Deep Switching Auto-Regressive Factorization: Application to Time Series Forecasting

Amirreza Farnoosh, Bahar Azari, Sarah Ostadabbas
Department of Electrical and Computer Engineering, Northeastern University,
Boston, Massachusetts, USA.
{afarnoosh, azari, ostadabbas}@ece.neu.edu

Abstract

We introduce deep switching auto-regressive factorization (DSARF), a deep generative model for spatio-temporal data with the capability to unravel recurring patterns in the data and perform robust short- and long-term predictions. Similar to other factor analysis methods, DSARF approximates high dimensional data by a product between time dependent weights and spatially dependent factors. These weights and factors are in turn represented in terms of lower dimensional latent variables that are inferred using stochastic variational inference. DSARF is different from the state-of-the-art techniques in that it parameterizes the weights in terms of a deep switching vector auto-regressive likelihood governed with a Markovian prior, which is able to capture the non-linear inter-dependencies among weights to characterize multimodal temporal dynamics. This results in a flexible hierarchical deep generative factor analysis model that (i) provides a collection of potentially interpretable states abstracted from the process dynamics, and (ii) performs short- and long-term vector time series prediction in a complex multi-relational setting. Our extensive experiments, which include simulated data and real data from a wide range of applications such as climate change, weather forecasting, traffic, infectious disease spread and non-linear physical systems attest the superior performance of DSARF in terms of long- and short-term prediction error, when compared with the state-of-the-art methods[1].

Introduction

Ever-improving sensing technologies offer fast and accurate collection of large-scale spatio-temporal data in various applications, ranging from medicine and biology to marketing and traffic control. In these domains, modeling the temporal dynamics and spatial relations of data have been investigated and analysed from different perspectives.

As these multivariate spatio-temporal data often exhibit high levels of correlation between dimensions, they can naturally be thought of as governed by a smaller number of underlying components. Tensor/matrix factorization frameworks are used to describe variability in these correlated dimensions in terms of potentially lower dimensional unobserved variables, namely temporal *weights* and spatial *factors*. Accordingly, Bayesian probabilistic global matrix/tensor factorization has been investigated in Salakhutdinov and Mnih (2008); Zhao, Zhang, and Cichocki (2015); Chen et al. (2019) for time series imputation. Besides, linear temporal dynamics have been adapted into this framework in Xiong et al. (2010); Charlin et al. (2015); Sun and Chen (2019). A number of non-Bayesian dynamical matrix factorization methods have been explored in Rogers, Li, and Russell (2013); Sun, Parthasarathy, and Varshney (2014); Bahadori, Yu, and Liu (2014); Cai et al. (2015); Yu, Rao, and Dhillon (2016); Takeuchi, Kashima, and Ueda (2017); Jing et al. (2018). Amongst these methods, some assume a linear vector auto-regressive model for the temporal weights (Bahadori, Yu, and Liu 2014; Yu, Rao, and Dhillon 2016), and spatial factors (Takeuchi, Kashima, and Ueda 2017) to model higher-order auto-regressive dependencies in multivariate time series data.

From another perspective, Bayesian switching linear state-space models, (Chang and Athans 1978; Hamilton 1990; Ghahramani and Hinton 2000; Murphy 1998; Fox et al. 2009; Linderman et al. 2017; Nassar et al. 2019; Becker-Ehmck, Peters, and Van Der Smagt 2019), have provided a more flexible structure for modeling temporal dynamics characterized by several modes. These models are specifically useful in the applications where complex dynamical behaviors can be broken down into simpler potentially interpretable units, which in turn provides additional insight into the rich processes generating complex natural phenomena. These models achieve globally nonlinear dynamics by composing linear systems through switching, (Sontag 1981). Besides, Gaussian state space models adapting neural networks have been used for approximating nonlinear first-order temporal dynamics in Krishnan, Shalit, and Sontag (2015); Watter et al. (2015); Karl et al. (2017); Krishnan, Shalit, and Sontag (2017); Fraccaro et al. (2017); Becker et al. (2019); Farnoosh et al. (2020).

In this paper, we introduce deep switching auto-regressive factorization (DSARF) in a Bayesian framework. Our method adds to the current body of knowledge by extending switching linear dynamical system models and Bayesian dynamical matrix factorization methods and combining their favorable properties. Specifically, for temporal dynamic modeling, we employ a non-linear vector auto-regressive latent model parameterized by neural networks and governed by a Markovian chain of discrete switches to capture higher-order multimodal latent dependencies. This will provide a more flexible model

[1]The source code and experiments are available at https://github.com/ostadabbas/DSARF

that expands prediction horizon and improves long- and short-term forecasting. In addition, we use a deep generative prior for estimation of multimodal distributions for spatial factors. We leverage the tensor/matrix factorization framework to make our model scalable to high dimensional data and solve this model efficiently with approximate variational inference.

Our hierarchical generative model with the help of corresponding learning and inference algorithm is able to handle missing data and to provide uncertainty measures for estimations. We demonstrate our model performance using experiments including simulated and real-world data from a wide range of application areas. Our experiments show that DSARF achieves better predictive performance on unseen data relative to current state-of-the-art methods when evaluated based on short- and long-term prediction errors. In the following section, we provide more motivation and background regarding our main contributions.

Background and Motivation

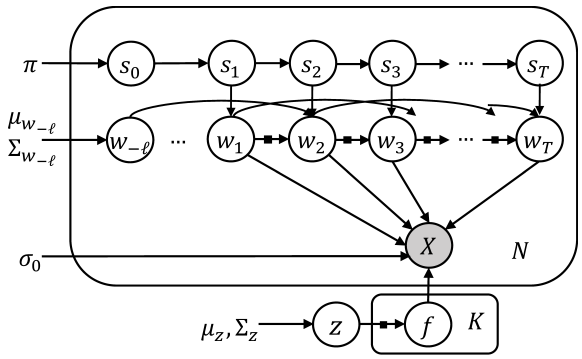
Our hierarchical model consists of three components: switching dynamical systems, non-linear vector auto-regression, and tensor/matrix factorization, for which we provide reviews and justifications in this section.

Linear Gaussian dynamical systems operating in Markov dependent switching environment have long been investigated in the literature, (Ackerson and Fu 1970; Chang and Athans 1978; Hamilton 1990; Ghahramani and Hinton 2000; Murphy 1998; Fox et al. 2009). These models, also known as switching linear dynamical system (SLDS), decompose nonlinear time series data into series of simpler, repeated dynamical modes. The SLDS model learns the underlying nonlinear generative process of the data as it breaks down the data sequences into coherent, potentially interpretable, discrete units, similar to the *piecewise affine* (PWA) framework in control systems (Sontag 1981; Juloski, Weiland, and Heemels 2005; Paoletti et al. 2007) . The generative process starts with sampling a discrete latent state $s_t \in \{1, \dots, S\}$ at each time $t = 1, \dots, T$ according to Markovian dynamics $s_t|s_{t-1}, \mathbf{\Phi} \sim \pi_{s_{t-1}}$, where $\mathbf{\Phi}$ is the Markov transition matrix and π_s is the categorical distribution parameter. Then, a continuous latent state $w_t \in \mathbb{R}^K$ is sampled from a normal distribution whose mean follows a conditionally linear dynamics as $w_t = \mathbf{A}_{s_t} w_{t-1} + \mathbf{b}_{s_t} + \nu_{t-1},\ \nu_{t-1} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{Q}_{s_t})$, for matrices $\mathbf{A}_s, \mathbf{Q}_s \in \mathbb{R}^{K \times K}$ and vectors $\mathbf{b}_s \in \mathbb{R}^K$ for $s = 1, 2, \dots, S$. Finally, a linear Gaussian observation $x_t \in \mathbb{R}^D$ is generated from the continuous latent state w_t according to $x_t = \mathbf{C}_{s_t} w_t + \mathbf{d}_{s_t} + \mu_t,\ \mu_t \overset{iid}{\sim} \mathcal{N}(0, \mathbf{G}_{s_t})$, for matrices $\mathbf{C}_s \in \mathbb{R}^{D \times K}, \mathbf{G}_s \in \mathbb{R}^{D \times D}$ and vectors $\mathbf{d}_s \in \mathbb{R}^D$. SLDS parameters are learned in a Bayesian inference approach. In this framework, the probabilistic dependencies are in such a way that $s_{t+1}|s_t$ is independent of the continuous state w_t, and hence the model cannot learn the transition of the discrete latent state when continuous latent state enters a particular region of state space. This problem is addressed in recurrent switching linear dynamical system (rSLDS), (Linderman et al. 2017; Nassar et al. 2019) by allowing the discrete state transition probabilities to depend on the preceding continuous latent state, i.e, $s_t|s_{t-1}, w_{t-1}$. rSLDS studies proposed to use auxiliary variable methods for approximate inference in a multi-stage training process. Nassar et al. (2019) extended rSLDS of Linderman et al. (2017) by enforcing a tree-structured prior on the switching variables in which subtrees share similar dynamics. Becker-Ehmck, Peters, and Van Der Smagt (2019) proposed to learn an rSLDS model through a recurrent variational autoencoder (rVAE) framework, and approximated switching variables by a continuous relaxation. This amortized inference compromised the applicability of their model on missing data, as they only included physics-simulated experiments.

The rSLDS prediction horizon is, however, limited as it adopts first-order linear Markovian dynamics, a prevalent model in the literature, for both discrete and continuous latent state. On the other hand, we advocate the use of higher-order dependencies in an auto-regressive approach, as in Bahadori, Yu, and Liu (2014); Wulsin, Fox, and Litt (2014); Sun and Chen (2019), namely we introduce deep generative **vector auto-regressive** priors for the continuous latent variable, w_t, which gives more flexibility to the model to expand its prediction horizon and capture higher-order non-linear auto-regressive relations amongst its continuous latents. More specifically, we use a weighted combination of a linear and a non-linear transformation as the relation between two continuous latent variables w_t and $w_{t-\ell}$, where ℓ is a lag set.

Bayesian tensor/matrix factorization constitutes the outermost layer of our hierarchical probabilistic model offering an effective approach to convert massive data into a lower-dimensional and computationally more tractable set of latent, e.g., temporal and spatial components. Tensor/matrix factorization frameworks, (Sun, Parthasarathy, and Varshney 2014; Bahadori, Yu, and Liu 2014; Cai et al. 2015; Zhao, Zhang, and Cichocki 2015; Yu, Rao, and Dhillon 2016; Takeuchi, Kashima, and Ueda 2017; Chen et al. 2019), are used to describe variability in high dimensional correlated data in terms of potentially lower dimensional unobserved variables called *factors*. In other words, given an observation matrix $X \in \mathbb{R}^{T \times D}$ of spatio-temporal data with T time points and D spatial locations, these methods decompose X into a set of $K \ll D$ temporal factors (weights) $W \in \mathbb{R}^{K \times T}$, and spatial factors $F \in \mathbb{R}^{K \times D}$ as $X \approx W^\top F$, where temporal dynamics are modeled in W as first-, e.g., in Sun, Parthasarathy, and Varshney (2014); Cai et al. (2015), or higher-order linear dependencies, e.g., in Bahadori, Yu, and Liu (2014); Yu, Rao, and Dhillon (2016); Takeuchi, Kashima, and Ueda (2017).

While the focus of this paper is on Bayesian *switching* dynamical modeling, several studies have employed neural networks for non-linear **Markovian state-space modeling**, (Krishnan, Shalit, and Sontag 2015; Watter et al. 2015; Karl et al. 2017; Krishnan, Shalit, and Sontag 2017; Fraccaro et al. 2017; Becker et al. 2019) (a.k.a. SSMs), and **multi-dimansional times series forecasting**, (Chang et al. 2018; Lai et al. 2018; Rangapuram et al. 2018; Li et al. 2019; Sen, Yu, and Dhillon 2019; Salinas et al. 2020) (denoted by fNNs here). Deep SSMs operate in an encoding/decoding framework (similar to VAEs), and are restricted to first-order Markovian dependencies. fNNs estimate model parameters from input data using recurrent neural networks (RNNs), e.g., in Chang et al. (2018); Lai et al. (2018); Salinas et al. (2020),



$p_\theta(w_t \mid w_{t-\ell}, s_t = s)$	$p_\theta(F \mid z)$
(1) : $w_{t-\ell} \in \mathbb{R}^{\lvert\ell\rvert \times K}$	(1) : $z \in \mathbb{R}^2$
(1) $\rightarrow \left[\text{FC}_{s,l}\ K \times K\right]_{l=1}^{\lvert\ell\rvert}$ PReLU $\rightarrow$ (2) $\frac{1}{\lvert\ell\rvert}\sum_l$ (2) $\rightarrow$ FC$_s$ $K \times K$ Sigmoid $\rightarrow$ (3) (3) : $g_s \in \mathbb{R}^K$ (1) $\rightarrow \left[\text{FC}_{s,l}\ K \times K\right]_{l=1}^{\lvert\ell\rvert}$ PReLU $\rightarrow$ (4) $\frac{1}{\lvert\ell\rvert}\sum_l$ (4) $\rightarrow$ FC$_s$ $K \times K \rightarrow$ (5) (5) : $\mu_{\theta_s}^w \in \mathbb{R}^K$ (1) $\rightarrow \left[\text{FC}_{s,l}\ K \times K\right]_{l=1}^{\lvert\ell\rvert} \rightarrow$ (6) $\frac{1}{\lvert\ell\rvert}\sum_l$ (6) : VAR$_{\theta_s} \in \mathbb{R}^K$ (5) $\rightarrow$ PReLU FC$_s$ $K \times K \rightarrow$ (7) (7) : $\log\Sigma_{\theta_s}^w \in \mathbb{R}^K$	(1) $\rightarrow$ FC 2×4 PReLU $\rightarrow$ (2) (2) $\rightarrow$ FC 4×8 PReLU $\rightarrow$ (3) (3) $\rightarrow$ FC $8 \times 2KD \rightarrow$ (4) (4) : $(\mu_\theta^F, \log\Sigma_\theta^F) \in \mathbb{R}^{KD,\,KD}$

Figure 1: Left: Graphical model representation for deep switching auto-regressive factorization (DSARF). Temporal weights, $w_{1:T}$ are generated according to a nonlinear auto-regressive model, *switched* by a Markovian chain of discrete states, $s_{1:t}$. Spatial factors, $f_{1:K}$, come from a shared low-dimensional latent, z. The solid black squares represent nonlinear functions. Right: Network architectures parameterizing the nonlinear mappings employed in DSARF. A fully connected (FC) layer is defined for each state $s \in \{1, \ldots, S\}$, and lag $l \in \ell$ as FC$_{s,l}$. These layers take as input $w_{t-\ell}$, and their outputs are aggregated in the succeeding layer: e.g., $\boldsymbol{\mu_{\theta_s}^w} = \text{FC}_s(\frac{1}{|\ell|}\sum_{l\in\ell}\text{PReLU}(\text{FC}_{s,l}(w_{t-l})))$.

Transformers, in Li et al. (2019), or temporal convolution networks (TCNs), in Sen, Yu, and Dhillon (2019). While linear vector auto-regression on high-dimensional input data is adopted in fNNs of Chang et al. (2018); Lai et al. (2018), most fNNs employ first-order autoregressive models. In addition, many SSMs and fNNs are not naturally tractable to data with missing values (without e.g., prior imputation or zero-filling), since, in training, target values, $x_{1:T}$, are fed directly to a neural network for model estimation, e.g, in fNNs, or variational estimation, i.e, amortized inference in SSMs (see Che et al. (2018); Ghazi et al. (2018); Rangapuram et al. (2018); Mattei and Frellsen (2019) for a discussion). As such, the two recent works, Sen, Yu, and Dhillon (2019); Salinas et al. (2020), excluded time series with missing values from their experiments. This is a major motivation for non-amortized inference in DSARF and some of the comparison baselines in this paper as the datasets in our experiments have up to 50% missing values. In the following section, we provide detailed formulation of DSARF model and its inference procedure.

Problem Formulation: Deep Switching Auto-Regressive Factorization

We consider a corpus of N spatio-temporal data $\{X_n\}_{n=1}^N$, where each $X_n \in \mathbb{R}^{T\times D}$ contains T time points and D spatial locations. We assume that X_n can be decomposed into a weighted summation of $K \ll D$ factors over time:

$$X_n \approx [w_1, \cdots, w_T]_n^\top [f_1; \cdots; f_K] = W_n^\top F, \tag{1}$$

where $f_k \in \mathbb{R}^D$ is the k^{th} spatial factor and $w_t \in \mathbb{R}^K$ are weights at time t. In order to model temporal dynamics, we assume that these low dimensional weights, $W = \{w_t\}_{t=1}^T$, are generated in accordance with a set of temporal lags, ℓ, through a deep probabilistic *switching* auto-regressive model, governed by a Markovian chain of *discrete* latent states, $\mathcal{S} = \{s_t\}_{t=1}^T$ as follows: $w_t \sim p(w_t|w_{t-\ell}, s_t)$, $s_t \sim p(s_t|s_{t-1})$. In other words, in the underlying state-space model of data, w_t is conditioned on $w_{t-\ell}$ (weights at the temporal lags specified in ℓ), and s_t (state of the model at time t). This encourages a *multimodal* distribution for the temporal generative model. We further assume that spatial factors, $F = \{f_k\}_{k=1}^K$, come from a shared low dimensional latent variable, z, which ensures the estimation of a multimodal distribution for the spatial factors as follows: $f_{1:K} \sim p(F|z)$, $z \sim p(z)$.

These assumptions define the graphical model for DSARF in Fig. 1. We train this hierarchical model using stochastic variational methods (Hoffman et al. 2013; Ranganath et al. 2013; Kingma and Welling 2014; Rezende and Mohamed 2015). These methods approximate the posterior $p_\theta(\mathcal{S}, W, z, F|X)$ using a variational distribution $q_\phi(\mathcal{S}, W, z, F)$ by maximizing a lower bound (known as ELBO) $\mathcal{L}(\theta, \phi) \leq \log p_\theta(X)$:

$$\begin{aligned}\mathcal{L}(\theta,\phi) &= \mathbb{E}_{q_\phi(\mathcal{S},W,z,F)}\left[\log \frac{p_\theta(X,\mathcal{S},W,z,F)}{q_\phi(\mathcal{S},W,z,F)}\right] \\ &= \log p_\theta(X) - \text{KL}(q_\phi(\mathcal{S},W,z,F)\,||\,p_\theta(\mathcal{S},W,z,F|X)).\end{aligned} \tag{2}$$

By maximizing this bound with respect to the parameters θ, we learn a deep generative model that defines a distribution over datasets $p_\theta(X)$. By maximizing the bound over the parameters ϕ, we perform Bayesian inference by approximating the distribution $q_\phi(\mathcal{S}, W, z, F) \simeq p_\theta(\mathcal{S}, W, z, F|X)$ over latent variables for each data point. Considering the proposed generative model, the joint distribution of observations and latents will be (denoting $\mathcal{Z} = \{W, z, F\}$ for brevity):

$$\begin{aligned}p_\theta(X,\mathcal{S},\mathcal{Z}) = p(F|z)p(z)\prod_{n=1}^N p(X_n|W_n,F)p(w_{n,-\ell})p(s_{n,0}) \\ \prod_{t=1}^T p(s_{n,t}|s_{n,t-1})p(w_{n,t}|w_{n,t-\ell},s_{n,t})\end{aligned} \tag{3}$$

We approximate the posterior distributions of latents with a fully factorized variational distribution:

$$q(\mathcal{S},\mathcal{Z}) = q(F)q(z)\prod_{n=1}^N q(w_{n,-\ell})q(s_{n,0})\prod_{t=1}^T q(s_{n,t})q(w_{n,t}) \tag{4}$$

The ELBO is then derived by plugging in $p_\theta(\cdot)$ and $q_\phi(\cdot)$ from Eq. (3) and Eq. (4) respectively into Eq. (2):

$$\mathcal{L}(\theta,\phi) = \sum_{n=1}^{N}\Big(\boldsymbol{\mathcal{L}_n^{\text{rec}}} + \boldsymbol{\mathcal{L}_n^{s_0,\,w_{-\ell}}} + \sum_{t=1}^{T}\big(\boldsymbol{\mathcal{L}_{t,n}^{\mathcal{S}}} + \boldsymbol{\mathcal{L}_{t,n}^{W}}\big)\Big) + \boldsymbol{\mathcal{L}^{\text{F}}},$$

$$\boldsymbol{\mathcal{L}_n^{\text{rec}}} = \mathbb{E}_{q(w_{n,1:T},F)}\Big[\log p(X_n|w_{n,1:T},F)\Big]$$

$$\boldsymbol{\mathcal{L}_n^{s_0,\,w_{-\ell}}} = -\text{KL}\big(q(s_{n,0})||p(s_0)\big) - \text{KL}\big(q(w_{n,-\ell})||p(w_{-\ell})\big)$$

$$\boldsymbol{\mathcal{L}_{n,t}^{\mathcal{S}}} = -\mathbb{E}_{q(s_{n,t-1})}\Big[\text{KL}\big(q(s_{n,t})||p(s_{n,t}|s_{n,t-1})\big)\Big]$$

$$\boldsymbol{\mathcal{L}_{n,t}^{\mathcal{W}}} = -\mathbb{E}_{q(s_{n,t})q(w_{n,t-\ell})}\Big[\text{KL}\big(q(w_{n,t})\|p\big(w_{n,t}|w_{n,t-\ell},s_{n,t}\big)\big)\Big]$$

$$\boldsymbol{\mathcal{L}^{F}} = -\mathbb{E}_{q(z)}\Big[\text{KL}\big(q(F)||p(F|z)\big)\Big] - \text{KL}\big(q(z)||p(z)\big). \quad (5)$$

In the following paragraphs, the parameterization of each term in Eq. (5) is described.

Latent States ($\mathcal{S}$) We assume that each data point at a specific time, $x_{n,t}$, belongs to a specific state out of S possible states. This is declared by the categorical variable $s_{n,t}$ in our temporal generative model. These discrete latents, $s_{n,1:T}$, are configured in a Markov chain and govern the state transitions over time as follows (n is dropped hereafter):

$$p_\theta(s_t|s_{t-1}) = \text{Cat}(\boldsymbol{\Phi}_\theta\,\boldsymbol{\pi}_{s_{t-1}}), \quad q_\phi(s_{t-1}) = \text{Cat}(\boldsymbol{\pi}_{s_{t-1}}), \quad (6)$$

where $\boldsymbol{\pi}_{s_{t-1}} = [\pi_1,\cdots,\pi_S]$ is the S-dimensional posterior parameter vector of s_{t-1}, representing probabilities of the categorical distribution, and $\boldsymbol{\Phi}_\theta \in \mathbb{R}^{S\times S}$ is a valid probability transition matrix. In practice, we pass $\boldsymbol{\Phi}_\theta\,\boldsymbol{\pi}_{s_{t-1}}$ from a softmax function to ensure a valid probability vector.

Temporal Latents (W) We adopt a switching Gaussian dynamic for the temporal latent transitions governed by the discrete latent states, s_t. In other words, we assume that the marginal distribution of temporal weights, w_t, follows a Gaussian mixture in the latent space, such that:

$$p_\theta(w_t|w_{t-\ell},s_t=s) = \mathcal{N}\Big(\mu_{\theta_s}^w(w_{t-\ell}),\Sigma_{\theta_s}^w(w_{t-\ell})\Big),$$

where $s\in\{1,\cdots,S\}$, and state-specific $\mu_{\theta_s}^w(\cdot)$ and diagonal $\Sigma_{\theta_s}^w(\cdot)$ are parameterized by multilayer perceptrons (MLPs), hence, follow a *nonlinear* vector auto-regressive model given $w_{t-\ell}$, temporal weights in accordance with a lag set ℓ, as input (e.g., w_{t-1}, w_{t-2}). Namely, we feed $w_{t-\ell}$ to a multi-head MLP for estimating the Gaussian parameters, e.g.,

$$\mu_{\theta_s}^w = \text{FC}_s(h_s), \quad h_s = \sum_{l\in\ell}\sigma(\text{FC}_{s,l}(w_{t-l})),$$

where FC denotes a fully connected layer and σ is a nonlinear activation function. We further combine a *linear* vector auto-regression (VAR) of $w_{t-\ell}$ with the estimated mean from MLP to support both linear and nonlinear dynamics:

$$\mu_{w_t|w_{t-\ell},\,s_t=s} = (\mathbf{1}-\mathbf{g}_s)\odot\text{VAR}_{\theta_s}(w_{t-\ell}) + \mathbf{g}_s\odot\mu_{\theta_s}^w(w_{t-\ell}),$$

where $\odot$ is an element-wise multiplication and $\mathbf{g}_s\in[0,1]$ is a gating vector estimated from $w_{t-\ell}$ using an MLP.

Spatial Factors (F) As with the temporal latents, we assume a diagonal Gaussian distribution for spatial factors parameterized with an MLP as:

$$p_\theta(F|z) = \mathcal{N}\big(\mu_\theta^F(z),\Sigma_\theta^F(z)\big),$$

where z itself is sampled from a normal distribution: $z\sim\mathcal{N}(0,I)$. Introducing z, as a low dimensional spatial embedding, encourages the estimation of a multimodal distribution among spatial factors. Namely, marginalizing $p(F,z) = p(F|z)p(z)$ over z leads to a Gaussian-mixture prior over K factors in F (given the *nonlinear* mapping from z that parameterizes the Gaussian $p(F|z)$). Whereas a matrix Normal prior on F, as in Sun and Chen (2019), naively assumes that $f_{1:K}$ have a unimodal distribution and uncorrelated elements, DSARF is able to encode such correlations by jointly estimating the factors from z.

Expected Log Likelihood Finally, having the temporal weights and spatial factors, we can recover the data by incorporating our initial factorization assumption from Eq. (1):

$$X_n \sim p_\theta(X_n|W_n,F) = \mathcal{N}\Big(\big[w_{n,1},\cdots,w_{n,T}\big]^\top F,\sigma_0^2\Big),$$

where σ_0 is a hyperparameter for observation noise.

Variational Parameters We introduce trainable variational parameters, ϕ, as mean and diagonal covariance of a Gaussian distribution for each data point to define a fully factorized variational distribution on the latents:

$$q(z;\phi^z),\ q(F;\phi^F),\ \Big\{q(w_{n,t};\phi_{n,t}^w)\Big\}_{n=1,\,t=-\ell}^{N,\,T}$$

We approximate variational parameters for discrete latents, $\big\{q(s_{n,t};\phi_{n,t}^s)\big\}_{n=1,\,t=1}^{N,\,T}$, with posteriors $\big\{p(s_{n,t}|w_{n,t})\big\}_{n=1,\,t=1}^{N,\,T}$ to compensate information loss induced by the mean-field approximation:

$$q(s_t;\phi_t^s) \simeq p(s_t|w_t) = \frac{\mathbb{E}_{q(s_{t-1})q(w_{t-\ell})}\Big[p(s_t|s_{t-1})p(w_t|w_{t-\ell},s_t)\Big]}{\sum_{s=1}^{S}\mathbb{E}_{q(s_{t-1})q(w_{t-\ell})}\Big[p(s_t=s|s_{t-1})p(w_t|w_{t-\ell},s_t=s)\Big]} \quad (7)$$

This approximation has a two-fold advantage: (1) spares the model additional trainable parameters for the variational distribution, and (2) further couples together the generative and variational parameters of *discrete* and *continuous* latents, and together with Eq. (6) resolve the *open loop* issue mentioned in Linderman et al. (2017) for these switching models as follows. The posterior on discrete state s_t is informed about the current value of the continuous latent w_t through Eq. (7):

$$q(s_t;\boldsymbol{\pi}_{s_t}) \simeq p_\theta(s_t|w_t) \propto p_\theta(w_t|s_t)p_\theta(s_t),$$
$$w_t \sim q_\phi(w_t),$$

hence, π_{s_t} is a function of w_t, i.e., $\boldsymbol{\pi}_{s_t} = f(w_t;\theta).p_\theta(s_t)$. This is then propagated to the generative model through Eq. (6):

$$p_\theta(s_{t+1}|s_t) = \text{Cat}(\boldsymbol{\Phi}_\theta\boldsymbol{\pi}_{s_t}) = \text{Cat}\left(\boldsymbol{\Phi}_\theta\,f(w_t;\theta).p_\theta(s_t)\right).$$

The latter modulates the prior on latent state, $p(s_{t+1})$, with w_t, whereas the rSLDS models explicitly allow the discrete switches to depend on the continuous latents.

Training DSARF

We compute the Monte-Carlo estimate of the gradient of ELBO in Eq. (5) with respect to generative, θ, and variational,

ϕ, parameters using a re-parameterized sample, (Kingma and Welling 2014), from the posterior of continuous latents, $\{W, z, F\}$. For the discrete latent, $\mathcal{S}$, we compute the expectations over $q_\phi(s_{n,t})$ by summing over the S possible states, hence no explicit sampling is performed, i.e, for each data point $\boldsymbol{\mathcal{L}}_t^{\mathcal{W}}$ would be:

$$-\sum_{s=1}^{S} q(s_t = s)\mathbb{E}_{q(w_{t-\ell})}\Big[\mathrm{KL}\big(q(w_t)\|p\big(w_t|w_{t-\ell}, s_t = s\big)\big)\Big]$$

This explicitly regularizes the S nonlinear auto-regressive priors based on their corresponding weighting. We can analytically calculate the Kullback-Leibler (KL) divergence terms of ELBO for both multivariate Gaussian and categorical distributions, which leads to lower variance gradient estimates and faster training as compared to e.g., noisy Monte Carlo estimates often used in literature. To handle missing entries in the data, we simply drop their corresponding likelihood terms from $\boldsymbol{\mathcal{L}}_n^{\mathbf{rec}}$ in the ELBO (see Eq. (5)).

Implementation Details

We implemented DSARF in PyTorch v1.3 (Paszke et al. 2017) and used the Adam optimizer (Kingma and Ba 2014) with learning rate of 0.01. We initialized all parameters randomly and adopted a linear KL annealing schedule, (Bowman et al. 2016), to increase from 0.01 to 1 over the course of 100 epochs. DSARF has $O(\mathrm{NK}\times(\mathrm{T+D}))$ variational parameters and $O(\mathrm{S}^2+\mathrm{S}|\ell|\mathrm{K}^2)$ parameters for the temporal generative model. The time complexity for training the MLPs is $O(\mathrm{NTS}|\ell|\mathrm{K}^2+\mathrm{KD})$ per epoch and the space complexity is $O(\mathrm{bTD})$ where b is the batch size. We learned and tested all the models on an Intel Core i7 CPU@3.7GHz with 8 GB of RAM. Per-epoch training time varied from 30ms in smaller datasets to 1.2 s in larger experiments and 500 epochs sufficed for most experiments.

Long- & Short-Term Prediction

We evaluated the performance of DSARF for both long- and short-term prediction tasks by adopting a rolling prediction scheme (Yu, Rao, and Dhillon 2016; Chen et al. 2019). For long-term prediction, we predict the test set sequentially using the generative model and spatial factors learned on the train set. For short-term prediction, we predict the next time point on the test set using the generative model and spatial factors learned on the train set: $\hat{X}_{t+1} = \hat{w}_{t+1}^\top F$, where $\hat{w}_{t+1} \sim p(\hat{w}_{t+1}|w_{t+1-\ell}, \hat{s}_{t+1})$, and $\hat{s}_{t+1} \sim p(\hat{s}_{t+1}|s_t)$. We then run inference on X_{t+1}, the actual observation at $t+1$ (if not missing), to obtain w_{t+1} and s_{t+1}, and add them to the historical data for prediction of the next time point $\hat{X}_{t+2}$ in the same way. We repeat these steps to make short-term predictions in a rolling manner across a test set. We keep the generative model and spatial factors fixed during the entire prediction. We report normalized root-mean-square error (NRMSE%) for both long- and short-term predictions. The test set NRMSE% we report for short-term predictions is related to the expected *negative test-set log-likelihood* for our case of Gaussian distributions, hence it is used for evaluating the predictive generative models.

Experimental Evaluation

We evaluated the performance of DSARF in modeling the temporal dynamics and discovering the underlying temporal states by conducting a number of controlled synthetic experiments, followed by a comprehensive real-world data assessment covering a wide range of application areas. To this end, we compared the predictive performance of DSARF against two established Bayesian switching state-space models, rSLDS (Nassar et al. 2019) and SLDS (Fox et al. 2009), three state-of-the-art dynamical matrix factorization methods, BTMF (Sun and Chen 2019), TRMF (Yu, Rao, and Dhillon 2016) and its Bayesian extension (B-TRMF), a deep state-space model, RKN (Becker et al. 2019), and a deep neural network-based time series forecasting method, LSTNet (Lai et al. 2018), which employs vector auto-regression, and allows long-term forecasting, in terms of short- and/or long-term prediction tasks throughout the experiments.

Model Evaluation using Synthetic Data

Toy Example Inspired by Ghahramani and Hinton (2000), we generated $N = 200$ spatio-temporal sequences, each with $T = 200$ time points and $D = 10$ spatial dimensions from $K = 2$ shared factors according to a simple nonlinear dynamical model in W which switched between two temporal models as follows:

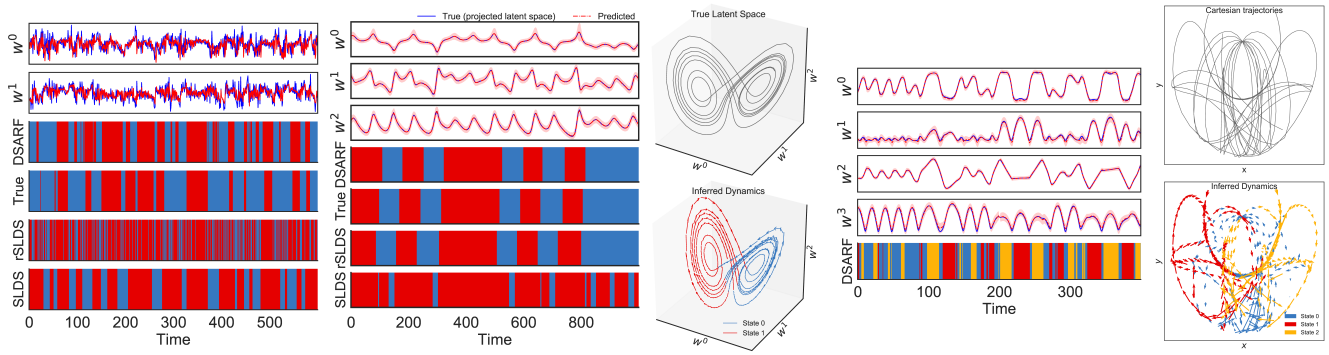
$$w_{t|s_t=0} = 0.9\, w_{t-1} + \tanh(0.5 w_{t-2}) + 3\sin(w_{t-3}) + \epsilon_t^0$$
$$w_{t|s_t=1} = 0.9\, w_{t-1} + \tanh(0.2 w_{t-2}) + \sin(w_{t-3}) + \epsilon_t^1$$
$$X_n = [w_1, \cdots, w_t]_n \,\mathrm{F} + \nu_n \quad \nu_n \sim \mathcal{N}(0, 0.1\,\mathbf{I}),\ \mathrm{F} \sim \mathcal{U}(-1, 1)$$

where $w_t \in \mathbb{R}^2$, $\mathrm{F} \in \mathbb{R}^{2\times 10}$ and $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$, and the switch state s_t was chosen using priors $\pi_1 = \pi_2 = 1/2$ and transition probabilities $\Phi_{11} = \Phi_{22} = 0.95$; $\Phi_{12} = \Phi_{21} = 0.05$. We picked 10 sequences for test, and trained DSARF on the rest with lag set $\ell = \{1, 2, 3\}$ for 200 epochs. We recovered temporal states on the entire dataset with an accuracy of $79.63\% \pm 3.89$ compared to $61.74\% \pm 9.13$ and $52.01\% \pm 11.19$ for rSLDS and SLDS respectively. We predicted the test set in short-term with NRMSE of 13.81% while rSLDS and SLDS only achieved 78.45% and 98.01% respectively. rSLDS and SLDS apparently failed in modeling the higher order temporal dependencies in this synthetic data. We have visualized short-term predictions along with recovered states for three test sequences in Fig. 2a.

Lorenz Attractor We applied DSARF to simulated data from a canonical nonlinear dynamical system, the *Lorenz attractor*, whose nonlinear dynamics are given by:

$$\frac{d\mathbf{w}}{dt} = \begin{bmatrix} \alpha(\mathrm{w}_2 - \mathrm{w}_1) \\ \mathrm{w}_1(\beta - \mathrm{w}_1) - \mathrm{w}_2 \\ \mathrm{w}_1\mathrm{w}_2 - \gamma \mathrm{w}_3 \end{bmatrix}$$

Though nonlinear and chaotic, we see that the Lorenz attractor roughly traces out ellipses in two opposing planes (see Fig. 2b top-right). We simulated $T = 2000$ time points and left the second half for test. Rather than directly observing the states, $\mathrm{w}_{1:T}$, we projected them into a $D = 10$ dimensional space: $X = W^\top F$. Fitting DSARF, we found that the model separates these two planes into two distinct states



(a) Toy example (b) Lorenz attractor (c) Double pendulum

Figure 2: Test set results for synthetic experiments. (a) DSARF with $\ell = \{1, 2, 3\}$ outperforms rSLDS and SLDS in recovering the actual temporal states as these baselines failed in modeling the higher order temporal dependencies in this synthetic data. (b) DSARF separates the two planes in lorenz attractor into two distinct states each with rotational dynamics similar to rSLDS, while SLDS completely failed in this task. (c) DSARF with $\ell = \{1, 2\}$ outperforms baselines in short-term prediction. This is expected as the motion of pendulums are governed by a set of coupled second-order ordinary differential equations. We observe that for $S = 3$ the dynamical trajectory is roughly segmented along the deflection angle of the first pendulum. We have visualized the true (blue) and predicted (red) latent space for all the experiments. Red shaded regions correspond to prediction uncertainty.

(accuracy of 92.90%), each with rotational dynamics as depicted in Fig. 2b bottom-left. SLDS failed in detecting the true states, while rSLDS performed close in terms of state estimation (see Fig. 2b). However, DSARF predicted the test set in short-term with NRMSE of 0.88% compared to 1.14% and 2.70% for rSLDS and SLDS respectively.

Double Pendulum A double pendulum is another simple nonlinear physical system that exhibits rich dynamic behavior with a strong sensitivity to initial conditions. The motion of a double pendulum is governed by a set of coupled second-order ordinary differential equations and is chaotic (Levien and Tan 1993):

$$2\ddot{\theta}_1 + \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) + 2g \sin(\theta_1) = 0$$
$$\ddot{\theta}_2 + \ddot{\theta}_1 \cos(\theta_1 - \theta_2) + \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) + g \sin(\theta_2) = 0,$$

where θ_1 and θ_2 are the deflection angles of the pendulums, and g is the gravitational acceleration. We simulated the system for $T = 20,000$ time points and recorded the locations of the two pendulums. We observed these locations through a linear projection with $D = 10$ just like the previous experiments. We kept the last 400 time points for test, and fit DSARF with $S = 3$ once with lag set $\ell = \{1, 2\}$ and another time with $\ell = \{1\}$ on the train set. DSARF with $\ell = \{1, 2\}$ predicted the test set in short-term with NRMSE of 4.38%, while DSARF with $\ell = \{1\}$ achieved 9.79%, and rSLDS and SLDS achieved 10.42% and 15.53% respectively. This is expected as the second derivatives of location (i.e., acceleration) appear in the Euler-Lagrange differential equation for double pendulum. We have visualized short-term predictions of the test set along with inferred states and dynamics in Fig. 2c. While this system could potentially be segmented to more states, we observed that for $S = 3$ the dynamical trajectory is roughly segmented along the deflection angle of the first pendulum. Increasing $S = 10$ would further improve test set prediction error to 4.16%.

Dataset (missing%)	Resolution	NT×D	T_{test}
Birmingham[2] (14.89)	q30min for 77d	1386×30	126 (7d)
Guangzhou[3] (1.29)	q10min for 61d	8784×214	720 (5d)
Hangzhou[4]	q10min for 25d	2700×80	540 (5d)
Seattle[5]	q5min for 28d	8064×323	1440 (5d)
PST[6]	qmt for 33yr	396×(30×84)	60 (5yr)
Flu[7] (9.53)	qwk for 13yr	658×29	84 (2yr)
Dengue[8] (4.89)	qwk for 13yr	658×10	84 (2yr)
Bat (32.55) in 2015	q33msec	3303×(34×3)	700
Precipitation[9] (50.65)	qd for 5yr	1462×239	305 (1yr)
Apnea in 1994; 2000	q500msec	2000×1	1000

q: every, d: days, yr: years, mt: months, wk: weeks.
The last T_{test} time points are held out for test.

Table 1: Description of real-world datasets.

Model Evaluation using Real-World Data

We give a brief description of each real-world data as well as the train/test splits we used for our evaluation in Table 1. In the following paragraphs, we describe our quantitative comparison results summarized in Table 2 and Table 3.

Short-term prediction results For the sleep Apnea dataset, we observed that DSARF is able to segment the respiration signal for both train and test sets into instances of apnea (small variability in chest volume, followed by bursts)

[2]https://data.birmingham.gov.uk/dataset/birmingham-parking
[3]https://doi.org/10.5281/zenodo.1205229
[4]https://tianchi.aliyun.com/competition/entrance/231708/
[5]https://github.com/zhiyongc/Seattle-Loop-Data
[6]http://iridl.ldeo.columbia.edu/
[7]https://www.google.org/flutrends/about/
[8]https://www.google.org/flutrends/about/
[9]https://www.ncdc.noaa.gov/

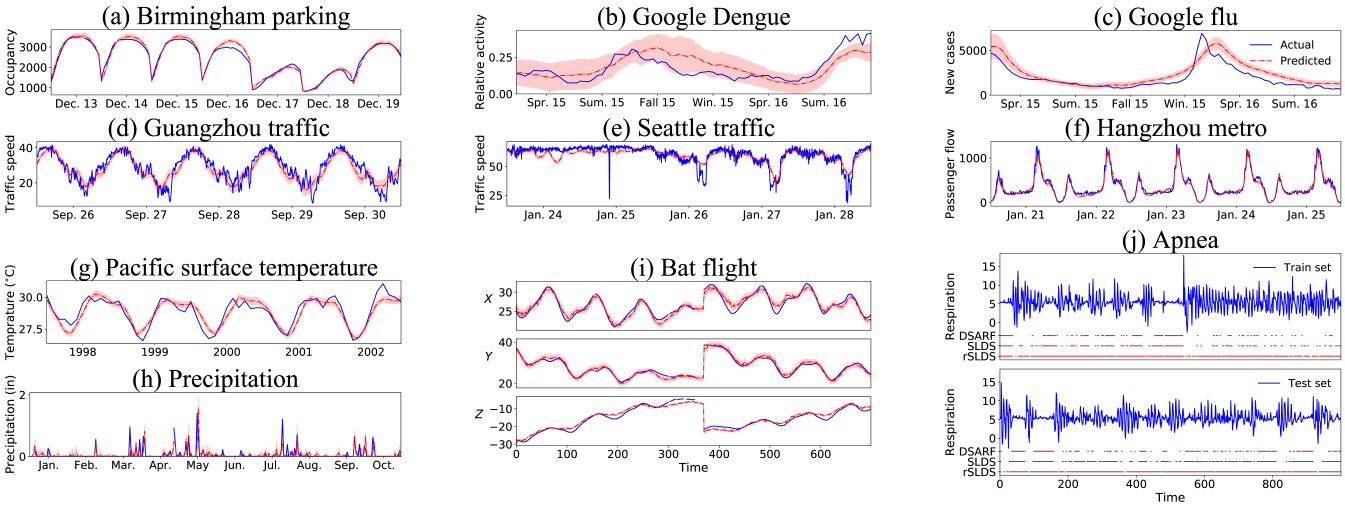


Figure 3: (a)-(g): Long-term predictions of test sets for real data. (h), (i): Short-term predictions of test sets for Precipitation and Bat flight datasets. Red shaded regions correspond to prediction uncertainty. One spatial dimension per dataset is visualized. (j): DSARF segments the respiration signal for both train and test sets into instances of apnea, outperforming SLDS and rSLDS.

Dataset \ Model	DSARF w/ switch	DSARF w/o switch	rSLDS w/ switch	rSLDS w/o switch	SLDS w/ switch	SLDS w/o switch	BTMF	B-TRMF	TRMF	RKN	LSTNet
Birmingham (K=10)	**5.70**	**5.76***	14.23	14.69	8.69	14.52	15.27	15.84	17.13	11.92	9.32
Guangzhou (K=30)	10.21	10.20	**10.10***	10.11	10.86	10.16	10.30	10.75	10.83	10.33	**9.17**
Hangzhou (K=10)	17.31	**15.55**	16.57	17.27	17.20	17.27	16.67	18.26	17.86	**16.39***	16.40
Seattle (K=30)	7.60	**7.52**	7.54	**7.52**	8.44	7.53	7.69	8.10	8.30	7.61	7.73
PST (K=50)	1.96	2.12	1.80	**1.75**	1.94	**1.74***	3.35	23.95	2.81	2.25	2.16
Flu (K=10)	**16.51***	16.77	18.98	18.41	19.50	18.25	22.59	19.01	**15.54**	24.03	17.78
Dengue (K=5)	35.29	**34.67***	43.78	41.48	40.92	39.91	37.23	**33.91**	35.92	37.02	36.39
Bat (K=5)	**7.74**	**8.08***	9.91	10.11	11.13	10.19	8.89	8.16	9.02	18.59	16.55
Precipitation (K=20)	**67.41**	69.52	**67.70***	67.70	68.62	68.44	70.48	96.07	98.01	78.80	74.35
Apnea	**23.86**	–	27.35	–	28.06	–	31.47	30.10	30.08	**27.13***	27.23

The two best results are highlighted in bold fonts without and with asterisk, respectively.

Table 2: Comparison of short-term prediction error (NRMSE%) on the test sets of real data.

with an accuracy of 86% on the test set compared to 71% for SLDS, while rSLDS completely fails on this task as depicted in Fig. 3j. In addition, DSARF predicts the test set in short-term with NRMSE of 23.86% outperforming all the other baselines (see Table 2. RKN with 27.13% is the second best). For the other datasets, as summarized in the short-term section of Table 2, DSARF outperforms in short-term prediction of test sets in Birmangham, Hangzhou, Seattle, Bat flight and precipitation datasets, while closely following the state-of-the-art in Google flu and Dengue datasets (where TRMF and B-TRMF perform the best respectively). LSTNet and rSLDS perform better in Guangzhou and PST datasets respectively, while rSLDS follows DSARF closely in Seattle and Precipitation datasets. We reported the results with (w/) and without (w/o) the switching feature for DSARF, rSLDS and SLDS to explore the impact of these switching latents. We observed that the predictions for Birmingham, Bat, PST, Precipitation and Google flu datasets improved when the switching feature was employed in DSARF. We used lag set $\ell = \{1, 2\}$ for DSARF on all short-term prediction experiments (set accordingly for BTMF, TRMF, B-TRMF, and LSTNet). Sample short-term predictions of test set for precipitation and bat flight data are depicted in Fig. 3h, i respectively.

Long-term prediction results We excluded the Bat flight, Precipitation, and Apnea datasets from long-term prediction task as these datasets hardly show periodic behaviours and/or are chaotic, e.g., in precipitation data (Buizza 2002). On the other hand, we see some extent of long-term recurrence, e.g., in traffic data and seasonal diseases spread, over calendar dates (days, weeks, seasons, etc.). For this reason, we used the lag set $\ell = \{1, 2, 3, T_0, T_0+1, T_0+2, 7T_0, 7T_0+1, 7T_0+2\}$ for traffic datasets as in Sun and Chen (2019) (where T_0 is the time points per day), $\ell = \{1, 2, 52, 52+1, 2\times 52, 2\times 52+1\}$ weeks for Google flu and Dengue datasets and $\ell = \{1, 2, 12, 12+1, 6\times 12, 6\times 12+1\}$ months for the PST dataset. We also excluded SLDS, rSLDS and RKN from this comparison as these baselines do not allow for long historical conditioning, hence are intractable for long-term prediction

Dataset \ Model	DSARF	BTMF	B-TRMF	TRMF	LSTNet
Birmingham (K=10)	**15.05**	18.02	28.71	22.65	23.38
Guangzhou (K=30)	13.01	**12.83**	16.03	14.75	15.76
Hangzhou (K=10)	**15.64**	18.33	20.92	17.85	16.68
Seattle (K=30)	**14.14**	14.33	22.51	16.79	16.30
PST (K=50)	**2.53**	7.43	6.91	3.49	3.17
Flu (K=10)	**34.96**	94.31	51.12	40.87	42.11
Dengue (K=5)	**52.83**	63.85	61.04	57.34	60.46

The best result is highlighted in bold fonts.

Table 3: Comparison of long-term prediction error (NRMSE%) on the test sets of real data.

and diverge very fast. As summarized in the long-term section of Table 3, DSARF outperforms in long-term prediction of the test sets in Birmingham, Hangzhou, Seattle, PST, Google flu and Dengue datasets, while closely follows BTMF on the Guangzhou dataset. We have visualized sample long-term predictions of test set (one spatial dimension per dataset) along with prediction uncertainty and ground-truth values in Fig. 3a-g. Note that part of the error is sourced from the sparse factorization.

Spatial generative model DSARF resulted in spatial factors with higher test-set log-likelihood in all of our real-data experiments when compared to a widely used matrix Normal prior, with -1.02 nats versus -1.37 nats (on average), respectively.

Conclusion

We introduced deep switching auto-regressive factorization (DSARF) in a Bayesian framework. Our method extends switching linear dynamical system models and Bayesian dynamical matrix factorization methods by employing a non-linear vector auto-regressive latent model switched by a Markovian chain of discrete latents to capture higher-order multimodal latent dependencies. This expands prediction horizon and improves long- and short-term forecasting as demonstrated by our extensive synthetic and real data experiments. DSARF proves scalable to high-dimensional data due to the incorporation of factorization framework, is tractable on missing data, provides uncertainty measures for estimations, and lends itself to an efficient inference algorithm.

Ethics Statement

The model we proposed in this work is a step toward better understanding of high dimensional time series data that appear in a variety of real-world settings. Analysing and more importantly *forecasting* these times series naturally embrace a very broad range of applications from healthcare management, disease spread prediction and infection diagnosis to traffic control and weather and financial forecasting, which are where we see the potential for a broader impact. Although, these time series data often show long- and short-term recurring patterns, they occasionally exhibit sophisticated behaviours or are chaotic. Subsequently, we need appropriate tools and legitimate assumptions for analysing them. While we understand that, as George Box wrote in Box (1979), "all models are wrong but some are useful," we hope that in a wide range of applications with proper assumptions and prior knowledge, our DSARF model be useful in providing a means to analyzing high dimensional spatio-temporal data.

References

Ackerson, G.; and Fu, K. 1970. On state estimation in switching environments. *IEEE transactions on automatic control* 15(1): 10–17.

Bahadori, M. T.; Yu, Q. R.; and Liu, Y. 2014. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in neural information processing systems*, 3491–3499.

Becker, P.; Pandya, H.; Gebhardt, G.; Zhao, C.; Taylor, C. J.; and Neumann, G. 2019. Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces. In *International Conference on Machine Learning*, 544–552.

Becker-Ehmck, P.; Peters, J.; and Van Der Smagt, P. 2019. Switching linear dynamics for variational bayes filtering. In *International Conference on Machine Learning*, 553–562. PMLR.

Bergou, A. J.; Swartz, S. M.; Vejdani, H.; Riskin, D. K.; Reimnitz, L.; Taubin, G.; and Breuer, K. S. 2015. Falling with style: bats perform complex aerial rotations by adjusting wing inertia. *PLoS Biol* 13(11): e1002297.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. *CoNLL 2016* 10.

Box, G. E. 1979. Robustness in the strategy of scientific model building. In *Robustness in statistics*, 201–236.

Buizza, R. 2002. Chaos and weather prediction January 2000. *European Centre for Medium-Range Weather Meteorological Training Course Lecture Series ECMWF* .

Cai, Y.; Tong, H.; Fan, W.; Ji, P.; and He, Q. 2015. Facets: Fast comprehensive mining of coevolving high-order time series. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 79–88.

Chang, C.-B.; and Athans, M. 1978. State estimation for discrete systems with switching parameters. *IEEE Transactions on Aerospace and Electronic Systems* (3): 418–425.

Chang, Y.-Y.; Sun, F.-Y.; Wu, Y.-H.; and Lin, S.-D. 2018. A memory-network based solution for multivariate time-series forecasting. *arXiv preprint arXiv:1809.02105* .

Charlin, L.; Ranganath, R.; McInerney, J.; and Blei, D. M. 2015. Dynamic poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, 155–162.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8(1): 1–12.

Chen, X.; He, Z.; Chen, Y.; Lu, Y.; and Wang, J. 2019. Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. *Transportation Research Part C: Emerging Technologies* 104: 66–77.

Farnoosh, A.; Rezaei, B.; Sennesh, E. Z.; Khan, Z.; Dy, J.; Satpute, A.; Hutchinson, J. B.; van de Meent, J.-W.; and Ostadabbas, S. 2020. Deep Markov Spatio-Temporal Factorization. *arXiv preprint arXiv:2003.09779* .

Fox, E.; Sudderth, E. B.; Jordan, M. I.; and Willsky, A. S. 2009. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in neural information processing systems*, 457–464.

Fraccaro, M.; Kamronn, S.; Paquet, U.; and Winther, O. 2017. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, 3601–3610.

Ghahramani, Z.; and Hinton, G. E. 2000. Variational learning for switching state-space models. *Neural computation* 12(4): 831–864.

Ghazi, M. M.; Nielsen, M.; Pai, A. S. U.; Cardoso, M. J.; Modat, M.; Ourselin, S.; and Sorensen, L. 2018. Robust training of recurrent neural networks to handle missing data for disease progression modeling. In *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101(23): e215–e220.

Hamilton, J. D. 1990. Analysis of time series subject to changes in regime. *Journal of econometrics* 45(1-2): 39–70.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *The Journal of Machine Learning Research* 14(1): 1303–1347.

Jing, P.; Su, Y.; Jin, X.; and Zhang, C. 2018. High-order temporal correlation model learning for time-series prediction. *IEEE transactions on cybernetics* 49(6): 2385–2397.

Juloski, A. L.; Weiland, S.; and Heemels, W. 2005. A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control* 50(10): 1520–1533.

Karl, M.; Soelch, M.; Bayer, J.; and van der Smagt, P. 2017. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *stat* 1050: 3.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *stat* 1050: 1.

Krishnan, R. G.; Shalit, U.; and Sontag, D. 2015. Deep Kalman Filters. *stat* 1050: 25.

Krishnan, R. G.; Shalit, U.; and Sontag, D. 2017. Structured inference networks for nonlinear state space models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104.

Levien, R.; and Tan, S. 1993. Double pendulum: An experiment in chaos. *American Journal of Physics* 61(11): 1038–1044.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*.

Linderman, S.; Johnson, M.; Miller, A.; Adams, R.; Blei, D.; and Paninski, L. 2017. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, 914–922.

Mattei, P.-A.; and Frellsen, J. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, 4413–4423.

Murphy, K. 1998. Switching kalman filters. Dept. of Computer Science. *University of California, Berkeley, Tech. Rep* 78.

Nassar, J.; Linderman, S.; Bugallo, M.; and Park, I. 2019. Tree-Structured Recurrent Switching Linear Dynamical Systems for Multi-Scale Modeling. In *International Conference on Learning Representations (ICLR)*.

Paoletti, S.; Juloski, A. L.; Ferrari-Trecate, G.; and Vidal, R. 2007. Identification of hybrid systems a tutorial. *European journal of control* 13(2-3): 242–260.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. *NIPS 2017 Workshop Autodiff Decision* .

Ranganath, R.; Wang, C.; David, B.; and Xing, E. 2013. An adaptive learning rate for stochastic variational inference. In *International Conference on Machine Learning*, 298–306.

Rangapuram, S. S.; Seeger, M. W.; Gasthaus, J.; Stella, L.; Wang, Y.; and Januschowski, T. 2018. Deep state space models for time series forecasting. In *Advances in neural information processing systems*, 7785–7794.

Rezende, D. J.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 1530–1538.

Rigney, D. R. 1994. Multichannel physiological data description and analysis. *Time Series Prediction* .

Rogers, M.; Li, L.; and Russell, S. J. 2013. Multilinear dynamical systems for tensor time series. In *Advances in Neural Information Processing Systems*, 2634–2642.

Salakhutdinov, R.; and Mnih, A. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, 880–887.

Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36(3): 1181–1191.

Sen, R.; Yu, H.-F.; and Dhillon, I. S. 2019. Think globally, act locally: A deep neural network approach to high-dimensional

time series forecasting. In *Advances in Neural Information Processing Systems*, 4837–4846.

Sontag, E. 1981. Nonlinear regulation: The piecewise linear approach. *IEEE Transactions on automatic control* 26(2): 346–358.

Sun, J. Z.; Parthasarathy, D.; and Varshney, K. R. 2014. Collaborative kalman filtering for dynamic matrix factorization. *IEEE Transactions on Signal Processing* 62(14): 3499–3509.

Sun, L.; and Chen, X. 2019. Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *arXiv preprint arXiv:1910.06366* .

Takeuchi, K.; Kashima, H.; and Ueda, N. 2017. Autoregressive tensor factorization for spatio-temporal predictions. In *2017 IEEE International Conference on Data Mining (ICDM)*, 1105–1110. IEEE.

Watter, M.; Springenberg, J.; Boedecker, J.; and Riedmiller, M. 2015. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, 2746–2754.

Wulsin, D. F.; Fox, E. B.; and Litt, B. 2014. Modeling the complex dynamics and changing correlations of epileptic events. *Artificial intelligence* 216: 55–75.

Xiong, L.; Chen, X.; Huang, T.-K.; Schneider, J.; and Carbonell, J. G. 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *2010 SIAM international conference on data mining*, 211–222.

Yu, H.-F.; Rao, N.; and Dhillon, I. S. 2016. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in neural information processing systems*, 847–855.

Zhao, Q.; Zhang, L.; and Cichocki, A. 2015. Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE transactions on pattern analysis and machine intelligence* 37(9): 1751–1763.