Switching Poisson Gamma Dynamical Systems

Wenchao Chen¹, Bo Chen¹*, Yicheng Liu¹, Qianru Zhao¹, Mingyuan Zhou²

¹ National Laboratory of Radar Signal Processing, Xidian University, Xian, China
² McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA
wcchen_xidian@163.com, bchen@mail.xidian.edu.cn, moooooore66@gmail.com, zqr951122@163.com mingyuan.zhou@mccombs.utexas.edu

Abstract

We propose switching Poisson-gamma dynamical systems (SPGDS) to model sequentially observed multivariate count data. Different from previous models, SPGDS assigns its latent variables into mixture of gamma distributed parameters to model complex sequences and describe the nonlinear dynamics, meanwhile, capture various temporal dependencies. For efficient inference, we develop a scalable hybrid stochastic gradient-MCMC and switching recurrent variational inference, which is scalable to large scale sequences and fast in out-of-sample prediction. Experiments on both unsupervised and supervised tasks demonstrate that the proposed model not only has excellent fitting and prediction performance on complex dynamic sequences, but also separates different dynamical patterns within them.

1 Introduction

Temporal sequences are abundant in real world and analyzing them is always an utmost important task in machine learning. Among them, time-series count data has attracted wide attention to deal with a variety of real-world applications, such as text analysis, social network modeling and natural language processing. The widely-used dynamic models, such as hidden Markov models (HMMs) [Rabiner and Juang, 1986] and linear dynamic systems (LDSs) [Ghahramani and Roweis, 1999], have difficulty in modeling such data, which are often high dimensional, sparse, and overdispersed. To address this issue, several dynamic methods have been proposed, especially based on the Poisson-gamma structure [Zhou et al., 2012; Zhou et al., 2016]. Specifically, Poisson-gamma dynamical systems (PGDS) [Schein et al., 2016] is a typical dynamic model for count sequence analysis and performs well in capturing cross-factor temporal dependence, which has been wildly used in text analysis, international relation study and so on. Generally, current developments of dynamic models are always focusing on modeling more complex sequences.

*: Corresponding author

In this paper, we propose switching Poisson-gamma dynamical systems (SPGDS), which is a powerful dynamical model that can capture different dynamics among time-steps, so as to model complex sequential relationships efficiently. Specifically, we build a dynamic system with the assumption that the latent variables of each time-step are drawn from the gamma mixture distributions, whose shape parameter for each mixture component is factorized into a linear transformation of the latent unit of previous time-step. Combining the temporal structure and mixture model, SPGDS can not only benefit the transmitting of nonlinear and diverse temporal variation for ample representational capacity by gamma mixture distributions, but also enable our model to cluster diverse dynamics among time-steps into different patterns. We introduce a discrete indicator variable $z_t$, called switching variable, to guide how the latent state $\theta_t$ varies from time $t-1$ to time $t$, as illustrated in Fig. 1 (a). The switching mechanism in our model serves two benefits: (1) it results in better fitting and prediction performance on complex count sequences; (2) it provides an insight into which dynamical patterns contains within sequences, which can be of value in application and analysis [Fraccaro et al., 2017; Becker Ehmck et al., 2019]. With the assumption that the real-world data changes in dynamics at time $t$ are causality, we assign $z_t$ to non-linearly depend on history input $x_{1:t−1}$ via a Gumble-softmax base recurrent variational inference network. Based on this, we further develop a Weibull-distribution-based switching recurrent variational inference network for structured inference [Krishnan et al., 2017] of latent variable $\theta_t$. This structured inference network enables SPGDS to learn rich latent representations and fast in out-of-sample prediction. The detail structure of our inference network is shown in Fig. 1 (b). Moreover, we develop a mini-batch based stochastic inference algorithm that combines stochastic gradient MCMC (SG-MCMC) [Patterson and W, 2013; Ma et al., 2015] and autoencoding variational inference, which accelerates our model in both training and testing phase for large scale sequences. Furthermore, to prove the flexibility and compatibility of our model with prevalent deep learning structures, we extend SPGDS into supervised SPGDS (sSPGDS), which incorporates the label information into the model and extract discriminative features to achieve enhanced performance on both data representation and classification.
2 Related Work

To model count sequence data, several dynamical model based on the Poisson-gamma construction has been proposed. The gamma process dynamic Poisson factor analysis (GP-DPFA) [Acharya et al., 2012] assumes that the data comes from the Poisson distribution and models the count vector at each time step under Poisson factor analysis (PFA) [Zhou et al., 2012] as \( x_t \sim \text{Pois}(\Phi \theta_t) \). It further smooths the translation through time by assigning \( \theta_t \sim \text{Gam}(\theta_{t-1}, \beta) \). Poisson-gamma dynamical systems (PGDS) [Schein et al., 2016] further improves the ability to capture cross-factor temporal dependence by introducing a translation matrix as \( \theta_t \sim \text{Gam}(\Pi \theta_{t-1}, \beta) \). Moreover, to capture long-range temporal dependencies and model long sequences, several multilayer probabilistic dynamic models are proposed. For instance, deep dynamic Poisson factor analysis (DDPFA) [Gong and Huang, 2017] combines recurrent neural network (RNN) [Martens and Sutskever, 2011] with PFA, to capture long-range temporal dependencies of the latent factor via RNN. Deep temporal sigmoid belief network (DTSBN) [Zhe et al., 2015] is an extension of deep sigmoid belief network (DSBN) [Gan et al., 2015] with sequential feedback loops on each layer. However, DTSBN restricts its hidden units to be binary, which limits its representational power. [Guo et al., 2018] extends PGDS into deep PGDS (DPGDS), a model with deep hierarchical latent structure and captures the correlations between the features across layers and over times using the gamma belief network [Zhou et al., 2016].

Although these methods exhibit attractive performance in describing temporal dependencies, due to the weights sharing across different time steps, there are still some difficulties for them in modeling sequences with very complex and highly nonlinear dynamics. However, switching linear dynamical systems (SLDS) [Linderman et al., 2016; Fraccaro et al., 2017; Becker Ehmk et al., 2019] is a widely used method to model those complex sequences by breaking down them into multiple simpler units and modeling them separately. But, it is difficult for SLDS to model count sequences for it is under the Gaussian assumption. In this paper, the proposed SPGDS could be seen as a novel switching extension of Poisson-gamma structure models, which not only fit the count sequence, but also inherits various virtues of switching dynamic models. Moreover, our model is also fast in out-of-sample prediction with the help of the structure variational inference network.

3 Switching Dynamical Systems for Count Sequences Modeling

In this section, we first propose a switching PGDS model for analysing count sequences. Then, we propose switching recurrent inference network to map the observation directly to the switching variables and latent representation of SPGDS, by which our model has benefited fast inference for out-of-sampling prediction.

3.1 Switching Poisson Gamma Dynamical Systems

We demonstrate the graphical representation of SPGDS in Fig. 1 (a). Assuming that dataset of \( V \)-dimensional sequentially observed multivariate count data \( x_1, ..., x_T \) are represented as a \( V \times T \) count matrix \( X \). The generation process of SPGDS can be expressed as

\[
\begin{aligned}
    z_t &\sim \text{Categorical}(r_t), \\
    \theta_t &\sim \prod_{c=1}^C \left( \text{Gam}(\tau_c, \Pi_c, \theta_{t-1}) \right)^{z_{tc}}, \\
    x_t &\sim \text{Pois}(\delta_F \Phi \theta_t),
\end{aligned}
\]

where the latent factors \( \Phi, \theta, \{\Pi_c\}_{c=1}^C, \{\tau_c\}_{c=1}^C \) and \( z \) are all positive. The input count data \( x_t \) at time \( t \) are factored into the loading matrix \( \Phi \in \mathbb{R}_+^{V \times K} \) and the corresponding hidden states \( \theta_t \in \mathbb{R}_+^{K \times C} \). And \( \delta \in \mathbb{R}_+ \) is the scaling parameter. We characterize the relationship between the hidden units of adjacent time-steps as multiple gamma distribution, so as to characterize the nonlinear sequential relationship between time-steps. \( z_t = (z_{t1}, ..., z_{tC}) \) is a \( C \)-dimensional categorical variable which choosing parameters \( \Pi_c \) and \( \tau_c \) at time \( t \) from different gamma distribution, we call it switching variable. Marginalizing \( z_t \) in (1), we get

\[
p \left( \theta_t | \theta_{t-1}, \{\Pi_c, \tau_c\}_{c=1}^C \right) = \sum_{c=1}^C r_t^C \text{Gam}(\theta_t | \tau_c \Pi_c, \theta_{t-1}, \tau_c),
\]
which is a mixture of gamma distribution to characterize the complex and diverse sequential relationship between time-steps more accurately than other Poisson-gamma structure models. \( \{ \Pi_c \in \mathbb{R}^{K \times K}_+ \}_{c=1}^C \) are the latent transition matrices that capture the various temporal dependencies between components, and \( C \) represents the number of components in mixture distribution. We denote \( \{ r_c \in \mathbb{R}_+ \}_{c=1}^C \) as the scaling parameters that control the variances of temporal amplitude variation of the hidden states. \( r_t = (r_{t1}, \ldots, r_{tC})^T \) is the parameter for categorical distribution. In this way, we can also label sequences into segmentations that exhibit different dynamics with \( z_t \). The vector \( \theta_t \) has an expected value of 

\[
E \left[ \theta_t | \theta_{t-1}, \{ \Pi_c \}_{c=1}^C \right] = \sum_{c=1}^C r_t^c \Pi_c \theta_{t-1},
\]

which suggests that \( \{ \Pi_c \}_{c=1}^C \) play the role of transiting the latent presentation across time, and controlled by \( r_t \). Hence, our model is capable to capture the primitive patterns of temporal sequences concisely and accurately, when the data contains multiple dynamics.

To complete the dynamic model, we introduce \( K \) factor weights in each components of mixture of gamma distribution:

\[
\pi^c_k \sim \text{Dir}(\nu^c_1, \ldots, \nu^c_{K-1}, \nu^c_K), \quad \nu^c_k \sim \text{Gam} \left( \frac{20}{K}, \beta^c \right),
\]

where \( \pi^c_k = (\pi_{1k}, \ldots, \pi_{Kk}) \) is the \( k \)-th column of \( \Pi_c \) and \( \{ \pi_{1k} \}_{c=1}^C \) can be interpreted as the \( C \) different probabilities for transitioning from component \( k \) to component \( k_1 \) for different dynamics. For the latent state at the first time-step, we define its prior as \( \theta_1 \sim \text{Gam} \left( \gamma_0, \frac{1}{\gamma_0} \right) \). Moreover, we place Dirichlet priors over the feature factors and draw the other parameters from a noninformative gamma prior: \( \phi_k = (\phi_{1k}, \ldots, \phi_{Kk}) \sim \text{Dir} \left( \eta, \ldots, \eta \right) \) and \( \xi(c), \beta(c) \sim \text{Gam} \left( \xi_0, \xi_0 \right), \tau(c) \sim \text{Gam} \left( \alpha_0, 1/\alpha_0 \right) \).

In particular, when \( C = 1 \), SPGDS reduces to PGDS. For non-count observations, we use Bernoulli-Poisson distribution [Zhou, 2015] and Poisson randomized gamma distribution [Aaron et al., 2019] to link binary observation and Nonnegative-real-valued observation to latent Poisson count, respectively. More details can be found in [Schein et al., 2016].

### 3.2 Switching Recurrent Inference Network for SPGDS

For efficient out-of-sample prediction, we develop a switching recurrent inference network, which will be used in hybrid SG-MCMC and variational inference method described in section 4, to map the observation directly to the latent variables. Specifically, we use Concrete distribution [Madison et al., 2016] to approximate the categorical distributed switching variables. Weibull distribution [Zhang et al., 2018] to approximate gamma distributed conditioned latent representations.

**Gumble-softmax based recurrent variational inference network for \( z_{tn} \):** Assuming there are \( N \) count sequential data \( \{ \mathbf{x}_{1n}, \ldots, \mathbf{x}_{Tn} \}_{n=1}^N \). It is clear that categorical variable \( z_{tn} \) in control of the changes of latent states \( \theta_{tn} \) from time \( t-1 \) to \( t \). In more general setting, the changes in dynamics at time \( t \) depend on the history of system and determined by the input ranging from 0 to \( t-1 \) [Fraccaro et al., 2017; Becker-Ehmck et al., 2019]. Here, we let \( z_{tn} \) determined by a learnable function of \( \mathbf{x}_{1: t-1, n} \) and modulate it by a recurrent variational parameter inference network. We construct the autoencoding variational distribution as \( q(z_{tn}) = \text{Categorical}(r_{tn}) \) and transform the observation to \( r_{tn} \) using recurrent structure as:

\[
r_{tn} = \text{softmax} \left( W_{xe} \mathbf{x}_{t-1, n} + W_{cc} r_{t-1, n} + b_{xn} \right) \quad (2)
\]

To obtain samples from categorical distribution, and to backpropagate through the categorical latent variables, we use \( q(z_{tn}) = \text{Gumbel} - \text{softmax}(r_{tn}) \) to approximate \( q(z_{tn}) \) [Maddison et al., 2016; Lang et al., 2016]. It draws samples via

\[
z_{tn}^c = \frac{\exp \left( (\log r_{tn}^c + g_{tn}^c) / \lambda \right)}{\sum_{c=1}^C \exp \left( (\log r_{tn}^c + g_{tn}^c) / \lambda \right)} \quad \text{for } c = 1, \ldots, C,
\]

where \( \lambda \) denotes the softmax temperature and \( g_{tn}^c \) is the standard uniform variable. As \( \lambda \) approaches 0, samples from the Gumbel-softmax distribution become one-hot and the Gumbel-softmax distribution \( q(z_{tn}) \) becomes identical to the categorical distribution \( q(z_{tn}) \).

**Weibull distribution based switching recurrent variational inference network for \( \theta_{tn} \):** Following [Zhang et al., 2018], we approximate the gamma distributed conditional posterior of \( \theta_{tn} \) with Weibull distribution by assigning \( q(\theta_{tn} | z_{tn}) \sim \text{Weibull}(k_{tn}, \lambda_{tn}) \), the random sample \( \epsilon_{tn} \) can be obtained by transforming standard uniform variables \( \epsilon_{tn} \) as \( \theta_{tn} = \lambda_{tn} \left( -\ln \left( 1 - \epsilon_{tn} \right) \right)^{1/k_{tn}} \), where \( k_{tn} \) and \( \lambda_{tn} \) are the parameters of \( \theta_{tn} \) and they are nonlinearly transformed from the hidden units \( h_t \) as

\[
k_{tn} = \ln \left[ 1 + \exp \left( W_{hh} h_{tn} + b_1 \right) \right],
\]

\[
\lambda_{tn} = \ln \left[ 1 + \exp \left( W_{hh} h_{tn} + b_2 \right) \right],
\]

where \( W_{hh} \in \mathbb{R}^{K \times K}, W_{hb} \in \mathbb{R}^{K \times \mathbb{R}}, b_1 \in \mathbb{R}^K, b_2 \in \mathbb{R}^K \). A nonlinear transformation deterministically covert \( h_{tn} \) from \( \mathbf{x}_{tn} \) and \( h_{t-1, n} \). To exploit the various temporal information, we propose a switching recurrent inference network considering diverse temporal dependence across time-steps, as illustrated in Fig. 1 (b). Therefore, \( h_{tn} \) can be expressed as

\[
h_{tn} = \prod_{c=1}^C f (W_{zh} \mathbf{x}_{tn} + W_{hc}^c h_{t-1, n} + b_3^c)^{z_{tn}^c},
\]

where \( f \) is a nonlinear transformation function, \( W_{zh} \in \mathbb{R}^{V \times K} \) is an input-hidden weight matrix, and \( \{ W_{hc}^c \in \mathbb{R}^{K \times \mathbb{R}} \}_{c=1}^C \) denote hidden states connected matrices. The detail structure of our inference network is shown in Fig. 1 (b).

### 4 Hybrid SG-MCMC and Variational Inference

In this section, we provide a hybrid stochastic gradient MCMC and autoencoding variational inference for SPGDS, which
is scalable in training phase and fast in testing phase. Aforementioned in Section 3, we develop a switching recurrent autoencoding variational inference network for switching variable $z_t$ and latent representation $\theta_t$, which enables our model to be fast in testing phase. For inferring global parameters in SPGDS, including $\Phi, \{\Pi_c\}_{c=1}^C$ and $\{\tau_c\}_{c=1}^C$, the topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC algorithm [Cong et al., 2017; Zhang et al., 2018], which is proposed to sample simplex-constrained global parameters in a mini-batch learning setting and improve its sampling efficiency by using the Fisher information matrix (FIM), can be extended to our model. More specifically, after sampling auxiliary latent counts using augmentable techniques as in [Guo et al., 2018], $\pi_k^{(c)}$, $4$th column of the transition matrix $\Pi^c$ of component $c$, can be sampled as

$$\pi_k^{(c)} + 1 = \left[ \pi_k^{(c)} + 1 \right] + \frac{1}{M_k^c} \left[ (\rho \sum_t z_{tc} \tilde{Z}_{kt} + \eta_k^{(c)}) \right]
- \frac{1}{M_k^c} \left[ \rho \sum_t z_{tc} \tilde{Z}_{kt} + \eta_k^{(c)} \right] \right]
+ \mathcal{N}(0, \frac{2\varepsilon_n}{M_k^c} \left[ \text{diag}(\pi_k^{(c)}) - (\pi_k^{(c)})(\pi_k^{(c)})^T \right]) \varphi,$$

where $\tilde{Z}$ comes from the augmented latent counts and the definition of $\rho, \varepsilon_n, \varphi$, and $M_k^c$ are analogous to [Cong et al., 2017]’s setting. The update of $\Phi$ is the same as [Guo et al., 2018]. For $\{\tau_c\}_{c=1}^C$, which are one-dimensional non-negative data, it is efficient to sample them with stochastic gradient Langevin dynamics [Welling and Teh, 2011] as

$$\tau_c^{(n+1)} = \tau_c^{(n)} + \frac{\varepsilon_n}{2} \ln \left( \tau_c^{(n)} \right) - \beta_0 + \sum_{i=1}^{N_n} \frac{\Pi_{c}^i \theta(t) - \theta(t)}{\text{exp}((\alpha_0 - 1) \ln (\tau_c^{(n)})) + \mathcal{N}(0, \varepsilon_n I))}$$

Given the global parameters $\Phi, \{\Pi_c\}_{c=1}^C, \{\tau_c\}_{c=1}^C$, the task is to optimize the parameters of the switching recurrent inference network. As the usual strategy of autoencoding variational inference, this optimization can be achieved by minimizing the negative Evidence Lower Bound (ELBO). We can express the ELBO of our inference network as:

$$L = \sum_{n=1}^{N} \sum_{t=1}^{T} E_q(z_{tn}) \left[ E_q(\theta_{tn}) \ln p(x_{tn}|\Phi, \theta_{tn}) +KL(q(\theta_{tn})||p(\theta_{tn}|\Pi, \theta_{t-1,n}, z_{tn})) \right]$$

$$-KL(q(z_{tn})||p(z_{tn}|\theta_{t-1,n})).$$

Denoting $\Omega$ as the parameters of the inference network: $\Omega = \{W_{xe}, W_{cc}, W_{sh}, \{W_{hc}^c\}_{c=1}^C, W_{hh}, W_{hl}\}$, the corresponding hybrid SG-MCMC and switching recurrent autoencoding variational inference method for SPGDS is described in Algorithm 1.

Furthermore, our model can be extended to a supervised version called supervised SPGDS (sSPGDS) to handle the categorization task of sequential data. We achieve this by concatenating the latent states across all time-steps to construct a $T \times K$-dimensional latent feature and add the softmax classifier on it [Wang et al., 2019]. Then, the loss function of the entire framework should be modified as

$$L = -L_g + \xi L_s,$$

where $L_g$ refers to ELBO of generative model shown in (7), $L_s$ denotes the classification criterion, and $\xi$ is a tradeoff parameter to balance aspects of generation and classification.

### 5 Experiments

In this section, We examine the performance of the proposed model on both unsupervised and supervised tasks.

#### 5.1 Unsupervised Models

We first examine the performance on fitting and prediction tasks of our proposed model, both synthetic datasets and real-world datasets are exploited here. Besides, we compare our model with some existing dynamic methods introduced in section 2, including HMM [Rabiner and Juang, 1986], LDS [Ghahramani and Roweis, 1999], GP-DPFA [Acharya et al., ] TSBN [Zhe et al., 2015] and PGDS [Schein et al., 2016]. We set the hyperparameters of GP-PFA, TSBN and PGDS the same as their original settings.

**Synthetic Dataset**

Inspired by [Gong and Huang, 2017], we consider three different multi-dimensional synthetic datasets:

- **Toydata 1:** $f_1(t) = t, f_2(x) = 2 \exp(-t/15) + \exp(-(t-25)^2/10^2)$ and $f_3(t) = 5 \sin(t^2) + 3$ for $t = 1, \ldots, 100.$
temporal dependencies and modeling them separately in the latent space. Taking Toydata 2 as an example, SPGDs captures three dynamic patterns at different time-steps, which are $t = 1 : 7$, $t = 50 : 51$ and $t = 7 : 50$, 51 : 99, as shown in Fig. 2. Corresponding transition matrices and weights to these three dynamic patterns are shown in Tab. 2. For stable sequential variation as dynamic 1, the transition matrix more closely approaches a diagonal matrix and transition weight is approximately equal to one. Relatively, as dynamic 2 and 3, transition matrix and weight changes with various temporal dependencies.

**Real-world Datasets**

Following [Gong and Huang, 2017; Schein et al., 2016], five real-world datasets are used:

- **Global Database of Events, Language, and Tone (GDELT):** GDELT is an international relationship dataset, which is extracted from news corpora.

- **Integrated Crisis Early Warning System (ICEWS):** ICEWS is another international relationship dataset extracted from news corpora.

- **State-of-the-Union transcript (SOTU):** The SOTU dataset contains the text of the annual SOTU speech transcripts from 1790 to 2014.

- **DBLP conference abstract (DBLP):** DBLP corpus is a database of computer research papers.


For each of these datasets, we summarize it as a $V \times T$ matrix, as shown in Tab.3. Specifically, we set $V = 1000$ for all data by choosing the top 1000 most frequently used features. Similar as previous methods [Zhe et al., 2015], we evaluate the prediction performance of our model by calculating the precision and recall at top-M as in [Han et al., 2014], which is given by the fraction of the predicted top-M words, that matches the true ranking of the words. $M$ is set as 50 here. We use three criterion MP, MR and PP. MP and MR are mean precision and mean recall over all years that appears in the training set, PP is the prediction precision for the final year. Moreover, we employ the setup in [Zhe et al., 2015] that the entire data of the last year is held-out, while the words of the each document for the documents in the previous years are randomly partitioned into 80%20% split. The 80% portion is used to train the model, and the prediction at the next year is tested on the rest of 20% held-out words. We compare the proposed model with several related works, including G-PDPFA, TSBN and PGDS, and the results are summarized in Table 3.

<table>
<thead>
<tr>
<th>Data</th>
<th>Measure</th>
<th>SPGDS</th>
<th>PGDS</th>
<th>HMM</th>
<th>LDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy1</td>
<td>MSE</td>
<td>1.17</td>
<td>2.11</td>
<td>27.59</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>PMSE</td>
<td>1.92</td>
<td>2.47</td>
<td>85.72</td>
<td>7.08</td>
</tr>
<tr>
<td>Toy2</td>
<td>MSE</td>
<td>35.12</td>
<td>48.24</td>
<td>83.98</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>PMSE</td>
<td>42.56</td>
<td>65.18</td>
<td>250.69</td>
<td>104</td>
</tr>
<tr>
<td>Toy3</td>
<td>MSE</td>
<td>102.31</td>
<td>180.25</td>
<td>400.18</td>
<td>210.35</td>
</tr>
<tr>
<td></td>
<td>PMSE</td>
<td>3.15</td>
<td>4.47</td>
<td>15.83</td>
<td>9.65</td>
</tr>
</tbody>
</table>

Table 1: Results on Synthetic Data

<table>
<thead>
<tr>
<th>Parameters</th>
<th>translation matrices</th>
<th>translation weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic 1</td>
<td>$\Pi_1 = \begin{bmatrix} 0.9996 &amp; 6 \times 10^{-5} \ 0.0004 &amp; 0.9999 \end{bmatrix}$</td>
<td>$\tau_1 = 0.9879$</td>
</tr>
<tr>
<td>Dynamic 2</td>
<td>$\Pi_2 = \begin{bmatrix} 0.9267 &amp; 0.3057 \ 0.0733 &amp; 0.8943 \end{bmatrix}$</td>
<td>$\tau_2 = 1.1281$</td>
</tr>
<tr>
<td>Dynamic 3</td>
<td>$\Pi_3 = \begin{bmatrix} 0.6731 &amp; 0.6025 \ 0.3273 &amp; 0.3987 \end{bmatrix}$</td>
<td>$\tau_3 = 0.2743$</td>
</tr>
</tbody>
</table>

Table 2: Corresponding transition matrices and weights to these three dynamic patterns in Fig. 2

Figure 2: Visualization of data (Top), latent factors (Middle), reconstruction data (Bottom) inferred by SPGDS with three mixture components from Toydata 2. $f_1(t)$, $f_2(t)$ and $f_3(t)$ are the three row vectors of Toydata 2 matrix. $\theta_1$, $\theta_2$ represent the two row vectors of latent factor $\theta$. $f_1(t)$, $f_2(t)$, $f_3(t)$ are the reconstruction of $f_1(t)$, $f_2(t)$ and $f_3(t)$. Temporal regions with different dynamic patterns are indicated through different colors.
increase in the number of papers accepted by NIPS. The sustainable growth of topics on "learning, network" indicates the decline of research on neuron memory patterns from 1994 to 2002. The superiority of SPGDS in modeling nonlinear sequences is clearly shown in Figure 3. SPGDS outperforms other methods in terms of MSE and PMSE on NIPS dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Top@M</th>
<th>GDELT</th>
<th>ICEWS</th>
<th>SOTU</th>
<th>DBLP</th>
<th>NIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$T = 365, V \approx 1000$</td>
<td>$T = 365, V \approx 1000$</td>
<td>$T = 225, V = 1000$</td>
<td>$T = 14, V = 1000$</td>
<td>$T = 17, V = 1000$</td>
</tr>
<tr>
<td>GPDPFA</td>
<td>MP</td>
<td>0.611 ± 0.001</td>
<td>0.607 ± 0.002</td>
<td>0.379 ± 0.002</td>
<td>0.435 ± 0.009</td>
<td>0.843 ± 0.005</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.145 ± 0.002</td>
<td>0.235 ± 0.005</td>
<td>0.369 ± 0.002</td>
<td>0.254 ± 0.005</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.447 ± 0.014</td>
<td>0.465 ± 0.008</td>
<td>0.617 ± 0.013</td>
<td>0.581 ± 0.011</td>
<td>0.807 ± 0.006</td>
</tr>
<tr>
<td>PGDS</td>
<td>MP</td>
<td>0.679 ± 0.001</td>
<td>0.658 ± 0.001</td>
<td>0.375 ± 0.002</td>
<td>0.419 ± 0.004</td>
<td>0.864 ± 0.004</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.150 ± 0.001</td>
<td>0.245 ± 0.005</td>
<td>0.373 ± 0.002</td>
<td>0.252 ± 0.004</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.420 ± 0.017</td>
<td>0.455 ± 0.008</td>
<td>0.612 ± 0.018</td>
<td>0.566 ± 0.008</td>
<td>0.802 ± 0.020</td>
</tr>
<tr>
<td>TSBN</td>
<td>MP</td>
<td>0.594 ± 0.007</td>
<td>0.471 ± 0.001</td>
<td>0.360 ± 0.001</td>
<td>0.405 ± 0.012</td>
<td>0.785 ± 0.005</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.124 ± 0.001</td>
<td>0.158 ± 0.001</td>
<td>0.275 ± 0.001</td>
<td>0.194 ± 0.001</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.418 ± 0.019</td>
<td>0.445 ± 0.001</td>
<td>0.611 ± 0.001</td>
<td>0.527 ± 0.003</td>
<td>0.692 ± 0.017</td>
</tr>
<tr>
<td>DTSBN-3</td>
<td>MP</td>
<td>0.411 ± 0.001</td>
<td>0.431 ± 0.001</td>
<td>0.370 ± 0.008</td>
<td>0.390 ± 0.002</td>
<td>0.774 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.141 ± 0.001</td>
<td>0.189 ± 0.001</td>
<td>0.274 ± 0.001</td>
<td>0.252 ± 0.004</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.367 ± 0.011</td>
<td>0.451 ± 0.012</td>
<td>0.548 ± 0.013</td>
<td>0.510 ± 0.006</td>
<td>0.715 ± 0.009</td>
</tr>
<tr>
<td>DPGDS-3</td>
<td>MP</td>
<td>0.689 ± 0.002</td>
<td>0.660 ± 0.001</td>
<td>0.380 ± 0.001</td>
<td>0.431 ± 0.012</td>
<td>0.887 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.150 ± 0.001</td>
<td>0.244 ± 0.003</td>
<td>0.374 ± 0.002</td>
<td>0.255 ± 0.004</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.456 ± 0.015</td>
<td>0.478 ± 0.024</td>
<td>0.628 ± 0.021</td>
<td>0.600 ± 0.001</td>
<td>0.839 ± 0.007</td>
</tr>
<tr>
<td>SPGDS</td>
<td>MP</td>
<td>0.705 ± 0.003</td>
<td>0.675 ± 0.003</td>
<td>0.380 ± 0.002</td>
<td>0.425 ± 0.004</td>
<td>0.890 ± 0.004</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.150 ± 0.001</td>
<td>0.253 ± 0.004</td>
<td>0.377 ± 0.002</td>
<td>0.257 ± 0.004</td>
<td>0.050 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>0.440 ± 0.015</td>
<td>0.450 ± 0.008</td>
<td>0.634 ± 0.028</td>
<td>0.605 ± 0.018</td>
<td>0.840 ± 0.010</td>
</tr>
</tbody>
</table>

Table 3: Top-M Results on Real-world Text Data

5.2 Supervised Models

To evaluate how well sSPGDS leverages the label information for feature learning, we compare its classification performance with a variety of algorithms on sequential MNIST dataset and permuted sequential MNIST dataset. For sequential MNIST, the pixels of MNIST digits [LeCun et al., 1998] are presented sequentially to the network and classification is performed at the end. By permuting the image sequences by a fixed random order, we can get permuted sequential MNIST dataset. Since the MNIST image are $28 \times 28$ pixels, they are reshaped into $784 \times 1$ sequences in sequential MNIST.

We compare our model with (a) RNN (relu), that is an RNN with RELU activations, (b) IRNN [Le et al., 2015], that is an RNN with RELU activations and with the recurrent weight matrix initialized to the identity, (c) Unitary evolution RNN (uRNN) [Arjovsky et al., 2016], that uses orthogonal and unitary matrices in RNN, (d) Full-Capacity Unitary RNN [Wisdom et al., 2016], which is a modified version of uRNN, (e) Skip RNN [Campos et al., 2018], an extending of existing RNN models, that learns to skip state updates and shortens the effective size of the computational graph, (f) long short-term...
memory (LSTM) [Greff et al., 2016] and (g) the gated recurrent unit (GRU) [Cho et al., 2014], which are the variants of RNN to address the gradient problem, (h) supervised Poisson gamma dynamical system, a supervised extension of PGDS. The classification accuracy of different methods are shown in Table 4, where the results of compared methods are provided in their corresponding papers. The latent dimension of models are 100 and Tensor Recurrent autoencoding network is developed based on RNN (relu). Clearly, our proposed sSPGDS can achieve a comparable performance among these methods and we contribute to the two characteristics of our model: (1) uncertainty is included in our hidden states [Chung et al., 2015] and (2) mixture distribution is assigned to our latent variables. These two characters enable sSPGDS to be robustness when track with complex and nonlinear sequential data. In addition, we show various dynamics captured by sSPGDS in Fig. 5. The number of mixture components is set as three. We reshape $I(z_t)$, which is the index for $z_t$ in (1), from $t = 2$ to $t = 785$ into $28 \times 28$ pixel and each pixel represents the index for dynamical patterns from the corresponding pixel in data to its next pixel. We visualize it by assign different colors to $I(z_t)$ with different values: $I(z_t) = 1 \rightarrow$ blue, $I(z_t) = 2 \rightarrow$ green, $I(z_t) = 3 \rightarrow$ yellow, and show it in the second row of Fig. 5. As we can see, our model captures three different dynamical patterns including: changing within noise or target, changing from noise to target and changing from target to noise. Moreover, we also visualize the probability of $p(I(z_t) = 1)$, $p(I(z_t) = 2)$ and $p(I(z_t) = 3)$ in third to fifth row of Fig. 5.

### 6 Conclusion

In this paper, we propose switching Poisson gamma dynamical systems (SPGDS) that takes advantage of gamma mixture distributions to model complex and nonlinear temporal dependencies, while capturing various dynamic patterns. To model the dependency of dynamic patterns among different time-steps and achieve fast out-of-sample prediction, a switching recurrent variational inference network is developed to infer the switching variable and latent representation. A mini-batch based stochastic inference method that combines both stochastic-gradient MCMC and variational inference algorithm is developed to accelerate both training and testing for large scale sequences. In addition, we provide a supervised extension of SPGDS. Experiment results show that our model not only has an excellent fitting and prediction performance on unsupervised feature extraction tasks, but also achieves comparable classification performance on supervised tasks.

### Acknowledgements

B. Chen acknowledges the support of the Program for Young Thousand Talent by Chinese Central Government, the 111 Project (No. B18039), NSFC (61771361), NSFC for Distinguished Young Scholars (61525105), Shaanxi Innovation Team Project, and the Innovation Fund of Xidian University. M. Zhou acknowledges the support of the U.S. National Science Foundation under Grant IIS-1812699.

### References


