State of the Union data, which results in $M = 14,036$ distinct words.

TABLE II
AVERAGE PREDICTIVE PRECISION FOR NIPS ABSTRACTS.

Model	Mean Precision	Predictive precision
Dynamic PFA	0.876	0.664
TSBN	0.810	0.538

In this experiment, we focus on a quantitative evaluation using the performance metrics previously defined. Provided that TSBN consistently outperforms DRFM, GP-DPFA and TRBM [12], we only consider comparisons against TSBN here. In both models, the number of latent variables is set to $K = 50$. Predictive performance is summarized in Table II, from which we can see that our model outperforms TSBN by marked margins, using a 80/20% split for mean precisions and the last year, 2003, for predictive precisions.

d) Music: In this experiment we evaluate the model specification for binary time-series described in Section II. We employ the so-called music dataset [10]. This dataset consists of four pieces of polyphonic music sequences of piano, in which time-points, uniformly sampled in time, are encoded as 88-dimensional binary vectors indicating what keys of the piano, ranging from A0 to C8, are "pressed" at any given point in time. The number of simultaneous keys (polyphony) ranges from 0 (silence) to 15 with an average of 3.9 keys. The four pieces in the dataset, namely, Nottingham, Piano, Muse and JSB, correspond to folk tunes, a piano MIDI archive, orchestral music and chorales by J.S. Bach, respectively, and span different playing styles and tempo.

TABLE III
AUC FOR MUSIC DATA.

Dataset	Dynamic PFA	TSBN
Piano	0.9303	0.8784
Muse	0.9742	0.9466
JSB	0.9737	0.9686
Nottingham	0.9978	0.9964

Provided that the observed data is binary, we report area under the ROC curve (AUC) on the last time-point of each music piece, as performance metric. The last time point was not used during the learning phase. Table III shows once again that our dynamic PFA model consistently outperforms TSBN. It is worth mentioning that [12] reported results on these data, but using marginal log-likelihood estimates as performance metric. We opted for AUCs because we consider this a more fair metric, provided that for TSBN we will have to report a variational lower bound (conservative estimate) whereas in our case we will have to report a Monte Carlo estimate based on annealed important sampling, which is known to have the potential for overestimating the true marginal log-likelihood.

e) Microbiome: We conclude by considering a microbiome dataset composed of longitudinal measurements of human gut microbiota over time, from 6 subjects spanning 3 different studies, with 2 subjects per study. Details about sample collection and processing are detailed in [26]. Data are produced by DNA sequencing of microbiota samples, followed by processing and mapping of raw DNA reads into Operational Taxonomic Units (OTUs). Each OTU defines a species or a group of species, and is commonly used as unit of microbial diversity, and is represented in the data as a count. The total number of OTUs per subject is shown in Table IV. The sparsity level of these data is on average 85% (most OTUs are not observed at a given time point), and this sparsity is leveraged in the proposed model to yield significant computational acceleration.

Importantly, these data are sampled *nonuniformly* in time, and this complexity motivated several aspects of the proposed model (as detailed in Section II). All of the previous models against which we compared in the previous examples are not applicable here, as they assume uniform temporal sampling. Hence, we focus on results based upon our proposed model. As in the other experiments, we run 900 Gibbs iterations during learning, 100 collections samples during inference and we set the number of latent variables to $K = 50$. Each dataset, has time-series of different time lengths, sampling intervals and

TABLE IV

One-step ahead forecasting results on microbiome data, in terms of correlation. Error bars for dynamic PFA correspond to averages over 100 posterior samples.

Sample	M	T	Dynamic PFA	Naive
S1	5432	321	0.880±0.008	0.8614
S2	5432	189	0.755±0.044	0.378
S3	9371	30	0.989±0.003	0.990
S4	9371	30	0.964±0.006	0.960
S5	33750	332	0.943±0.003	0.935
S6	33750	129	0.975±0.002	0.943

OTU resolutions (see Table IV). In particular, data for which $T = 30$ were sampled once a day, whereas all the others where sampled nonuniformly over a total period of a year [26].

We first evaluate whether our model can make predictions at the last time-point, *i.e.*, one-step ahead forecasting, that correlate better with the ground truth measurements, compared to a *naive* approach in which we make predictions by assuming that observed OTUs at time T and $T-1$ do not change at all, *i.e.*, $\mathbf{x}_{nT} = \mathbf{x}_{nT-1}$, which is known to be a good assumption in some cases. Table IV shows Spearman correlations indicating that for 5 out of 6 time-series, modeling the dynamics of OTU changes over time has actual predictive value. We see these results on forecasting of OTU concentrations as interesting preliminary results that need to be further investigated.

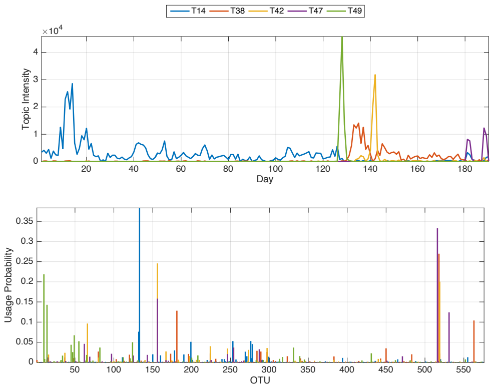


Fig. 3. Selected topics learned from microbiome data and particular to subject S5. Intensities, $\boldsymbol{\theta}_{nt}$, and OTU weights, ϕ_{mk}, for selected topics (T14, T38, T42, T47 and T49), in Top and Bottom panels, respectively. The bottom pannel of the figure shows weights for 5 different topics (colors), where each bar represent an element of ψ_k, a column of $\boldsymbol{\Psi}$.

Figure 3 show five selected topics from the model learned for subject S2, respectively. The proportion of non-zero intensities is 64%. We see that latent variable intensities, $\boldsymbol{\theta}_{nt}$, are nicely localized in time. We verified that Topic 49 is consistent with the onset of a *Salmonella* infection suffered by the subject (see [26]), while Topic 38 is related to its recovery period. Interestingly, we see that Topic 14, stably present up to the time of infection does not reappear even after recovery has taken place. Also interesting is that the five selected topics in Figure 3 only account for about 10% of the total OTUs in the sample ($\phi_{mk} > 10^{-4}$), which indicates that topics are not only localized in time but in OTU space. In particular, Topic 49 is enriched for *Proteobacteria*, and Topics 14 and 38 are enriched for *Firmicutes*, while Topic 42 is enriched for *Tenericutes*. All these results are consistent with the findings of [26], which were derived using a completely different, more biologically targeted approach.

VI. Conclusion and Future Work

We have introduced a dynamic time-series model based on Poisson factor analysis. The model allows for count and binary data, as well as for nonuniformly sampled time-series. Efficient inference using GPUs is developed, that scales with the number of non-zeros in the data and binary latent variables. Extensive results on benchmark data demonstrate the excellent performance of our simple yet elegant specification. Results on real microbiome data highlight the applicability of our model to interesting problems in modern computational biology.

References

[1] L. Rabiner and B. Juang, “An introduction to hidden Markov models,” in *ASSP Magazine, IEEE*, 1986.

[2] R. Kalman, “Mathematical description of linear dynamical systems,” in *J. SIAM, Series A: Control*, 1963.

[3] M. Hermans and B. Schrauwen, “Training and analysing deep recurrent neural networks,” in *NIPS*, 2013.

[4] J. Martens and I. Sutskever, “Learning recurrent neural networks with Hessian-free optimization,” in *ICML*, 2011.

[5] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *ICML*, 2013.

[6] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *ICML*, 2013.

[7] G. Taylor, G. Hinton, and S. Roweis, “Modeling human motion using binary latent variables,” in *NIPS*, 2006.

[8] I. Sutskever and G. Hinton, “Learning multilevel distributed representations for high-dimensional sequences,” in *AISTATS*, 2007.

[9] I. Sutskever, G. Hinton, and G. Taylor, “The recurrent temporal restricted Boltzmann machine,” in *NIPS*, 2009.

[10] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription,” in *ICML*, 2012.

[11] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee, “Structured recurrent temporal restricted Boltzmann machines,” in *ICML*, 2014.

[12] Z. Gan, C. Li, R. Henao, D. E. Carlson, and L. Carin, “Deep temporal Sigmoid belief networks for sequence modeling,” in *NIPS*, 2015.

[13] A. Mnih and K. Gregor, “Neural variational inference and learning in belief networks,” in *ICML*, 2014.

[14] Z. Gan, C. Chen, R. Henao, D. Carlson, and L. Carin, “Scalable deep Poisson factor analysis for topic modeling,” in *ICML*, 2015.

[15] R. Henao, Z. Gan, J. Lu, and L. Carin, “Deep Poisson factor modeling,” in *NIPS*, 2015.

[16] M. Zhou, “Infinite edge partition models for overlapping community detection and link prediction,” in *AISTATS*, 2015.

[17] M. Zhou, L. Hannah, D. Dunson, and L. Carin, “Beta-negative binomial process and Poisson factor analysis,” in *AISTATS*, 2012.

[18] M. D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *JASA*, vol. 90, no. 430, pp. 577–588, 1995.

[19] M. Zhou and L. Carin, “Negative binomial process count and mixture modeling,” *PAMI*, vol. 37, no. 2, pp. 307–320, 2015.

[20] W. W. Piegorsch, “Complementary log regression for generalized linear models,” *The American Statistician*, vol. 46, no. 2, pp. 94–99, 1992.

[21] D. Collett, *Modelling binary data*. CRC Press, 2002.

[22] R. Neal, “Connectionist learning of belief networks,” in *Artificial intelligence*, 1992.

[23] Z. Gan, R. Henao, D. Carlson, and L. Carin, “Learning deep sigmoid belief networks with data augmentation,” in *AISTATS*, 2015.

[24] A. Acharya, J. Ghosh, and M. Zhou, “Nonparametric Bayesian factor analysis for dynamic count matrices,” in *AISTATS*, 2015.

[25] S. Han, L. Du, E. Salazar, and L. Carin, “Dynamic rank factor model for text streams,” in *NIPS*, 2014.

[26] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, “Host lifestyle affects human microbiota on daily timescales,” *Genome Biology*, vol. 15, no. 7, 2014.