Multimodal Teacher Forcing for Reconstructing Nonlinear Dynamical Systems

Anonymous submission

Abstract

Many, if not most, systems of interest in science are naturally described as nonlinear dynamical systems (DS). Empirically, we commonly access these systems through time series measurements, where often we have time series from different types of data modalities simultaneously. For instance, we may have event counts in addition to some continuous signal. While by now there are many powerful machine learning (ML) tools for integrating different data modalities into predictive models, this has rarely been approached so far from the perspective of uncovering the underlying, data-generating DS (aka, DS reconstruction). Recently, sparse teacher forcing (TF) has been suggested as an efficient control-theoretic method for dealing with exploding loss gradients when training ML models on chaotic DS. Here we incorporate this idea into a novel recurrent neural network (RNN) training framework for DS reconstruction based on multimodal variational autoencoders (MVAE). The forcing signal for the RNN is generated by the MVAE which integrates different types of simultaneously given time series data into a joint latent code optimal for DS reconstruction. We show that this training method achieves significantly better reconstructions on multimodal datasets generated from DS benchmarks, like the chaotic Lorenz or Rössler system, than a previously proposed method based on sequential variational autoencoders.

Introduction

For many temporally evolving phenomena in physics, biology, or the social sciences, we have only limited knowledge about the generating dynamical mechanisms. Inferring these from data is a core interest in any scientific discipline. It is also practically highly relevant for predicting important changes in system dynamics, like tipping points in climate systems (Patel and Ott 2022; Bury et al. 2021). In recent years, a variety of methods for recovering dynamical systems (DS) directly and automatically from time series observations has been proposed (Brunton, Proctor, and Kutz 2016; Vlachas et al. 2018; Durstewitz 2017), mostly based on recurrent neural networks (RNNs) for approximating the unknown governing equations of the true DS. However, almost all of these methods assume that observed time series come as continuous signals with Gaussian noise, except for one recent study that also considered non-Gaussian, like categorical or sparse count data for DS reconstruction (Kramer et al. 2022). Time series data from discrete random processes are in fact quite commonplace in many areas, e.g., in the medical domain (electronic health records, Smartphone-based data) (Koppe et al. 2019a), neuroscience (behavior + neural measurements), or climate science (event counts) (Peyre et al. 2020). While it was shown that, in principle, integrating such different measurement modalities can help to improve DS reconstruction, major issues remain. Arguably one of the most challenging aspects is the 'exploding and vanishing gradient problem (EVGP)', a problem that cannot be avoided even in principle when attempting to capture natural systems with chaotic dynamics (Mikhaeil, Monfared, and Durstewitz 2022).

Thus, rather than trying to conquer the EVGP by RNN design, control-theoretic methods based on sparse teacher forcing (TF) were recently suggested to address the EVGP directly in BPTT-based training (Mikhaeil, Monfared, and Durstewitz 2022). The idea here is to guide the training process by pulling the reconstruction algorithm 'back on track' at times determined from the observed system's maximal Lyapunov exponent, i.e. by replacing the RNN's latent states by states inferred from the current observations. While this turned out to be a simple yet powerful remedy for reconstructing systems with chaotic dynamics, outperforming many other state-of-the-art algorithms (Brenner et al. 2022), it remained unclear how to harvest this idea for non-Gaussian observations as it relies on an invertible linear-Gaussian observation model. Hence, one major contribution of the present work is to make the sparse-TF approach amenable to other than continuous Gaussian observations.

Another contribution is a novel formulation of the multimodal data integration problem for DS reconstruction: Building on the success of variational autoencoders (VAEs) for multimodal integration (Baltrušaitis, Ahuja, and Morency 2017; Wu and Goodman 2018; Sutter, Daunhawer, and Vogt 2021), which naturally accommodate different distributional models, we use multimodal VAEs (MVAEs) to construct a common latent representation. However, rather than constructing a sequential VAE process that directly operates on this latent code, as in Kramer et al. (2022), here the purpose of the MVAE is to provide a TF signal that acknowledges all recorded data modalities. Thus, the MVAE learns a variational approximation to a latent distribution that yields the joint distribution over the observed multimodal data, which is then used as a TF signal in the latent space of a simul-

taneously trained RNN. This makes the insights derived from sparse-TF applicable to a fully probabilistic multimodal MVAE training framework, called MVAE-TF in here, which – unlike classical BPTT-TF – also explicitly allows for process (dynamical) noise and yields uncertainty estimates. At the same time, it leaves the distributional assumptions of the individual observed modalities intact.

Related Work

Dynamical Systems Reconstruction In DS reconstruction the goal is to infer or learn a model of the nonlinear DS that produced a set of observed quantities such that invariant temporal and topological properties of the data-generating DS are preserved. This is highly relevant for both science and engineering. In general, nonlinear DS, especially when high-dimensional, noisy, and chaotic, are challenging to learn and difficult to analyze (Champion et al. 2019). RNNs are popular machine learning tools for modeling DS (Park et al. 2022), and various RNN architectures such as Reservoir Computing (RC) (Pathak et al. 2018), Long-Short-Term-Memory networks (LSTMs) (Zheng et al. 2017), or piecewise-linear RNNs (PLRNNs) (Koppe et al. 2019b; Schmidt et al. 2021) have been employed for DS reconstruction. However, with few exceptions (Kramer et al. 2022), almost all of the previously designed methods for DS reconstruction have focused solely on the case where all observations are continuous with Gaussian noise. RNN models trained with classical BPTT suffer from the EVGP (Bengio, Simard, and Frasconi 1994; Pascanu, Mikolov, and Bengio 2013). Recent theoretical work (Mikhaeil, Monfared, and Durstewitz 2022) has shown that this problem is particularly severe when training on time series from chaotic systems (as almost always the case for complex physical or biological systems), as in this case loss gradients inevitably diverge. While many specific RNN architectures have been proposed to deal with the EVGP (Kerg et al. 2019; Kag, Zhang, and Saligrama 2020; Chang et al. 2019; Rusch and Mishra 2021), most of them are not suitable for DS reconstruction as their specific form or parameterization prevents many important DS phenomena (like chaos) by design. Instead, Mikhaeil, Monfared, and Durstewitz (2022) suggested a variant of TF (Williams and Zipser 1989; Pearlmutter 1990), called sparse TF, that strikes an optimal balance between learning relevant long time scales while avoiding too furiously diverging gradients. At present, this method, however, relies on a continuous linear-Gaussian assumption for connecting the data to the latent process.

Generative Models for Multimodal Data Integration Variational generative models are powerful methods for learning latent representations of joint distributions across many data types in an unsupervised fashion. Variational autoencoders (VAE) (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014) are one popular variant which naturally lends itself to multimodal settings (Baltrušaitis, Ahuja, and Morency 2017; Wu and Goodman 2018; Sutter, Daunhawer, and Vogt 2021) and a sequential formulation (Bayer et al. 2021; Girin et al. 2021; Bai, Wang, and Gomes 2021). Longitudinal autoencoders have been proposed (Ramchandran et al. 2021) to model temporal correlations in latent space with Gaussian process priors, and have also been extended to multimodal data (Öğretir et al. 2022). Other generative models for sequential data include state space models that have been applied for posterior inference of latent state paths $\boldsymbol{z}_t \sim p(\boldsymbol{z}_t|\boldsymbol{x}_{1:T})$ of DS given time series observations $\{\boldsymbol{x}_{1:T}\}$ (Pandarinath et al. 2018; Kim et al. 2021; Zhao and Park 2020; Ghahramani and Roweis 1998; Durstewitz 2017). Through their probabilistic formulation, they can account for uncertainty in the model formulation or latent process itself and yield the full distribution over latent space variables (Karl et al. 2017). Many natural and engineered systems are observed through multiple channels with distinct statistical properties simultaneously. For instance, in climate science we may have simultaneous records of temperatures and counts of extreme weather events like tornados. Multimodal data integration can improve model inference and reveal interesting connections between observed modalities (Liang et al. 2015). Multimodal models have also been developed for time series forecasting (Antelmi et al. 2018; Bhagwat et al. 2018; Dezfouli et al. 2018; Shi et al. 2021; Sutter, Daunhawer, and Vogt 2021). DS reconstruction, however, goes beyond forecasting in that we also require an approximation to the governing equations which captures the temporal and geometrical structure of the original system. Such a model enables further analysis and mechanistic insight into the underlying dynamics (Strogatz 2018). Kramer et al. (2022) recently proposed a nonlinear state space model embedded within a sequential VAE for reconstructing DS from multimodal time series data. Their work demonstrates the advantages of exploiting various data modalities simultaneously for DS reconstruction, including non-Gaussian like categorical series. Here we develop a more efficient approach that takes into account recent insights on training RNNs for DS reconstruction (Mikhaeil, Monfared, and Durstewitz 2022).

Method

Our approach to DS reconstruction from multimodal time series rests on three components: 1) A specific type of dynamically interpretable and mathematically tractable RNN, the recently introduced 'dendritic PLRNN (dendPLRNN)' (Brenner et al. 2022); 2) a specific TF algorithm, sparse identity-TF, for guiding the training process (Brenner et al. 2022; Mikhaeil, Monfared, and Durstewitz 2022); 3) a MVAE for producing a multimodal TF signal, trained jointly with the dendPLRNN for dynamics reconstruction through a combined loss. The whole procedure is illustated in Fig. 1.

dendPLRNN The dendPLRNN (Brenner et al. 2022) used for DS reconstruction is defined by the M-dimensional latent process equation

$$\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}\phi(\boldsymbol{z}_{t-1}) + \boldsymbol{h} + \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \quad (1)$$

which describes the temporal evolution of M-dimensional latent state vector $\boldsymbol{z}_t = (z_{1t} \dots z_{Mt})^T$, with a linear diagonal matrix term $\boldsymbol{A} \in \mathbb{R}^{M\times M}$, off-diagonal matrix $\boldsymbol{W} \in \mathbb{R}^{M\times M}$, and diagonal noise covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$. The nonlinearity is given by the 'dendritic' spline expansion

$$\phi(\boldsymbol{z}_{t-1}) = \sum_{b=1}^{B} \alpha_b \max(0, \boldsymbol{z}_{t-1} - \boldsymbol{h}_b), \quad (2)$$

with slopes $\alpha_b \in \mathbb{R}$ and thresholds $\boldsymbol{h}_b \in \mathbb{R}^M$ (Brenner et al. 2022). This formulation retains analytical access to fixed points and cycles (Durstewitz 2017) and allows for a transformation into a continuous-time representation (Monfared and Durstewitz 2020). To infer the latent process equation jointly from multiple data modalities, the dendPLRNN is connected to different decoder models that take the distinct distributional properties of each modality into account (see next section and Appendix). For normally distributed data, for example, this may take the simple linear Gaussian form

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\eta}_t, \tag{3}$$

with factor loading matrix $\boldsymbol{B} \in \mathbb{R}^{N\times M}$, and Gaussian observation noise $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$ with diagonal covariance $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$.

Identity Teacher Forcing Without any purpose-tailoring of the RNN architecture for dealing with exploding gradients, sparse TF (Mikhaeil, Monfared, and Durstewitz 2022), a variant of TF (Williams and Zipser 1989; Pearlmutter 1990), can minimize the problem when training on chaotic systems. It does so by balancing loss and trajectory divergence with the need to capture relevant long time scales. To apply sparse TF to the latent space, the observation model needs to be inverted. Most simply, this is achieved if we have a linear-Gaussian model as Eq. 3 and choose an "identity-mapping" for the observation model $\hat{\boldsymbol{x}}_{\boldsymbol{t}} = \mathcal{I}\boldsymbol{z}_{\boldsymbol{t}}$, where $\mathcal{I} \in \mathbb{R}^{N\times M}$ with $\mathcal{I}_{kk} = 1$ if $k \leq N$ and zeroes everywhere else.[1]

Consider an observed time series $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_T}\}$ generated by a DS we want to reconstruct. For times $l\tau + 1,\ l \in \mathbb{N}_0$, with forcing interval $\tau \geq 1$, we replace the first N latent states by observations $\hat{z}_{k,l\tau+1} = x_{k,l\tau+1},\ k \leq N$. The observations are thereby mapped one-to-one onto a subset of latent 'readout' states, while the remaining latent states, $\hat{z}_{k,l\tau+1} = z_{k,l\tau+1},\ k > N$, remain unaffected. Replacing latent states with observations significantly helps to stabilize training while allowing unimpeded gradient flow through the non-forced states. As shown in Mikhaeil, Monfared, and Durstewitz (2022), the best tradeoff between exploding gradients and capturing relevant long-term dependencies is achieved when choosing the forcing interval τ according to the system's maximal Lyapunov exponent. However, for non-continuous data (like counts), this cannot readily be determined. In this case the forcing interval τ may also be considered as a hyperparameter, and optimal settings found via grid search. With $\mathcal{F} = \{l\tau + 1\}_{l\in\mathbb{N}_0}$, the dendPLRNN updates can then be written as

$$\boldsymbol{z_{t+1}} = \begin{cases} dendPLRNN(\tilde{\boldsymbol{z}}_{\boldsymbol{t}}) & \text{if } t \in \mathcal{F} \\ dendPLRNN(\boldsymbol{z_t}) & \text{else} \end{cases}. \tag{4}$$

The squared error loss used within the BPTT-TF algorithm is given by $\mathcal{L}_{MSE} = \sum_{t=2}^{T} \|\boldsymbol{x_t} - \mathcal{I}\boldsymbol{z_t}\|_2^2$ and is calculated prior to forcing for every time step. When sampling from the trained model, the dendPLRNN runs freely without forcing.

[1]Note that by 'invertible' we are referring to the pseudo-inverse here.

Multimodal Variational Autoencoder (MVAE) While our model formulation is general and can work with any combination of data modalities, for the present exposition we consider time series of multivariate Gaussian, ordinal, and count nature of length T, $\tilde{\boldsymbol{X}} = \{\{\boldsymbol{x_1}, \cdots, \boldsymbol{x_T}\}; \{\boldsymbol{o_1}, \cdots, \boldsymbol{o_T}\}; \{\boldsymbol{c_1}, \cdots, \boldsymbol{c_T}\}\}$. We employ an MVAE for inferring a joint latent representation over these data. We denote encoded states at time t by $\tilde{\boldsymbol{z}}_{\boldsymbol{t}} \in \mathbb{R}^K$ to avoid confusion with the latent dynamical process $\boldsymbol{z_t} \in \mathbb{R}^M$. To this end, we maximize the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta, \phi; \tilde{\boldsymbol{x}}) := \mathbb{E}_{q_\phi}[\log(p_\theta(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{z}})] - \mathrm{KL}[q_\phi(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})||p_\theta(\tilde{\boldsymbol{z}})] \tag{5}$$

using the reparameterization trick (Kingma and Welling 2014). For the encoding step, we concatenate all input modalities and use convolutional neural networks (CNNs) for parameterizing the encoder model $q_\phi(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})$, with hyperparameters as proposed in Brenner et al. (2022). This allows the recognition model to embed context into the latent representation (see e.g. Cui, Chen, and Chen (2016)). As the latent dendPLRNN process itself is conditionally Gaussian, we make a Gaussian assumption for the approximate posterior $q_\phi(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\tilde{\boldsymbol{x}}), \boldsymbol{\Sigma}_\phi(\tilde{\boldsymbol{x}}))$ to estimate the true unknown posterior $p(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})$, where mean and covariance are functions of the data. We further use the common mean field approximation to factorize $q_\phi(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})$ across time. For the decoding step, the first K states of the latent model are connected to the output via modality-specific observation models.

Observations are assumed to be conditionally independent given the latent state $\tilde{\boldsymbol{z}}_{\boldsymbol{t}}$. A linear decoder as in Eq. 3 is employed for Gaussian data, a cumulative link model for ordinal data, and a log-link function for Poisson data (but recall the present framework can deal with any set of distributional models):

$$\boldsymbol{x_t} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{t}} \sim \mathcal{N}(\boldsymbol{B}\tilde{\boldsymbol{z}}_{\boldsymbol{t}}, \boldsymbol{\Gamma}) \tag{6}$$

$$\boldsymbol{o_t} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{t}} \sim \mathrm{Ordinal}(\boldsymbol{\beta}\tilde{\boldsymbol{z}}_{\boldsymbol{t}}, \epsilon) \tag{7}$$

$$\boldsymbol{p_t} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{t}} \sim \mathrm{Poisson}(\lambda) \tag{8}$$

Due to the conditional independence given the latent states, the likelihood terms related to the observations are simply summed, $\mathcal{L}_{rec} = \sum_{t=1}^{T}(\log p(\boldsymbol{x_t}|\tilde{\boldsymbol{z}}_{\boldsymbol{t}}) + \log p(\boldsymbol{o_t}|\tilde{\boldsymbol{z}}_{\boldsymbol{t}}) + \log p(\boldsymbol{c_t}|\tilde{\boldsymbol{z}}_{\boldsymbol{t}}))$.

Multimodal Teacher Forcing The MVAE provides a flexible and convenient framework for mapping any combination of observed data modalities into the process model's latent space, and is employed here to generate a TF signal for the dendPLRNN. We call this the MVAE-TF approach. For training, observed time series $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_{\boldsymbol{1}}, \cdots, \tilde{\boldsymbol{x}}_{\boldsymbol{T}}\}$ from possibly many different and non-Gaussian modalities are first converted to a common latent code $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_{\boldsymbol{1}}, \cdots, \tilde{\boldsymbol{z}}_{\boldsymbol{T}}\}$ using the MVAE. The initial condition $\boldsymbol{z_1}$ of the latent process is inferred from the encoded state $\tilde{\boldsymbol{z}}_{\boldsymbol{1}}$. If $K < M$, the $M - K$ remaining states are randomly sampled from a uniform normal distribution. The dendPLRNN is then iterated forward from $\boldsymbol{z_1}$ across the length of the observed time series T to obtain a latent path $\boldsymbol{Z} = \{\boldsymbol{z_1}, \boldsymbol{z_2}, \cdots, \boldsymbol{z_T}\}$, while in analogy to identity-TF, at times $l\tau+1,\ l \in \mathbb{N}_0$ the first K latent states are replaced by the encoded states $z_{k,l\tau+1} = \tilde{z}_{k,l\tau+1},\ k \leq K$.

The first K states of the predicted latent trajectory $\boldsymbol{Z}$ of length T are then used to compute the modality-specific likelihoods for the observed multimodal time series $\tilde{\boldsymbol{X}}$, giving $\mathcal{L}_{dec}$. The general scheme for the MVAE-TF is visualised in Fig. 1. Note that the MVAE-TF does not require that encoder dimension K and latent process dimension M are equal. Preliminary experiments have shown that leaving a subset of the latent states unforced leads to better results, which we conjecture is due to the unrestricted gradient flow through these states across the whole length T of a training sequence. The second term in the ELBO, Eq. 5, quantifies the consistency between encoded samples $\tilde{\boldsymbol{Z}}$ and the first K latent states of the trajectory drawn from the model prior $\boldsymbol{Z} \sim p_\theta(\boldsymbol{Z})$, parameterized by the dendPLRNN, and is given by

$$\mathcal{L}_{lat} = \sum_{i=1}^{K} \sum_{t=2}^{T} \left(-\frac{1}{2} \left(z_{ti} - \tilde{z}_{ti} \right)^2 / \sigma_{ii}^2 \right). \tag{9}$$

The total loss is given by the reconstruction loss of the MVAE (first term in Eq. 5), decoded from $\tilde{\boldsymbol{Z}}$, the decoder loss from the likelihoods of the observed time series $\tilde{\boldsymbol{X}}$ given the first K states of predicted latent path $\boldsymbol{Z}$, and the latent loss Eq. 9, computed between encoder states $\tilde{\boldsymbol{Z}}$ and latent path $\boldsymbol{Z}$.

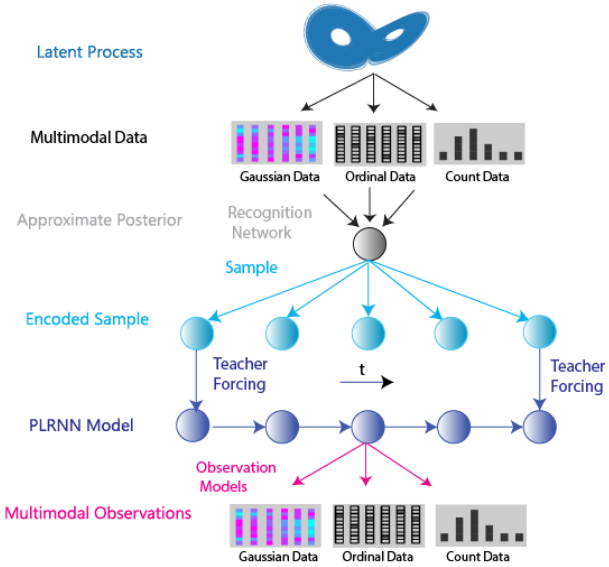

Figure 1: MVAE-TF setup. Multimodal observations are translated via an encoder into a common latent representation, which is used for sparse TF in the dendPLRNN's latent space. The latent trajectory is then mapped back into the multimodal observation space via modality-specific observation (decoder) models.

Experiments

Benchmark Datasets We sample a training set and a test set of $100,000$ time steps from a Lorenz-63 system ($dt = 0.05, \rho = 28.0, \sigma = 10.0, \beta = 8.0/3.0$) and a Rössler system ($dt = 0.1, a = 0.2, b = 0.2, c = 5.7$) in

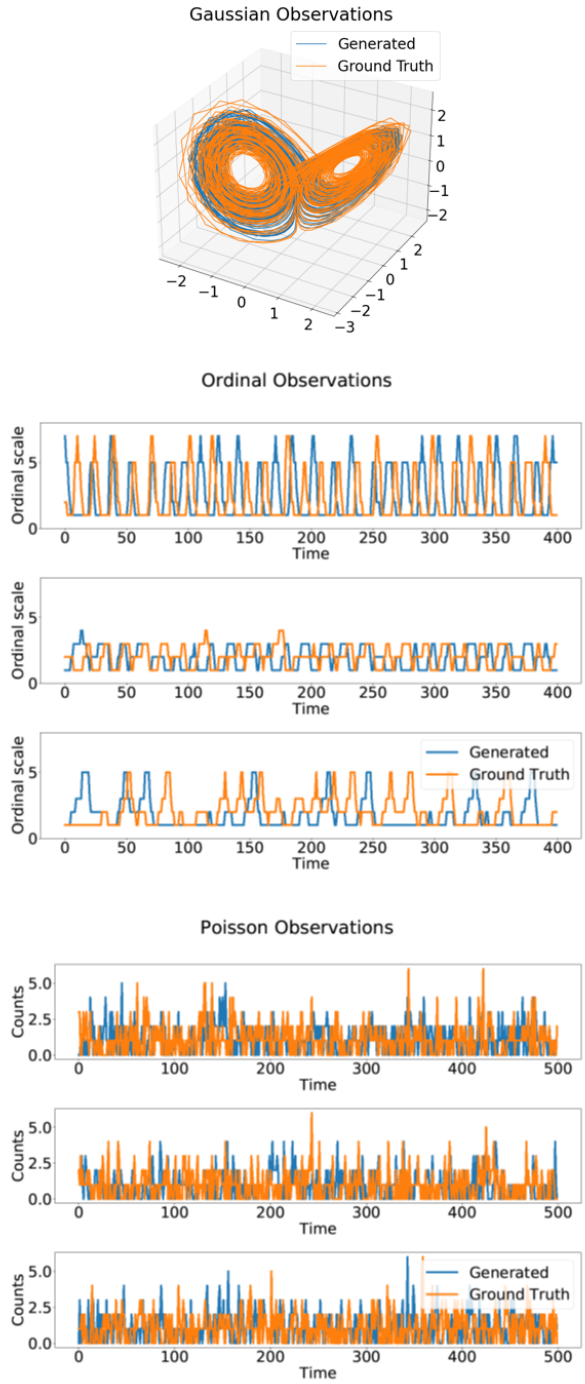

Figure 2: Freely generated example trajectories from a dendPLRNN ($M = 20, K = 15, \tau = 10, B = 10$), trained with MVAE-TF jointly on normal, ordinal, and count data sampled from a ground-truth Lorenz-63 system. Simulated latent trajectories faithfully capture the geometry of the Lorenz attractor, as well as the temporal structure of the ground truth data when projected back into observation space.

Dataset	Method	$D_{stsp}\downarrow$	PSC $\uparrow$	PE $\downarrow$	OPE $\downarrow$	SCE $\downarrow$	OAC $\downarrow$	CAC $\downarrow$
Lorenz	MVAE-TF	1.37 ± 0.14	0.988 ± 0.001	$1.2\text{e}{-2} \pm 1.1\text{e}{-3}$	0.056 ± 0.01	0.023 ± 0.002	0.011 ± 0.002	$5.6\text{e}{-5} \pm 2.4\text{e}{-6}$
	SVAE	6.7 ± 0.7	0.89 ± 0.03	$3.4\text{e}{-1} \pm 2.1\text{e}{-2}$	0.46 ± 0.03	0.14 ± 0.01	0.13 ± 0.03	$9.5\text{e}{-5} \pm 1.2\text{e}{-5}$
Rössler	MVAE-TF	2.43 ± 0.5	0.93 ± 0.01	$2.4\text{e}{-3} \pm 6.1\text{e}{-4}$	0.04 ± 0.016	0.03 ± 0.027	0.006 ± 0.001	$6.4\text{e}{-5} \pm 5.1\text{e}{-6}$
	SVAE	9.7 ± 1.5	0.83 ± 0.01	$1.2\text{e}{-1} \pm 3.4\text{e}{-2}$	0.20 ± 0.06	0.12 ± 0.06	0.10 ± 0.03	$1.1\text{e}{-4} \pm 2.3\text{e}{-5}$

Table 1: Comparison of dendPLRNN (Ours) trained by the MVAE-TF ($M = 20, K = 15, \tau = 15$) and with a SVAE, based on (Kramer et al. 2022), on multivariate normal, ordinal and count data, derived from the Lorenz and Rössler attractor. Values are mean $\pm$ SEM, averaged over 10 models.

their chaotic regimes, using `scipy.odeint`, including 1% process noise (more details are given in the Appendix). From these, ordinal and count observations are obtained using Eqs. 16 and 18 with randomly initialized parameters, and 1% observation noise is added for the Gaussian observations.

Performance Measures In DS reconstruction, we are primarily interested in capturing *invariant* properties of the underlying DS like its geometrical and asymptotic temporal structure. While mean-squared errors (MSE) may be used to asses short-term ahead prediction, especially in chaotic systems they are not suited for evaluating the system's longer-term behavior (Koppe et al. 2019b; Mikhaeil, Monfared, and Durstewitz 2022). To assess the quality of reconstructions, we compile a set of general as well as modality-specific performance measures (for more details, see Appendix):

- To assess overlap in *attractor geometries*, we employ a Kullback-Leibler divergence in state space (Koppe et al. 2019b; Brenner et al. 2022).
- To assess the agreement in asymptotic *temporal structure*, in the *continuous-Gaussian* case power spectra were first computed through the Fast Fourier Transform (Cooley and Tukey 1965) on all observed time series dimensions, slightly smoothed with Gaussian kernels to remove noise, and correlated with those generated by the model. To assess the agreement in temporal structure for *ordinal* and *count data*, we reverted to the auto-correlation function (Wiener 1930) computed for up to 200 time lags based on the Spearman rank correlation coefficient for discrete ordinal data (see Appx. Fig. 5, OAC for ordinal and CAC for count data in Table 1).
- To assess *short-term behavior*, for *Gaussian* data the classical 10-step-ahead PE along test set trajectories was used. For *ordinal data*, we computed a linear (L1) prediction error (OPE) instead, as suggested in Öğretir et al. (2022).[2]

Multimodal Data We first evaluate the MVAE-TF's ability to combine information from different observed data modalities (Gaussian, ordinal, and count data) for inferring a common latent nonlinear DS model on two benchmark datasets. Example reconstructions of the MVAE-TF are displayed in Fig. 2. For comparison, the same dendPLRNN was trained as proposed in Kramer et al. (2022) with a sequential VAE, optimizing its multimodal ELBO. This to our knowledge is currently the only other approach specifically designed for DS reconstruction from various data modalities observed simultaneously. Results are given in Table 1.

[2]Alternatively, one may use a χ^2 measure to assess agreement in counts.

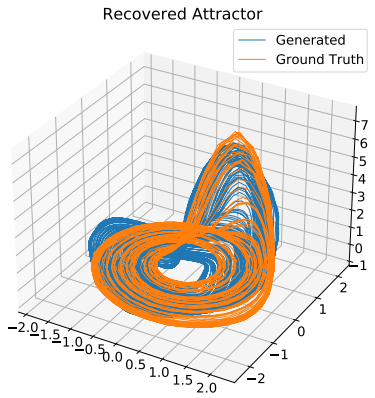

Figure 3: Reconstructed Roessler attractor, recovered solely from an 8-dimensional set of ordinal observations with 7 categories each ($M = 20, K = 15, \tau = 15, B = 10$). The figure was obtained by freely sampling a trajectory in the latent space of the reconstructed system and then applying a geometry-preserving linear transformation, yielding $D_{stsp} = 3.22$ for this example.

Ordinal Data As another test, we probed the reconstruction from an *eight dimensional ordinal time series*, using seven ordered categories (i.e., $o_{nt} \in \{1 \dots 7\}$ for each dimension n). This is profoundly more challenging than the multimodal setting with Gaussian data, since the ordinal process considerably coarse-grains the underlying continuous dynamical process. Since in this case we do not have a direct linear mapping between the ground truth state space and the latent model's one (which in the case of Gaussian observations is simply given by Eq. 3), we construct one post-hoc by optimizing a linear operator (given by a linear dimensionality reduction concatenated with a rotation operation) to find a mapping between true and latent spaces that only allows for transformations that preserve geometrical congruency (see Fig. 3). Table 2 shows that our approach also clearly outperforms the SVAE in this setting.

Conclusions

In the present work we introduced a novel training method for DS reconstruction from multimodal time series data based on dynamically interpretable RNNs. While DS reconstruction is meanwhile a large field in scientific ML (Brunton and Kutz

Dataset	Method	$D_{stsp}\downarrow$	OPE $\downarrow$	SCE $\downarrow$	OAC $\downarrow$
Lorenz	MVAE-TF	8.8 ± 0.59	0.24 ± 0.015	0.085 ± 0.02	0.016 ± 0.04
	SVAE	14.7 ± 0.7	0.8 ± 0.03	0.17 ± 0.02	0.23 ± 0.02
Rössler	MVAE-TF	7.9 ± 0.8	0.093 ± 0.007	0.051 ± 0.009	0.051 ± 0.009
	SVAE	11.5 ± 1.3	0.39 ± 0.02	0.23 ± 0.05	0.18 ± 0.04

Table 2: Comparison of dendPLRNN (Ours) trained by MVAE-TF ($M = 20, K = 15, \tau = 15$), and with a SVAE (Kramer et al. 2022) with $M = 20$, on 8 randomly sampled ordinal observations with seven ordered categories, derived from the Lorenz and Rössler attractor. As a approximation for the agreement in state space geometry, we employ a measure based on the first three principal components of the sampled latent trajectory, compared to the ground-truth attractor (see Appendix). Values are mean $\pm$ SEM, averaged over 10 models.

2022), reconstruction based on multimodal, especially non-continuous/ non-Gaussian, data has hardly been addressed so far, although such scenarios are commonplace in many areas like medicine, neuroscience, or climate research. Here we utilize recent insights on guiding the training process by control signals (Mikhaeil, Monfared, and Durstewitz 2022) within a novel multimodal data integration framework for DS reconstruction. In our approach, a sparse TF signal is generated by an MVAE that translates many different data modalities into a common latent code. This yields a flexible inference framework for recovering DS from multimodal data while avoiding common training issues associated with chaotic systems (Mikhaeil, Monfared, and Durstewitz 2022). It further leaves the underlying statistical assumptions of the observed data intact by retaining modality-specific observation models. We show that on two common benchmark systems, and assessed through a number of different performance measures, training the dendPLRNN with the MVAE-TF approach clearly outperformed a SVAE-based approach previously suggested by Kramer et al. (2022). We conjecture that this is due to the fact that MVAE-TF allows latent trajectories to evolve freely for longer time spans during training. In contrast, in classical SVAEs temporal consistency is ensured only through the 'same-time' terms in the ELBO. Moreover, the MVAE-TF algorithm was able to recover the underlying attractor based on ordinal observations alone. That attractor geometries can be faithfully reconstructed from discrete random variables alone to our knowledge has indeed never been shown before. A further advantage of our method is that it is modular, that is, subcomponents of the algorithm, such as the encoder or latent model, can easily be replaced within the same overall training framework. More sophisticated encoder models, e.g. combinations of mixtures-of-experts or products-of-experts (Wu and Goodman 2018; Shi et al. 2019), could be used to find more effective embeddings of multimodal data. Optimal weighing schemes for the different loss terms making up the total loss (Bakarji et al. 2022) may further improve performance. While here we presented first results on DS reconstruction for discrete data, whether and how much of the original state space topology of a data-generating DS can be recovered from non-continuous, non-Gaussian random variables remains an important topic for future theoretical and empirical research.

References

Antelmi, L.; Ayache, N.; Robert, P.; and Lorenzi, M. 2018. Multi-channel Stochastic Variational Inference for the Joint Analysis of Heterogeneous Biomedical Data in Alzheimer's Disease. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, 15–23. Cham: Springer International Publishing. ISBN 978-3-030-02628-8.

Bai, J.; Wang, W.; and Gomes, C. P. 2021. Contrastively Disentangled Sequential Variational Autoencoder. *Advances in Neural Information Processing Systems*, 34: 10105–10118.

Bakarji, J.; Champion, K.; Kutz, J. N.; and Brunton, S. L. 2022. Discovering Governing Equations from Partial Measurements with Deep Delay Autoencoders. arXiv:2201.05136 [cs, math].

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv:1705.09406 [cs]*.

Bayer, J.; Soelch, M.; Mirchev, A.; Kayalibay, B.; and van der Smagt, P. 2021. Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models. *arXiv:2101.07046 [cs, stat]*.

Bengio, Y.; Simard, P.; and Frasconi, P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2): 157–166.

Bhagwat, N.; Viviano, J. D.; Voineskos, A. N.; Chakravarty, M. M.; and Initiative, A. D. N. 2018. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLOS Computational Biology*, 14(9): e1006376. Publisher: Public Library of Science.

Brenner, M.; Hess, F.; Mikhaeil, J. M.; Bereska, L. F.; Monfared, Z.; Kuo, P.-C.; and Durstewitz, D. 2022. Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2292–2320. PMLR.

Brunton, S. L.; and Kutz, J. N. 2022. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press. ISBN 978-1-00-909848-9. Google-Books-ID: rxNkEAAAQBAJ.

Brunton, S. L.; Proctor, J. L.; and Kutz, J. N. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc Natl Acad Sci U S A*, 113(15): 3932–3937.

Bury, T. M.; Sujith, R. I.; Pavithran, I.; Scheffer, M.; Lenton, T. M.; Anand, M.; and Bauch, C. T. 2021. Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, 118(39): e2106140118. Publisher: Proceedings of the National Academy of Sciences.

Champion, K.; Lusch, B.; Kutz, J. N.; and Brunton, S. L. 2019. Data-driven discovery of coordinates and governing equations. *Proc Natl Acad Sci USA*, 116(45): 22445–22451.

Chang, B.; Chen, M.; Haber, E.; and Chi, E. H. 2019. AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks. *International Conference on Learning Representations*.

Cooley, J. W.; and Tukey, J. W. 1965. An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 19(90): 297–301. Publisher: American Mathematical Society.

Cui, Z.; Chen, W.; and Chen, Y. 2016. Multi-Scale Convolutional Neural Networks for Time Series Classification. *Computing Research Repository*, abs/1603.06995.

Dezfouli, A.; Morris, R.; Ramos, F. T.; Dayan, P.; and Balleine, B. 2018. Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Durstewitz, D. 2017. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLoS Comput. Biol.*, 13(6): e1005542.

Ghahramani, Z.; and Roweis, S. T. 1998. Learning nonlinear dynamical systems using an EM algorithm. In *Advances in Neural Information Processing Systems 11*.

Girin, L.; Leglaive, S.; Bie, X.; Diard, J.; Hueber, T.; and Alameda-Pineda, X. 2021. Dynamical Variational Autoencoders: A Comprehensive Review. *Foundations and Trends® in Machine Learning*, 15(1): 1–175.

Kag, A.; Zhang, Z.; and Saligrama, V. 2020. RNNs Incrementally Evolving on an Equilibrium Manifold: A Panacea for Vanishing and Exploding Gradients? *International Conference on Learning Representations*, 24.

Kantz, H.; and Schreiber, T. 2004. *Nonlinear time series analysis*, volume 7. Cambridge university press.

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. In *Proceedings of the 5th International Conference on Learning Representations*.

Kerg, G.; Goyette, K.; Touzel, M. P.; Gidel, G.; Vorontsov, E.; Bengio, Y.; and Lajoie, G. 2019. Non-normal Recurrent Neural Network (nnRNN): learning long time dependencies while improving expressivity with transient dynamics. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 11.

Kim, T. D.; Luo, T. Z.; Pillow, J. W.; and Brody, C. 2021. Inferring Latent Dynamics Underlying Neural Population Activity via Neural Differential Equations. In *International Conference on Machine Learning*, 5551–5561. PMLR. ISSN: 2640-3498.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Koppe, G.; Guloksuz, S.; Reininghaus, U.; and Durstewitz, D. 2019a. Recurrent Neural Networks in Mobile Sampling and Intervention. *Schizophrenia Bulletin*, 45(2): 272–276.

Koppe, G.; Toutounji, H.; Kirsch, P.; Lis, S.; and Durstewitz, D. 2019b. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLOS Computational Biology*, 15(8): e1007263.

Kramer, D.; Bommer, P. L.; Tombolini, C.; Koppe, G.; and Durstewitz, D. 2022. Reconstructing Nonlinear Dynamical Systems from Multi-Modal Time Series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 11613–11633. PMLR.

Lambert, D. 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1): 1–14. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Liang, M.; Li, Z.; Chen, T.; and Zeng, J. 2015. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4): 928–937.

Liddell, T. M.; and Kruschke, J. K. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79: 328–348.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140: 55–55.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *Proceedings of the Eighth International Conference on Learning Representations*.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2): 130–141.

McCullagh, P. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2): 109–142. Publisher: [Royal Statistical Society, Wiley].

Mikhaeil, J. M.; Monfared, Z.; and Durstewitz, D. 2022. On the difficulty of learning chaotic dynamics with RNNs. *arxiv*.

Monfared, Z.; and Durstewitz, D. 2020. Transformation of ReLU-based recurrent neural networks from discrete-time to continuous-time. In *Proceedings of the 37th International Conference on Machine Learning*.

Pandarinath, C.; O'Shea, D. J.; Collins, J.; Jozefowicz, R.; Stavisky, S. D.; Kao, J. C.; Trautmann, E. M.; Kaufman, M. T.; Ryu, S. I.; Hochberg, L. R.; Henderson, J. M.; Shenoy, K. V.; Abbott, L. F.; and Sussillo, D. 2018. Inferring single-trial neural population dynamics using sequential autoencoders. *Nature Methods*, 15(10): 805–815.

Park, Y.; Gajamannage, K.; Jayathilake, D. I.; and Bollt, E. M. 2022. Recurrent Neural Networks for Dynamical Systems: Applications to Ordinary Differential Equations, Collective Motion, and Hydrological Modeling. *arXiv:2202.07022 [cs, math]*.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*.

Patel, D.; and Ott, E. 2022. Using Machine Learning to Anticipate Tipping Points and Extrapolate to Post-Tipping Dynamics of Non-Stationary Dynamical Systems. arxiv:2207.00521 [physics].

Pathak, J.; Hunt, B.; Girvan, M.; Lu, Z.; and Ott, E. 2018. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Phys. Rev. Lett.*, 120(2): 024102.

Pearlmutter, B. 1990. Dynamic recurrent neural networks.

Peyre, M.; Zhang, Y.; Huc, M.; Roger, F.; and Kerr, Y. H. 2020. Chaos theory applied to the outbreak of COVID-19: an ancillary approach to decision making in pandemic context. *Epidemiology and Infection*, 148. Publisher: Cambridge University Press (CUP).

Ramchandran, S.; Tikhonov, G.; Kujanpaa, K.; Koskinen, M.; and Lahdesmaki, H. 2021. Longitudinal Variational Autoencoder. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 11.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*.

Rusch, T. K.; and Mishra, S. 2021. Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*.

Rössler, O. E. 1976. An equation for continuous chaos. *Physics Letters A*, 57(5): 397–398.

Schmidt, D.; Koppe, G.; Monfared, Z.; Beutelspacher, M.; and Durstewitz, D. 2021. Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies. In *Proceedings of the 9th International Conference on Learning Representations*.

Shi, Y.; Paige, B.; Torr, P. H. S.; and Siddharth, N. 2021. Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models. *arXiv:2007.01179 [cs, stat]*.

Shi, Y.; Siddharth, N.; Paige, B.; and Torr, P. H. S. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1408, 15718–15729. Curran Associates Inc.

Strogatz, S. H. 2018. *Nonlinear Dynamics and Chaos*. CRC Press, 1 edition. ISBN 978-0-429-96111-3.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized Multimodal ELBO. *arXiv:2105.02470 [cs, stat]*.

Talathi, S. S.; and Vartak, A. 2016. Improving performance of recurrent neural network with relu nonlinearity. In *Proceedings of the 4th International Conference on Learning Representations*.

Vlachas, P. R.; Byeon, W.; Wan, Z. Y.; Sapsis, T. P.; and Koumoutsakos, P. 2018. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proc. R. Soc. A.*, 474(2213): 20170844.

Wiener, N. 1930. Generalized harmonic analysis. *Acta Mathematica*, 55: 117–258. Publisher: Institut Mittag-Leffler.

Williams, R. J.; and Zipser, D. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2): 270–280.

Wu, M.; and Goodman, N. 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning. *arXiv:1802.05335 [cs, stat]*.

Zhao, Y.; and Park, I. M. 2020. Variational Online Learning of Neural Dynamics. *Front. Comput. Neurosci.*, 14.

Zheng, X.; Zaheer, M.; Ahmed, A.; Wang, Y.; Xing, E. P.; and Smola, A. J. 2017. State Space LSTM Models with Particle MCMC Inference. *arXiv preprint arXiv:1711.11179*.

Öğretir, M.; Ramchandran, S.; Papatheodorou, D.; and Lähdesmäki, H. 2022. A Variational Autoencoder for Heterogeneous Temporal and Longitudinal Data. *arXiv:2204.09369 [cs, stat]*.

Appendix

Performance Measures

Geometrical Measure D_{stsp} measures the match between the ground truth distribution $p_{\text{true}}(\boldsymbol{x})$ and the generated distribution $p_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})$ across *state space* through a discrete binning approximation: (Koppe et al. 2019b)

$$D_{\text{stsp}}\left(p_{\text{true}}(\boldsymbol{x}), p_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})\right) \approx \sum_{k=1}^{K} \hat{p}_{\text{true}}^{(k)}(\boldsymbol{x}) \log\left(\frac{\hat{p}_{\text{true}}^{(k)}(\boldsymbol{x})}{\hat{p}_{\text{gen}}^{(k)}(\boldsymbol{x} \mid \boldsymbol{z})}\right). \tag{10}$$

Here, K is the total number of bins. $\hat{p}_{\text{true}}^{(k)}(\boldsymbol{x})$ and $\hat{p}_{\text{gen}}^{(k)}(\boldsymbol{x} \mid \boldsymbol{z})$ are estimated by sampling 100 trajectories with randomly drawn initial conditions and 1000 time steps each. A range of $2\times$ the data standard deviation on each dimension was partitioned into m bins, leading to a total of $K = m^N$ bins, where N is the dimension of the ground truth system. Initial transients are removed after sampling trajectories to ensure that the measure is evaluated on the limit set.

Power Spectrum Correlation The power spectrum correlation (PSC) was obtained by first sampling a time series of $100,000$ time steps and computing dimension-wise Fast Fourier Transforms (using `scipy.fft`) for both the ground truth system and simulated time series. The noise dominated high-frequency tails of the spectra were cut off, and the power spectra were slightly smoothed with a Gaussian kernel and normalized. The obtained correlations between smoothed power spectra were then averaged to obtain the PSC values given in Table 1.

Mean Squared Prediction Error A mean squared prediction error (PE) was computed across test sets of length $T = 10000$ by initializing the trained dendPLRNN with the test set time series up to some time point t, from where it was then iterated forward by n time steps to yield a prediction at time step $t + n$. The n-step PE is then defined as the MSE between predicted and true observations:

$$PE(n) = \frac{1}{N(T-n)} \sum_{t=1}^{T-n} \sum_{i=1}^{N} (x_{i,t+n} - \hat{x}_{i,t+n})^2 \tag{11}$$

Due to exponential divergence of initially close trajectories in chaotic systems, the PE is sensible only for a limited number of time steps (Koppe et al. 2019b)). The rate of exponential divergence is quantified through the system's maximal Lyapunov exponent (Kantz and Schreiber 2004).

Ordinal Prediction Error The ordinal PE was computed similarly as the mean squared PE above, but – as pointed out in Öğretir et al. (2022) – due to the non-metric nature of ordinal data taking the absolute ($L1$) deviation between observed and predicted values is more sensible:

$$OPE(n) = \frac{1}{N(T-n)} \sum_{t=1}^{T-n} \sum_{i=1}^{N} |x_{i,t+n} - \hat{x}_{i,t+n}| \tag{12}$$

Spearman Cross-Correlation Error To assess whether the global cross-correlation structure between the different ordinal time series is preserved by the reconstruction method, the Spearman correlation between each pair of ordinal time series was computed based on $100,000$ time steps long samples, using `scipy.stats.spearmanr`, for both generated and test set data (see Fig. 4. The squared error between all elements of the correlation matrices was then taken and averaged across dimensions.

Spearman Autocorrelation Function To assess the temporal agreement between generated and ground truth *ordinal* and *count* observations, we computed a measure based on the average Spearman autocorrelation function (ACF). To this end we first sampled a time series of $100,000$ time steps and compute the dimension-wise Spearman autocorrelation for time lags up to 200 for both generated data and test set data (see Fig. 4). The squared error between the resulting ACFs was then averaged across all dimensions.

Geometric reconstruction measure in the absence of continuous observations If the underlying DS was observed only through time series of discrete random variables, we lack a direct mapping between the true and reconstructed continuous state spaces. To construct a mapping for such cases that does not fundamentally alter the geometry of the reconstructed attractor (i.e., does not introduce additional degrees of freedom for the mapping of geometries), we used a two-step procedure resulting in a geometry-preserving linear operator. First, the dendPLRNN's latent state space was reduced to the same dimensionality N as that of the ground truth system (which usually is of lower dimensionality) simply by Principal Component Analysis (PCA). In a second step, a rotation matrix was then determined to rotate the latent state space such as to minimize the same Kullback-Leibler measure D_{stsp} that was used to assess agreement in attractor geometries, see Fig. 7 for an example. This was done simply by grid search over the space of rotation matrices, as we found numerical optimization to often yield inferior results. Note that this operation does not change the geometry of objects in the latent space but merely rotates them such that they are best aligned with their ground truth counterparts (more generally, any reconstruction geometrically similar to the original attractor object can be considered valid). Comparing different grid- and step sizes in preliminary runs, we fixed parameters such that a single grid

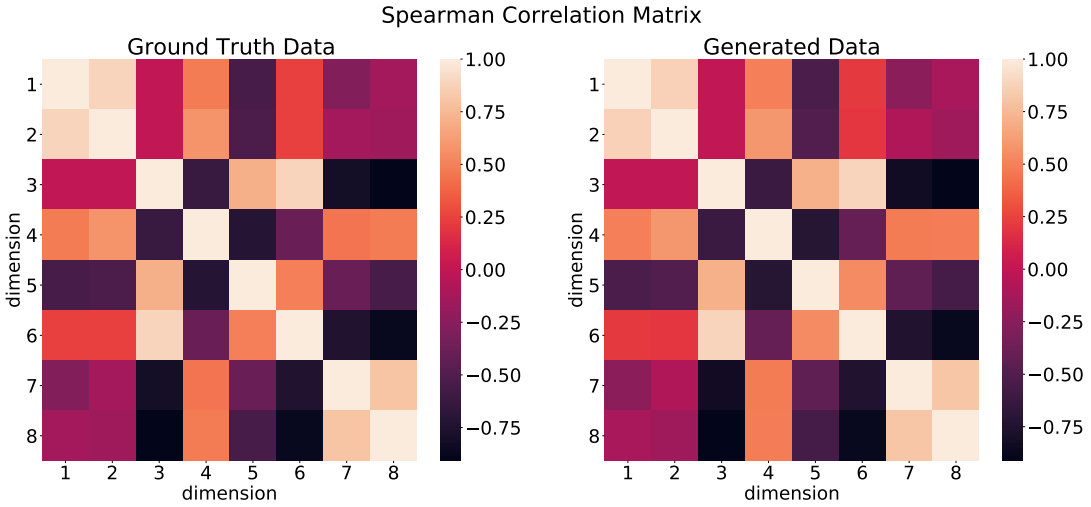

Figure 4: The Spearman correlation matrix for freely generated ordinal trajectories from a reconstructed Lorenz-63 system, and for the ground truth ordinal data. The correlation structure of simulated data closely resembles those of the ground truth data.

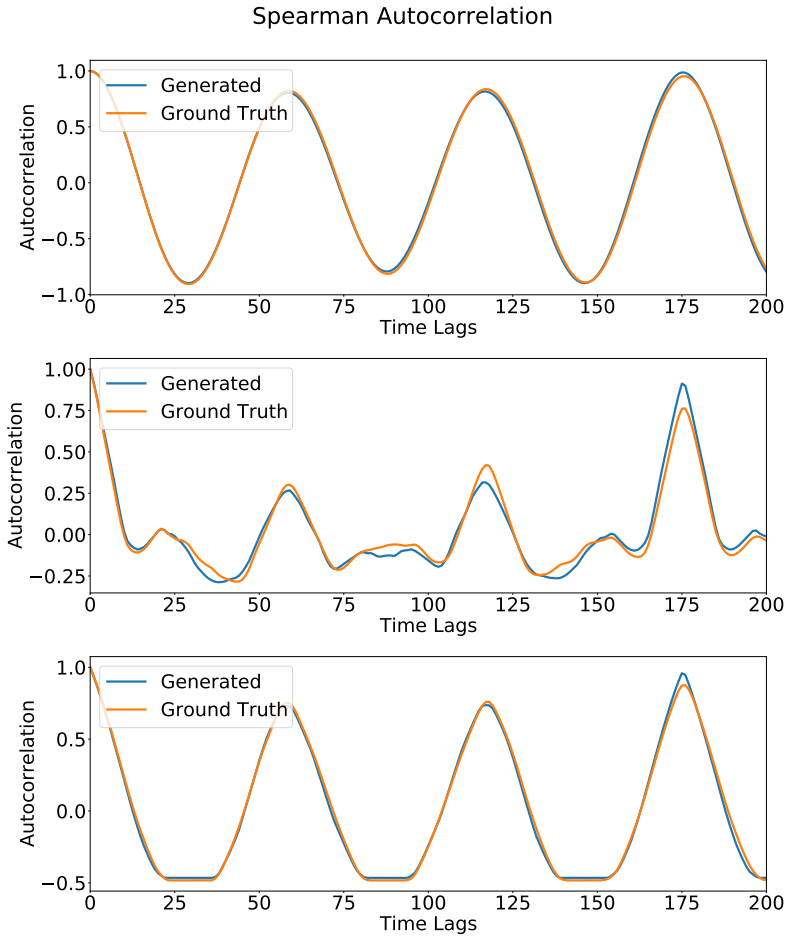

Figure 5: The Spearman ACF for three simulated ordinal time series from a reconstructed Rössler system, compared to that of the (ground truth) test data.

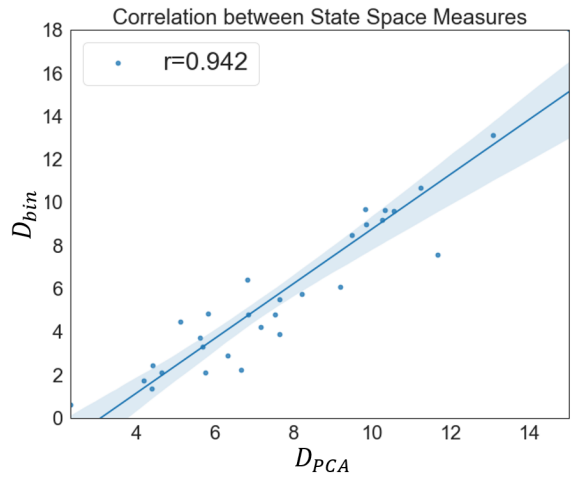

Figure 6: Correlation between geometrical reconstruction measures computed for the Rössler system in observation space from continuous Gaussian observations (D_{bin}) and on the first three principal components of latent space after optimally rotating the cloud of points (D_{PCA}). This gave a correlation of 0.941 for data from 30 trained models.

search takes no more than $30 - 60$ seconds on a single CPU. To confirm that this procedure yields results in agreement with those obtained from a co-trained linear-Gaussian model fed with continuous observations, we compared D_{stsp} computed in observation space ('D_{bin}') as outlined above with D_{stsp} ('D_{PCA}') obtained from ordinal data alone using our PCA+rotation method. As shown in Fig. 6, these two measures were indeed highly correlated, $r \approx 0.94$. An example reconstruction from solely ordinal observations for a Rössler system is given in Fig. 3).

Details on Dynamical Systems Benchmarks

Lorenz-63 System The $3d$ chaotic Lorenz-63 attractor, originally proposed in Lorenz (1963), is defined by

$$\begin{aligned} \mathrm{d}x &= (\sigma(y-x))\mathrm{d}t + \mathrm{d}\epsilon_1(t), \\ \mathrm{d}y &= (x(\rho - z) - y)\mathrm{d}t + \mathrm{d}\epsilon_2(t), \\ \mathrm{d}z &= (xy - \beta z)\mathrm{d}t + \mathrm{d}\epsilon_3(t). \end{aligned} \tag{13}$$

Parameters used for producing ground truth data in the chaotic regime were $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. Process noise was injected into the system by drawing from a Gaussian term $d\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.01^2 dt \times \mathbf{I})$. For both training and test data, a trajectory of $100,000$ time steps was sampled. To obtain multimodal observations, the observation models from Eqs. 16 and 18 were randomly initialized and the trajectories drawn from the ground truth systems used as latent factors.

Rössler System The Rössler attractor was introduced by Otto Rössler in 1976 in Rössler (1976) as a simpler version of the Lorenz attractor, and is given by

$$\begin{aligned} \mathrm{d}x &= (-y - z)\mathrm{d}t + \mathrm{d}\epsilon_1(t), \\ \mathrm{d}y &= (x + ay)\mathrm{d}t + \mathrm{d}\epsilon_2(t), \\ \mathrm{d}z &= (b + z(x - c))\mathrm{d}t + \mathrm{d}\epsilon_3(t). \end{aligned} \tag{14}$$

Parameters used for producing ground truth data in the chaotic regime were $a = 0.2, b = 0.2$, and $c = 5.7$. Process noise was added by drawing from $d\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.01^2 dt \times \mathbf{I})$. Training and test data was sampled as described above for the Lorenz-63 system.

Observation Models for Multimodal Data

Ordinal Model Ordinal data are not associated with a metric space, but there is a natural ordering between variables, e.g. in the context of survey data in economy or psychology commonly assessed by the Likert scale (Likert 1932). Treating ordinal data as metric data can lead to a variety of problems, as pointed out in (Liddell and Kruschke 2018). Ordinal observations are coupled

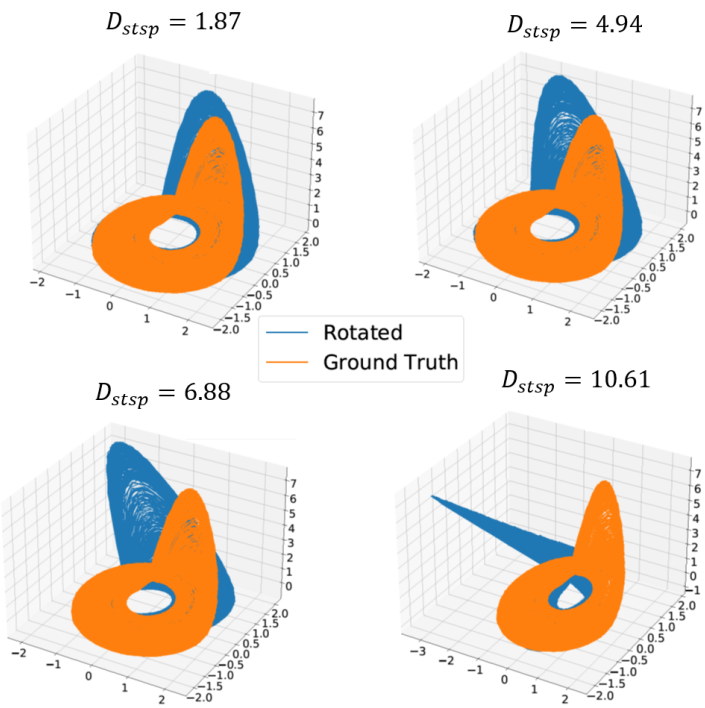

Figure 7: Ground truth and rotated attractors for the Roessler system with associated D_{stsp}-values.

to latent states via a generalized linear model (McCullagh 1980). Here, specifically, we assume that the ordinal observations are derived from an underlying continuous variable x_{ti}, which is linked to the latent states z_t via a linear model

$$x_{ti}^* = \boldsymbol{\beta}_i^T \mathbf{z}_t + \epsilon_{ti}, \tag{15}$$

where $\boldsymbol{\beta}_i^T \in \mathbb{R}^M$ are the model parameters and ϵ_{ti} is an independently distributed noise term. Inverting the link function leads to an expression for the cumulative probabilities:

$$p\left(x_{ti} \leq k \mid \mathbf{z}_t\right) = \frac{\exp\left(\beta_{ik}^0 - \boldsymbol{\beta}_i^T \mathbf{z}_t\right)}{1 + \exp\left(\beta_{ik}^0 - \boldsymbol{\beta}_i^T \mathbf{z}_t\right)} \tag{16}$$

The individual probabilities $p\left(x_{ti} = k \mid \mathbf{z}_t\right)$ follow from the cumulative probabilities via $p\left(x_{ti} = k \mid \mathbf{z}_t\right) = p\left(x_{ti} \leq k \mid \mathbf{z}_t\right) - p\left(x_{ti} \leq k-1 \mid \mathbf{z}_t\right)$, from which we can compute the log-likelihood as

$$\log p_\theta(\mathbf{x} \mid \mathbf{z}) = \sum_i^N \sum_t^T \sum_k^K \left[x_{ti} = k\right] \log p\left(x_{ti} = k \mid \mathbf{z}_t\right) \tag{17}$$

Poisson Model For count observations $\{\mathbf{c}_t\}_{t=1}^T$, with $\mathbf{c}_t = \left(c_{1t}, \ldots, c_{Lt}\right)^T$, we employ a Poisson observation model (Lambert 1992). The probability of an observed count under a Poisson model is given by

$$p\left(q_{lt} \mid \mathbf{z}_t\right) = \frac{\lambda_{lt}^{q_{lt}}}{q_{lt}!} e^{-\lambda_{lt}} \tag{18}$$

The probability is related to the latent states via a log-link function by $\log \lambda_{lt} = \gamma_0^{(l)} + \sum_{m=1}^M \gamma_m^{(l)} z_{mt}$, where $\gamma^{(l)}$ is a $1\text{x}M$ matrix. Thus, $\lambda_{lt} = e^{\gamma^{(l)} z_t}$ is the expected count for the $l-th$ observation at time t.

Hyperparameter Settings

To train the dendPLRNN with MVAE-TF, RAdam (Liu et al. 2020) with an initial learning rate of 10^{-3} that was iteratively reduced during training down to 10^{-5} was used. For each epoch, we randomly sampled sequences of length $T_{seq} = 300$ from the total training data pool of each dataset with a batch size of 16. Parameters $\boldsymbol{A}$, $\boldsymbol{W}$ and $\boldsymbol{h}$ were initialized according to Talathi and Vartak (2016). To train the dendPLRNN with the SVAE from (Kramer et al. 2022), we followed the implementation of the encoder model as provided on https://github.com/DurstewitzLab/mmPLRNN, training with a sequence-length of 100 time steps per batch, a hidden dimension of 20, and other model parameters kept similar to the runs with $MVAE-TF$. Hyperparameter settings for all runs were chosen such as to approximately keep the number of total parameters similar to that used for MVAE-TF.