Multimodal Teacher Forcing for Reconstructing Nonlinear Dynamical Systems

Manuel Brenner[1 2], **Georgia Koppe**[1 4 *], **Daniel Durstewitz**[1 2 3 *]

[1]Dept. of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty, Heidelberg University, Germany
[2]Faculty of Physics and Astronomy, Heidelberg University, Germany
[3]Interdisciplinary Center for Scientific Computing, Heidelberg University, Germany
[4]Dept. of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty, Heidelberg University, Germany

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 chaotic DS benchmarks than various alternative methods.

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 (Bury et al. 2021; Patel and Ott 2022). 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; Koppe et al. 2019b), 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 (behavioral responses) (Durstewitz, Koppe, and Thurm 2022), 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 (Kramer et al. 2022), major issues remain. Arguably one of the most challenging aspects is the 'exploding and vanishing gradient problem (EVGP)' (Bengio, Simard, and Frasconi 1994; Hochreiter and Schmidhuber 1997), a problem that cannot easily be avoided when attempting to capture natural systems with chaotic dynamics due to the exponential divergence of trajectories (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), we use multimodal VAEs (MVAEs) to construct a common latent representation from random variables following different distributional models. However, rather than constructing a sequential VAE process that directly operates on this latent code, as in Kramer et al. (2022), here

*These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org), AAAI 2023 Workshop "When Machine Learning meets Dynamical Systems: Theory and Applications" (MLmDS 2023). All rights reserved.

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 simultaneously trained RNN. This makes the insights derived from sparse-TF applicable to a fully probabilistic MVAE training framework, called MVAE-TF in here, which – unlike classical BPTT-TF – also explicitly allows for process (dynamical) noise and 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 geometrical 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) (Vlachas et al. 2018), 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; Hochreiter and Schmidhuber 1997; 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; Chang et al. 2019; Kag, Zhang, and Saligrama 2020; 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 $\{\boldsymbol{x}_{1:T}\}$ (Ghahramani and Roweis 1998; Durstewitz 2017; Pandarinath et al. 2018; Zhao and Park 2020; Kim et al. 2021). Through their probabilistic formulation, they can account for uncertainty in the model formulation or latent process itself and yield the full distribution over latent states (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 2015). 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) an MVAE for producing a multimodal TF signal, trained jointly with the dendPLRNN 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 a 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

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

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 translation into an equivalent 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 Appx.). For example, for normally distributed data 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 in 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

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

[1]Note that by 'invertible' we are referring to the pseudo-inverse of $\boldsymbol{B}$ here.

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.

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, $\boldsymbol{Y} = \{\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}; \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T\}; \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T\}\}$. We employ an MVAE for inferring a joint latent representation over these data. We denote encoded states at time t by $\tilde{\boldsymbol{z}}_t \in \mathbb{R}^K$ to avoid confusion with the latent dynamical process $\boldsymbol{z}_t \in \mathbb{R}^M$ generated by the dendPLRNN. To this end, we minimize the negative Evidence Lower Bound (ELBO)

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{Y}) = &-\mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}) + \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{Z}})] \\ &- \mathbb{H}_{q_{\boldsymbol{\phi}}}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{Y})\end{aligned} \tag{5}$$

using the reparameterization trick for latent random variables (Kingma and Welling 2014). For the encoding step, we concatenate all input modalities and use convolutional neural networks (CNNs) for parameterizing the encoder model $q_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Z}}|\boldsymbol{Y})$, with hyperparameters as proposed in Brenner et al. (2022). This allows the recognition model to embed temporal 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 variational density $q_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Z}}|\boldsymbol{Y}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{Y}), \boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\boldsymbol{Y}))$ to approximate the true posterior $p(\tilde{\boldsymbol{Z}}|\boldsymbol{Y})$, where mean and covariance are functions of the data. We further use the common mean field approximation to factorize $q_{\boldsymbol{\phi}}(\tilde{\boldsymbol{Z}}|\boldsymbol{Y})$ across time (Girin et al. 2021). Observations are assumed to be conditionally independent given the latent states $\tilde{\boldsymbol{z}}_t$. 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):

$$\begin{aligned}\boldsymbol{x}_t \mid \tilde{\boldsymbol{z}}_t &\sim \mathcal{N}(\boldsymbol{B}\tilde{\boldsymbol{z}}_t, \boldsymbol{\Gamma}) \\ \boldsymbol{o}_t \mid \tilde{\boldsymbol{z}}_t &\sim \text{Ordinal}(\boldsymbol{\beta}\tilde{\boldsymbol{z}}_t, \epsilon) \\ \boldsymbol{c}_t \mid \tilde{\boldsymbol{z}}_t &\sim \text{Poisson}(\lambda(\tilde{\boldsymbol{z}}_t))\end{aligned} \tag{6}$$

Due to the conditional independence given the latent states, the likelihood terms related to the observations simply sum up, $\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}) = \sum_{t=1}^{T} (\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\tilde{\boldsymbol{z}}_t) + \log p_{\boldsymbol{\theta}}(\boldsymbol{o}_t|\tilde{\boldsymbol{z}}_t) + \log p_{\boldsymbol{\theta}}(\boldsymbol{c_t}|\tilde{\boldsymbol{z}}_t))$.

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 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ from possibly many different and non-Gaussian modalities are first converted to a common latent code $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_1, \cdots, \tilde{\boldsymbol{z}}_T\}$ using the MVAE. The initial condition $\boldsymbol{z}_1$ of the latent process is inferred from the encoded state $\tilde{\boldsymbol{z}}_1$. If $K < M$, the $M - K$

remaining states are randomly sampled from a standard 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 generated latent trajectory $\boldsymbol{Z}$ (using the unforced states) are then used to compute the modality-specific negative log-likelihoods for the observed multimodal time series, $\mathcal{L}_{\text{PLRNN}} = -\sum_{t=1}^{T}(\log p_{\boldsymbol{\theta}}(\boldsymbol{x_t}|\boldsymbol{z}_{1:K,t}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{o}_t|\boldsymbol{z}_{1:K,t}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{c}_t|\boldsymbol{z}_{1:K,t}))$, using the decoder models from Eq. 6 (hence, importantly, the latent states $\boldsymbol{z}_t$ and $\tilde{\boldsymbol{z}}_t$ of the dendPLRNN and MVAE, respectively, are *both* coupled to the observations through the very *same* set of decoder models with same parameters).

A crucial assumption of the MVAE-TF framework now is that the process prior $p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{Z}})$ in Eq. 5 comes from the dendPLRNN. For the second term in Eq. 5 we assume

$$\mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{Z}})] \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{t=1}^{T} -\frac{1}{2}\big(\log|\boldsymbol{\Sigma}| + (\tilde{\boldsymbol{z}}_t^{(l)} - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}_t^{(l)} - \boldsymbol{\mu}_t) + const.\big), \tag{7}$$

where the expectation value is approximated by L Monte Carlo samples $\tilde{\boldsymbol{z}}_t^{(l)} \sim q_{\boldsymbol{\phi}}(\tilde{\boldsymbol{z}}_t)$. To connect the latent codes of the MVAE and the dendPLRNN, the model prior of the MVAE is instantiated through the dendPLRNN by taking $\boldsymbol{\mu}_t = \boldsymbol{z}_{1:K,t}$ (i.e., the first K states of the generated latent sequence $\{\boldsymbol{z}_t\}$). As the initial state $\boldsymbol{z}_1$ is estimated directly from the encoded state $\tilde{\boldsymbol{z}}_1$, the term for $t = 1$ evaluates to zero. Setting $L = 1$, we thus obtain a consistency loss between encoded and generated latent state paths as

$$\mathcal{L}_{cons} = \frac{1}{2}\sum_{t=2}^{T}\big(\log|\boldsymbol{\Sigma}| + (\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_{1:K,t})^\top \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_{1:K,t})\big) \tag{8}$$

The general scheme for the MVAE-TF is visualised in Fig. 1. The total MVAE-TF loss is thus given by the reconstruction loss of the MVAE (first and third term in Eq. 5), the latent loss in Eq. 8 that ensures consistency between the latent codes of the MVAE and dendPLRNN, and the dendPLRNN loss from the likelihoods of the observed time series $\boldsymbol{Y}$ given the predicted latent path $\boldsymbol{Z}$ (above Eq. 7):

$$\mathcal{L}_{total} = \mathcal{L}_{MVAE} + \mathcal{L}_{cons} + \mathcal{L}_{PLRNN} \tag{9}$$

Experiments

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 because of exponential trajectory divergence (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):

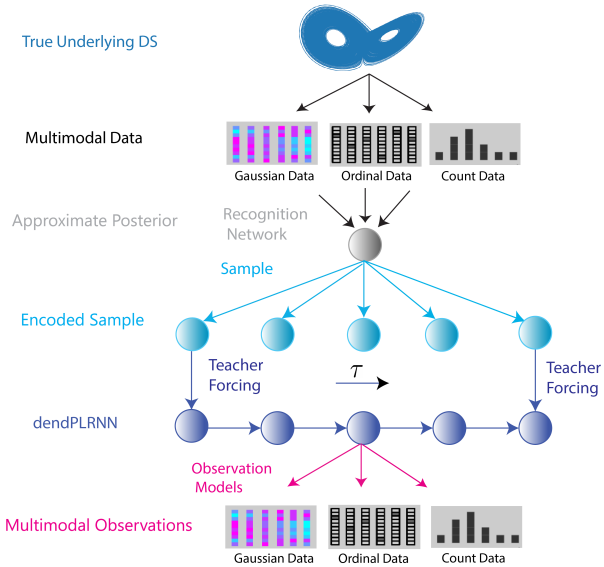

Figure 1: MVAE-TF setup. Multimodal observations are translated via an encoder into a common latent representation, which is used for sparse TF in the dendPLRNN's latent space. The latent trajectory is then mapped back into the multimodal observation space via modality-specific observation (decoder)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

- To assess overlap in *attractor geometries*, we employ a Kullback-Leibler divergence in state space, D_{stsp} (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 and slightly smoothed with Gaussian kernels to remove noise. The Hellinger distance D_H between empirical and model-generated power spectra was then computed. To assess the agreement in temporal structure for *ordinal* and *count data*, we determined the MSE between autocorrelation functions (Wiener 1930) computed for up to 200 time lags based on the Spearman rank correlation coefficient for discrete ordinal data (see Appx. Fig. 9, OACF for ordinal and CACF for count data in Table 1). We also assessed how well the Spearman cross-correlation structure between observed ordinal variables was preserved within the generated time series (SCC, see Appx. Fig. 4).
- To assess *short-term behavior*, for *Gaussian* data the classical 10-step-ahead prediction error (PE) along test set trajectories was used. For *ordinal data*, we computed a linear L_1 PE (OPE) instead, as suggested in Öğretir et al. (2022).

Benchmark Comparisons on Multi-Modal Time Series We first evaluate the MVAE-TF's ability to combine different

Dataset	Method	$D_{stsp}\downarrow$	$D_H\downarrow$	PE $\downarrow$	OPE $\downarrow$	SCC $\downarrow$	OACF $\downarrow$	CACF $\downarrow$
Lorenz	MVAE-TF	$\mathbf{3.4 \pm 0.35}$	$\mathbf{0.30 \pm 0.06}$	$\mathbf{1.3\text{e}{-2} \pm 2\text{e}{-4}}$	$\mathbf{0.12 \pm 0.03}$	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$	$\mathbf{6.6\text{e}{-5} \pm 8.1\text{e}{-6}}$
	SVAE	11.1 ± 0.6	0.82 ± 0.05	$6.3\text{e}{-1} \pm 5.1\text{e}{-2}$	0.68 ± 0.03	0.14 ± 0.01	0.18 ± 0.02	$8.5\text{e}{-5} \pm 1.6\text{e}{-5}$
	GVAE-TF	4.3 ± 0.3	0.47 ± 0.07	$3.6\text{e}{-1} \pm 1.5\text{e}{-3}$	X	X	X	X
	BPTT-TF	8.8 ± 1.9	0.86 ± 0.05	$4.4\text{e}{-1} \pm 2.2\text{e}{-2}$	X	X	X	X
	MS	4.5 ± 1.5	0.61 ± 0.08	X	X	0.14 ± 0.04	0.11 ± 0.02	$\mathbf{6.5\text{e}{-5} \pm 3.8\text{e}{-6}}$
Rössler	MVAE-TF	$\mathbf{1.45 \pm 0.71}$	$\mathbf{0.32 \pm 0.03}$	$\mathbf{1.9\text{e}{-3} \pm 7.1\text{e}{-5}}$	$\mathbf{0.08 \pm 0.02}$	$\mathbf{0.04 \pm 0.004}$	$\mathbf{0.017 \pm 0.003}$	$\mathbf{6.5\text{e}{-5} \pm 1.2\text{e}{-5}}$
	SVAE	10.7 ± 1.5	0.66 ± 0.05	$1.5\text{e}{-1} \pm 3.1\text{e}{-2}$	0.24 ± 0.02	0.17 ± 0.03	0.13 ± 0.02	$1.1\text{e}{-4} \pm 1.4\text{e}{-5}$
	GVAE-TF	12.1 ± 0.5	0.55 ± 0.04	$4.9\text{e}{-2} \pm 3.4\text{e}{-3}$	X	X	X	X
	BPTT-TF	8.9 ± 1.4	0.64 ± 0.07	$2.8\text{e}{-1} \pm 1.8\text{e}{-3}$	X	X	X	X
	MS	3.99 ± 1.1	0.59 ± 0.04	X	X	0.08 ± 0.04	0.09 ± 0.02	$1.6\text{e}{-4} \pm 5.9\text{e}{-5}$
Lewis-Glass	MVAE-TF	$\mathbf{0.27 \pm 0.07}$	$\mathbf{0.33 \pm 0.02}$	$\mathbf{2.1\text{e}{-3} \pm 7\text{e}{-5}}$	$\mathbf{0.11 \pm 0.01}$	0.12 ± 0.03	$\mathbf{0.05 \pm 0.02}$	$2.3\text{e}{-4} \pm 2.0\text{e}{-5}$
	SVAE	2.6 ± 0.5	0.52 ± 0.03	$8.0\text{e}{-2} \pm 4\text{e}{-3}$	0.26 ± 0.01	0.4 ± 0.05	0.18 ± 0.03	$7.5\text{e}{-3} \pm 4.7\text{e}{-3}$
	GVAE-TF	$\mathbf{0.28 \pm 0.08}$	0.44 ± 0.02	$4.6\text{e}{-3} \pm 4\text{e}{-4}$	X	X	X	X
	BPTT-TF	2.51 ± 0.71	0.43 ± 0.03	$2.6\text{e}{-2} \pm 3\text{e}{-3}$	X	X	X	X
	MS	$\mathbf{0.33 \pm 0.06}$	$\mathbf{0.35 \pm 0.01}$	X	X	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.04 \pm 0.002}$	$\mathbf{1.9\text{e}{-4} \pm 7.5\text{e}{-6}}$

Table 1: Comparison of dendPLRNN trained by MVAE-TF (proposed method), by a SVAE based on Kramer et al. (2022), an VAE-TF approach similar to MVAE-TF except that all data modalities were 'Gaussianized' (GVAE-TF), BPTT-TF as in Brenner et al. (2022) using Gaussianized data, and a multiple-shooting (MS) approach (see Appx. for details). Training was performed on multivariate normal, ordinal, and count data produced by the chaotic Lorenz system, Rössler system, and Lewis-Glass model. Observation noise with 10% of the data variance was added to the Gaussian observations. Values are mean $\pm$ SEM, averaged over 15 trained models. X = value cannot be computed for this model (e.g., because resp. decoder model is not present).

observed data modalities (Gaussian, ordinal, and count data) for inferring a common latent DS model on three ground-truth datasets, and compare its performance to a variety of other methods. We generated a training and a test set of $100,000$ time steps from a Lorenz-63 and a Rössler system with 1% process noise, and a 6d Lewis-Glass network model (Lewis and Glass 1992; Gilpin 2022), all in their chaotic regimes (see Appx. for detailed parameter settings and numerical integration). From the simulated trajectories we then sample ordinal and count observations using Eqs. 18 and 20 (with randomly drawn parameters), as well as continuous observations with 10% Gaussian noise. Example reconstructions of the MVAE-TF are in Fig. 9. 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. We also included three other comparisons: One naive approach is to transform all data modalities to approximately Gaussian (via Box-Cox & Gaussian kernel smoothing, see Appx.), and then either train the dendPLRNN via standard BPTT-TF (Brenner et al. 2022) or by VAE-TF, as proposed here, but without multi-modal integration (labeled GVAE-TF in Table 1). A third approach that can deal with multi-modal observation models but, unlike TF, does not require model inversion, is 'multiple shooting (MS)', a method suggested in the dynamical systems literature (Voss, Timmer, and Kurths (2004); see Appx. for more details). As evidenced in Table 1, MVAE-TF outperforms all these other possible model setups. We also tested the situation where faithful reconstruction should be possible from the Gaussian modality alone, with just 1% Gaussian noise. As seen in Appx. Table 4, even in this case MVAE-TF outperforms all other methods (possibly because here count and ordinal data may actually tend to mislead DS reconstruction for some of the other models).

Challenging Data Situations To really challenge the algorithm, we next tested a scenario where continuous observations from the Lorenz-63 system were heavily distorted by Gaussian noise with 50% of the data variance. At the same time, ordinal observations with 8 variables divided into 7 ordered categories each, $o_{nt} \in \{1 \dots 7\}$, $n = 1...8$, were sampled using Eq. 18. We then trained the dendPLRNN via MVAE-TF once with, and once without, ordinal observations. Figure 2(a) proves that with ordinal observations on board, DS reconstruction is, in principle, possible even under these challenging conditions. The cumulative histograms of the geometric measure D_{stsp} in Figure 2(b) and the temporal measure D_H in Figure 8 (b), comparing runs in the unimodal and multimodal settings, furthermore shows that inclusion of ordinal observations significantly improves reconstruction of the underlying system.

Motivated by these results, we pushed the system even further and attempted DS reconstruction *solely from ordinal data* (created as above), i.e. completely omitting continuous observations. 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 ground truth state space and that of the trained dendPLRNN (which in the case of Gaussian observations would simply be given by Eq. 3), we construct one post-hoc by optimizing a linear operator given by a linear dimensionality reduction (PCA) concatenated with a geometry-preserving rotation operation (see Appx.). Fig. 3 shows that, using MVAE-TF, successful DS reconstruction is, in principle, even feasible in this situation. Comparable results could not be achieved by the multimodal SVAE (see Appx., Table 3).

Conclusions

In the present work we introduced a novel training method for DS reconstruction from multimodal time series data based on

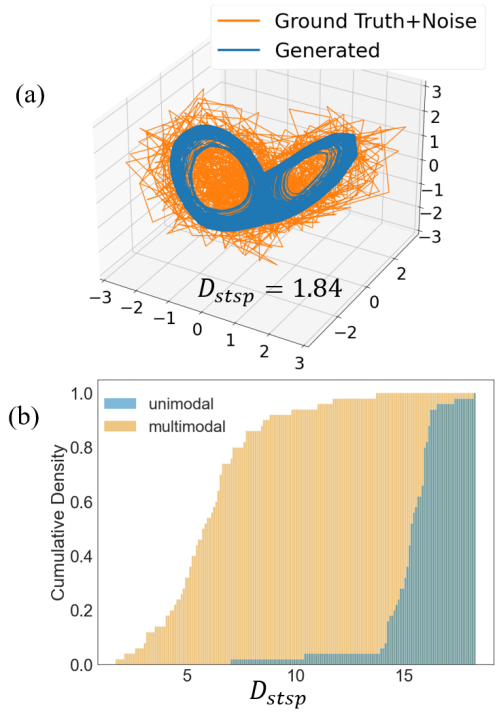

Figure 2: DS reconstruction from heavily distorted continuous observations (Gaussian observation noise of 50% of the data variance) and simultaneously provided ordinal observations. a) Example of a successful reconstruction ($D_{stsp} = 1.84$) of the butterfly wing structure of the Lorenz attractor by the MVAE-TF ($M = 20, K = 15, \tau = 10, B = 10$). b) Normalized cumulative densities of geometrical attractor disagreement (D_{stsp}) between reconstructed and ground-truth system.

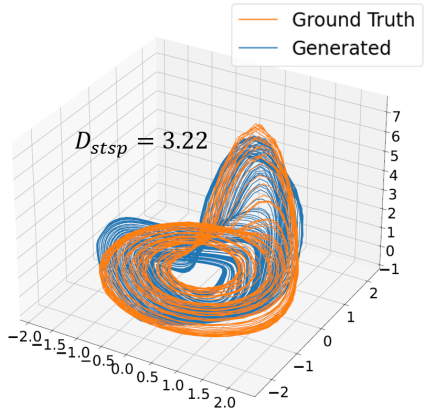

Figure 3: Rössler attractor reconstructed by the MVAE-TF solely from an 8-dimensional set of ordinal observations with 7 categories each ($M = 20, K = 15, \tau = 10, B = 10$). Comparison of attractor geometries was performed by reducing latent space dimensionality by PCA followed by a rotation operator that did not alter geometry.

dynamically interpretable RNNs. While DS reconstruction is meanwhile a large field in scientific ML (Brunton and Kutz 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). We show that for various chaotic benchmarks and sampling conditions, assessed by a variety of DS statistics, training the dendPLRNN by MVAE-TF clearly outperformed several other model formulations, including 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 one-step ahead prediction terms in the ELBO. Moreover, the MVAE-TF algorithm was able, in principle, 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.

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation Programme under grant agreement 945263 (IMMERSE), by the German Research Foundation (DFG) within the collaborative research center TRR-265 (project A06), and by a living lab grant by the Federal Ministry of Science, Education and Culture (MWK) of the state of Baden-Württemberg, Germany (grant number 31-7547.223-7/3/2), to DD and GK.

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.

Bock, H. G.; and Plitt, K. J. 1984. A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems*. *IFAC Proceedings Volumes*, 17(2): 1603–1608.

Box, G. E. P.; and Cox, D. R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2): 211–252. Publisher: [Royal Statistical Society, Wiley].

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.

Durstewitz, D.; Koppe, G.; and Thurm, M. I. 2022. Reconstructing Computational Dynamics from Neural Measurements with Recurrent Neural Networks. Pages: 2022.10.31.514408 Section: New Results.

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

Gilpin, W. 2022. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

Gower, J. C. 1975. Generalized procrustes analysis. *Psychometrika*, 40(1): 33–51.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

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.

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.

Lewis, J. E.; and Glass, L. 1992. Nonlinear Dynamics and Symbolic Dynamics of Neural Networks. *Neural Computation*, 4(5): 621–642.

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].

Meng, F.; Richer, M.; Tehrani, A.; La, J.; Kim, T. D.; Ayers, P. W.; and Heidar-Zadeh, F. 2022. Procrustes: A python library to find transformations that maximize the similarity between matrices. *Computer Physics Communications*, 276(108334): 1–37.

Mikhaeil, J. M.; Monfared, Z.; and Durstewitz, D. 2022. On the difficulty of learning chaotic dynamics with RNNs. In *Advances in Neural Information Processing Systems*.

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 auto-encoders. *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. 2015. *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.

Voss, H. U.; Timmer, J.; and Kurths, J. 2004. Nonlinear dynamical system identification from uncertain and indirect measurements. *International Journal of Bifurcation and Chaos*, 14(6): 1905–1933. Publisher: World Scientific Publishing Co.

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.

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 To assess the (dis-)agreement D_{stsp} between the data distribution $p_{\text{true}}(\boldsymbol{x})$ and the generated distribution $p_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})$ across *state space*, $\hat{p}_{\text{true}}(\boldsymbol{x})$ and $\hat{p}_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})$ are estimated by sampling 100 trajectories with randomly drawn initial conditions and 1000 time steps each. Transients are removed from each sampled trajectory to ensure that the measure is evaluated on the limit set. The match between distributions is then approximated by binning the state space into discrete bins (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. 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. Due to the exponential scaling of the number of bins with system dimensionality, for the 6-dimensional Lewis-Glass network model we instead used an approximation of $p_{\text{true}}(\boldsymbol{x})$ and $p_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})$ based on Gaussian mixture models placed along trajectories, as described in Brenner et al. (2022).

Power Spectrum Hellinger Distance The power spectrum Hellinger distance (D_H) 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. We then computed the Hellinger distance (Mikhaeil, Monfared, and Durstewitz 2022) between smoothed power spectra of ground-truth, $F(\omega)$, and generated, $G(\omega)$, trajectories given by

$$H(F(\omega), G(\omega)) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{F(\omega)G(\omega)}d\omega} \in [0, 1] \tag{11}$$

The dimension-wise Hellinger distances were then averaged to yield the D_H values from Tables 1 and 4.

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:

$$PE(n) = \frac{1}{N(T-n)} \sum_{t=1}^{T-n} \sum_{i=1}^{N} (x_{i,t+n} - \hat{x}_{i,t+n})^2 \tag{12}$$

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).

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 (L_1) deviation between observed and predicted values is more sensible:

$$OPE(n) = \frac{1}{N(T-n)} \sum_{t=1}^{T-n} \sum_{i=1}^{N} |o_{i,t+n} - \hat{o}_{i,t+n}| \tag{13}$$

Spearman Cross-Correlation (SCC) 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.

Spearman Autocorrelation Function (SACF) To assess the temporal agreement between generated and ground truth *ordinal* and *count* observations, we computed a measure based on the average SACF. 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. 9). The squared error between the resulting SACFs 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, we aimed for a linear operator that does not introduce additional degrees of freedom for modifying the reconstructed attractor geometry, but consists only of 1) a projection into a space of same dimensionality (and re-standardization of variables) followed by 2) a (geometry-preserving) rotation. This was to ensure that the quality of geometrical agreement can be attributed solely to the reconstruction method and not to any post-hoc fitting. For the first step, we simply used Principal Component Analysis (PCA) to reduce the dendPLRNN's latent space to the same dimensionality

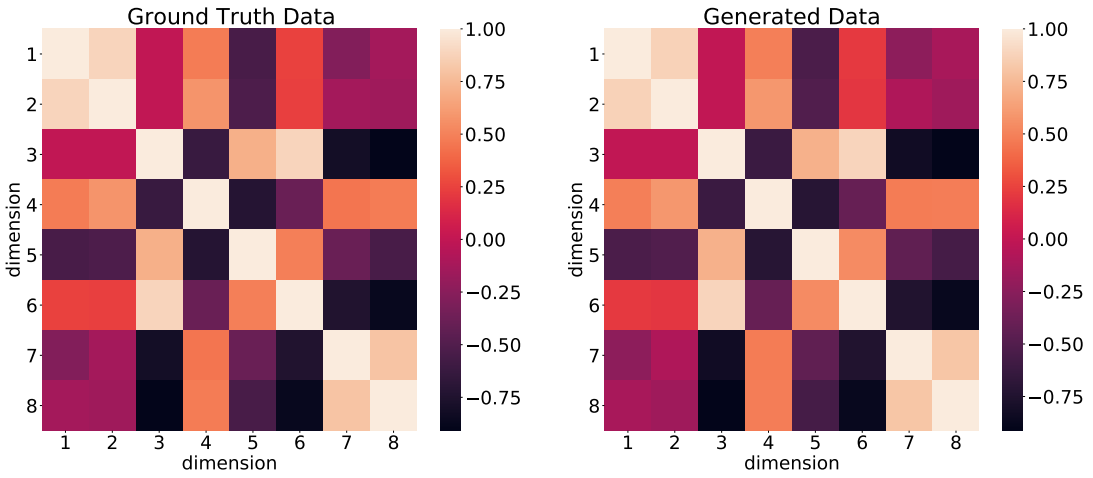

Figure 4: Spearman cross-correlation matrix for the ground truth ordinal data (left) and for freely generated ordinal trajectories (right) from a reconstructed Lorenz-63 system. The correlation structure of simulated data closely resembles that of the ground truth data.

N as that of the ground truth system (which usually is of lower dimensionality). Afterwards, all axes were re-standardized (as for the original system). In the 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. 5 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 alter the geometry* of objects in the latent space but merely rotates them such that they are best aligned with their ground truth counterparts (we also attempted Procrustes analysis (Gower 1975), using the Procrustes library (Meng et al. 2022), to determine the best affine mapping between spaces directly, but found this generally to be inferior despite actually being less conservative than our approach). Comparing different grid- and step sizes in preliminary runs, we fixed parameters such that a single grid 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} obtained from ordinal data alone using our PCA+rotation method ('D_{PCA}'). 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. For the 6-dimensional Lewis-Glass chaotic network model, performing a grid search over rotation matrices in the observation space was unfortunately no longer computationally feasible, such that in this case we needed to resort to Procrustes analysis (see above; generally, the Procrustes method aims to superimpose two data sets by optimally translating, rotating, and scaling them, preserving geometric similarity). In this case, the correlation between the D_{stsp} measures obtained by a co-trained linear model and the one obtained post-hoc via the Procrustes-transformed space dropped to $r \approx 0.57$, but was still significant.

Details on Dynamical Systems Benchmarks

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

$$\begin{aligned} \mathrm{d}x &= (\sigma(y-x))\mathrm{d}t + \mathrm{d}\epsilon_1(t), \\ \mathrm{d}y &= (x(\rho - z) - y)\mathrm{d}t + \mathrm{d}\epsilon_2(t), \\ \mathrm{d}z &= (xy - \beta z)\mathrm{d}t + \mathrm{d}\epsilon_3(t). \end{aligned} \tag{14}$$

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 $\mathrm{d}\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.01^2 \mathrm{d}t \times \boldsymbol{I})$. For both training and test data, a trajectory of $100,000$ time steps was sampled, performing numerical integration with `scipy.odeint` ($\mathrm{d}t = 0.05$; note this value differs from the one used in Mikhaeil, Monfared, and Durstewitz (2022), explaining the different τ values required). To obtain multimodal observations, trajectories drawn from the ground truth system were fed into the different types of observation models in Eq. 6, with randomly drawn parameters.

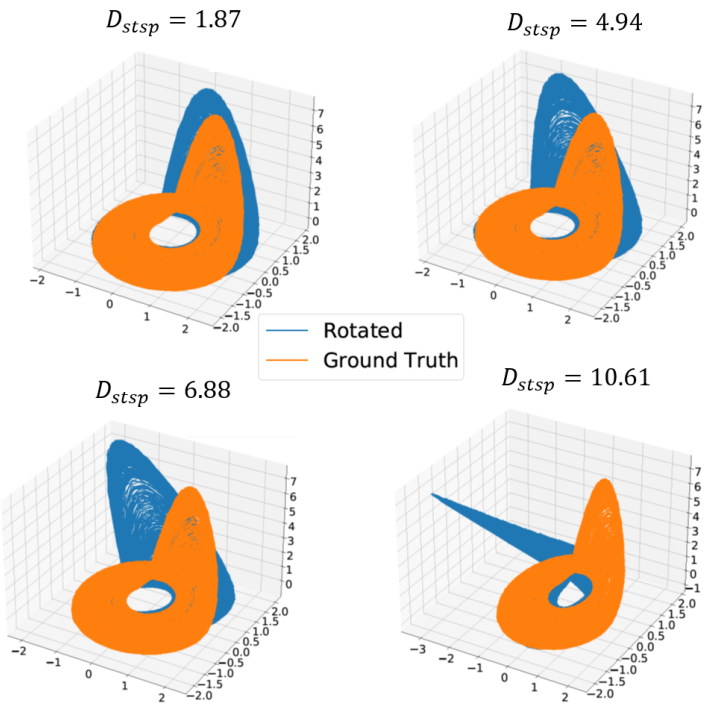

Figure 5: Ground truth and rotated attractors of the Rössler system with associated D_{stsp}-values.

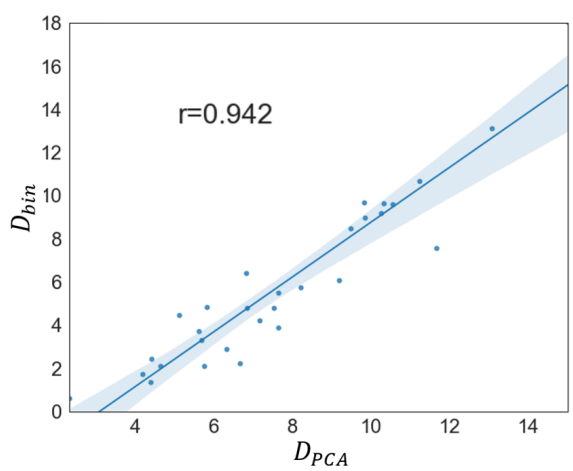

Figure 6: Correlation between geometrical reconstruction measures for the Rössler system directly in observation space given a co-trained linear (Gaussian) observation model (D_{bin}), and from a 3d PCA projection of latent space followed by an optimal rotation of the reconstructed attractor (D_{PCA}), based on a total of 30 trained models.

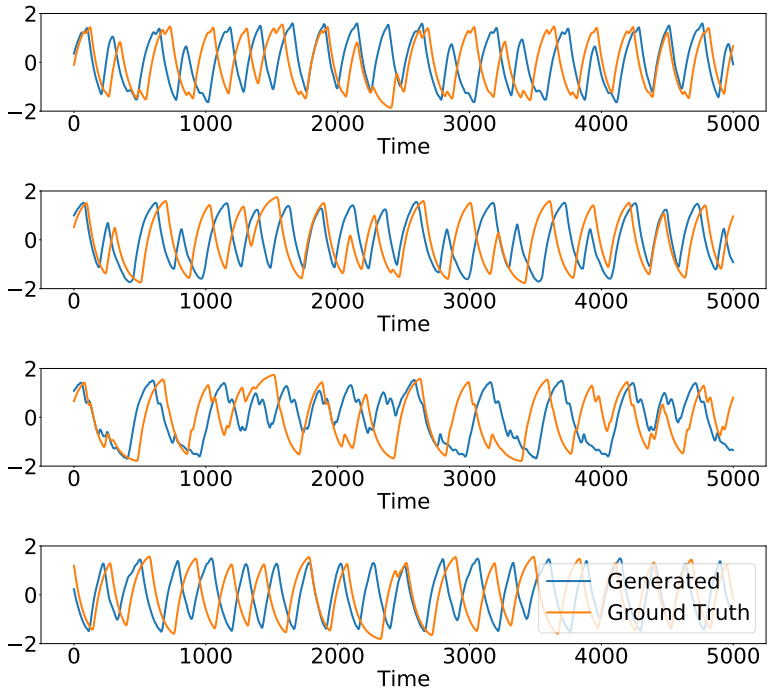

Figure 7: Example ground truth time series and freely generated series from a dendPLRNN ($M = 20, K = 15, \tau = 20, B = 10$) trained with MVAE-TF on the Lewis-Glass neural network model (Lewis and Glass 1992).

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

$$\begin{aligned} \mathrm{d}x &= (-y - z)\mathrm{d}t + \mathrm{d}\epsilon_1(t), \\ \mathrm{d}y &= (x + ay)\mathrm{d}t + \mathrm{d}\epsilon_2(t), \\ \mathrm{d}z &= (b + z(x - c))\mathrm{d}t + \mathrm{d}\epsilon_3(t). \end{aligned} \tag{15}$$

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 $\mathrm{d}\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.01^2\mathrm{d}t \times \boldsymbol{I})$. Training and test data was sampled as described above for the Lorenz-63 system, using $\mathrm{d}t = 0.1$.

Lewis-Glass Chaotic Network Model We simulate a 6-dimensional model of a neural network, originally introduced in Lewis and Glass (1992). Here the individual units of the network are endowed with a continuous gain function $G(x) = \frac{1+\tanh(-\alpha x)}{2}$, with the vector field given by

$$\mathrm{d}\boldsymbol{x}/\mathrm{d}t = \frac{-\boldsymbol{x}}{\tau} + G(\epsilon\, \boldsymbol{K}\boldsymbol{x}) - \beta \tag{16}$$

To sample from this system in the chaotic regime, we used the `Hopfield` model implementation in the Python package `dysts.flows`, based on Gilpin (2022). Here, $\alpha = -1, \beta = 0.5, \epsilon = 10, \tau = 2.5$, and

$$\boldsymbol{K} = \begin{bmatrix} 0 & -1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & -1 & 0 & 0 \end{bmatrix}$$

We generated training and test data by `maketrajectory`, and down-sampled the generated data by a factor of 30. We sampled ordinal and count data in the same way as for the other datasets. Example time series and reconstructions are displayed in Figure 7.

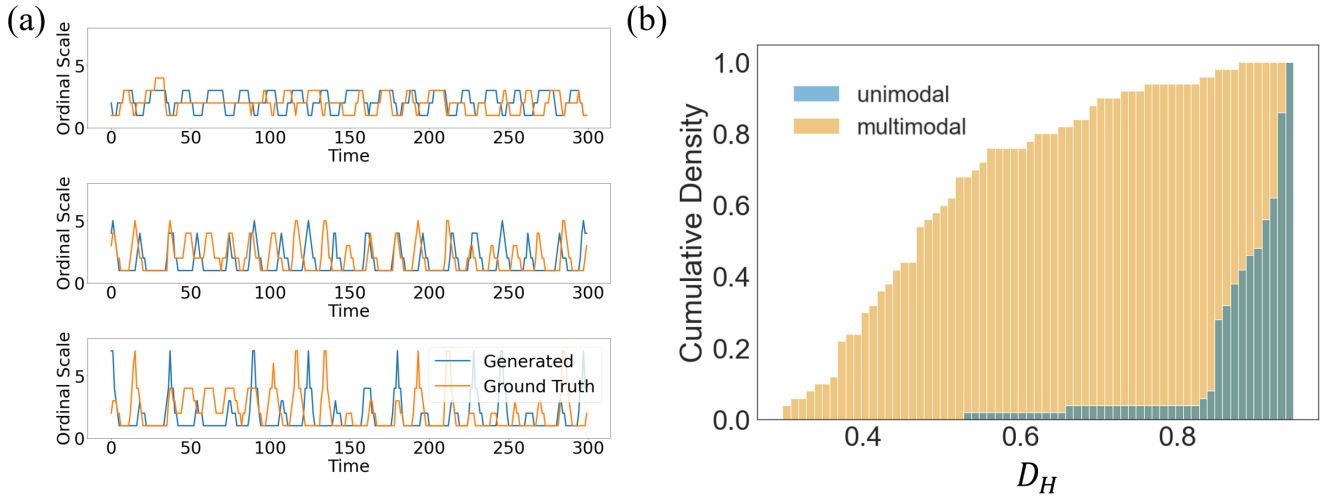

Figure 8: DS reconstruction with MVAE-TF from heavily distorted continuous observations and simultaneously provided ordinal observations (see also Figure 2). a) Example reconstructed ordinal observations with 7 categories each. b) Normalized cumulative densities of Hellinger distances between power spectra of reconstructed and ground-truth systems.

Observation Models for Multimodal Data

Ordinal Model Ordinal data are not associated with a metric space, but there is a natural ordering between variables, as e.g. in survey data in economy or psychology commonly assessed through Likert scales (Likert 1932). Treating ordinal data as metric can lead to a variety of problems, as pointed out in (Liddell and Kruschke 2018). Ordinal observations are coupled to latent states via a generalized linear model (McCullagh 1980). Here, specifically, we assume that the ordinal observations $\boldsymbol{o_t}$ are derived from an underlying continuous variable u_{it}, which is linked to the latent states $\boldsymbol{z}_t$ via a linear model

$$u_{it}^* = \boldsymbol{\beta}_i^T \boldsymbol{z}_t + \epsilon_{it}, \tag{17}$$

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

$$p\left(u_{it} \leq k \mid \boldsymbol{z}_t\right) = \frac{\exp\left(\beta_{ik}^0 - \boldsymbol{\beta}_i^T \boldsymbol{z}_t\right)}{1 + \exp\left(\beta_{ik}^0 - \boldsymbol{\beta}_i^T \boldsymbol{z}_t\right)} \tag{18}$$

The probability masses $p\left(u_{it} = k \mid \boldsymbol{z}_t\right)$ follow from the cumulative distribution via $p\left(u_{it} = k \mid \boldsymbol{z}_t\right) = p\left(u_{it} \leq k \mid \boldsymbol{z}_t\right) - p\left(u_{it} \leq k-1 \mid \boldsymbol{z}_t\right)$, from which we can compute the log-likelihood as

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{O} \mid \boldsymbol{Z}) = \sum_i^N \sum_t^T \sum_k^K \left[u_{it} = k\right] \log p\left(u_{it} = k \mid \boldsymbol{z}_t\right) \tag{19}$$

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

$$p_{\boldsymbol{\theta}}\left(c_{lt} \mid \boldsymbol{z}_t\right) = \frac{\lambda_{lt}^{c_{lt}}}{c_{lt}!} e^{-\lambda_{lt}} \tag{20}$$

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 $\boldsymbol{\gamma}^{(l)}$ is a $1\text{x}M$ vector. Thus, $\lambda_{lt} = e^{\gamma_0^{(l)} + \boldsymbol{\gamma}^{(l)} \boldsymbol{z_t}}$ is the expected count for the lth observation variable at time t.

Details on Training

Multiple Shooting Another technique from the dynamical systems literature for controlling trajectory divergence, similar in spirit to teacher forcing, is 'multiple shooting' (Bock and Plitt 1984). Multiple shooting aims to solve boundary value problems by dividing time into sub-intervals, treating each subinterval as a separate initial value problem, and then imposing continuity conditions between intervals. Multiple shooting has also been applied to DS reconstruction, where the initial conditions ('shooting

nodes') for each interval are model parameters and continuity across intervals is enforced through a penalty term in the loss (Voss, Timmer, and Kurths 2004). Hence, rather than controlling trajectory flows through a sparse teacher forcing signal applied after forcing intervals τ, alternatively one may reset the latent model trajectory to an inferred initial condition after τ time steps. The advantage is that this method does not require inversion of observation models and is hence naturally suited to handle different data modalities without further care (i.e., retaining the distributional properties of the original data). More specifically, the observed time series $\boldsymbol{Y}$ is partitioned into N_{seq} subsequences $\boldsymbol{Y}^s, s = 1 \dots N_{seq}$, of length L, and for each subsequence a new initial condition $\boldsymbol{\mu}_0^s$ is learned. During training trajectories are freely generated for L time steps from $\boldsymbol{\mu}_0^s$ for each subsequence, and likelihoods for the observed trajectories $\boldsymbol{Y}^s$ are computed using the observation models from Eq. 6. A consistency (penalty) term in the loss ensures continuity between subsequences according to

$$\mathcal{L}_{\text{MS}} = \lambda_{\text{MS}} \sum_{s=1}^{N_{seq}-1} ||F_\theta(\boldsymbol{z}_L^s) - \boldsymbol{\mu}_0^{s+1}||_2^2 \tag{21}$$

where F_θ in our case is the dendPLRNN, Eq. 1, λ_{MS} is a regularization parameter, and $F_\theta(\boldsymbol{z}_L^s) = F_\theta(F_\theta(\dots F_\theta((\boldsymbol{\mu}_0^s))) = F_\theta^L(\boldsymbol{\mu}_0^s)$. The sequence length L plays a similar role as the teacher forcing interval τ for MVAE-TF, controlling the times at which states and gradients are reset during training. Indeed, optimal settings for τ and L closely agreed for the datasets studied here (see Table 2).

Standard Unimodal Approach with Data 'Gaussianization' A naive approach for handling multi-modal observations with any type of DS reconstruction model would be to pre-process all modalities such as to bring them into approximate agreement with Gaussian assumptions. Thus, for training the dendPLRNN with standard BPTT-TF (Brenner et al. 2022) and GVAE-TF, we transformed ordinal and count observations into approximately Gaussian variables through a Box-Cox-transformation (Box and Cox 1964), z-scoring, and Gaussian kernel smoothing across the time series. For the optimal width of the Gaussian kernel, we performed a grid search over kernel sizes $\nu \in \{0, 0.01, 0.1, 1, 10, 15, 20, 25\}$. Optimal settings for the results displayed in Tables 1 and 4 are given in Table 2.

Hyperparameter Settings To train the dendPLRNN with MVAE-TF, RAdam (Liu et al. 2020) was used with a learning rate scheduler that iteratively reduced the learning rate from 10^{-3} to 10^{-5} during training. For each epoch, we randomly sampled sequences of length $T_{seq} = 300$ from the total training data with a batch size of 16. The network weights $\boldsymbol{A}$, $\boldsymbol{W}$ and $\boldsymbol{h}$ from Eq. 1 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 150 time steps per batch, a hidden dimension of 20, and other model parameters similar to the runs with the MVAE-TF. Hyperparameter settings were chosen such as to approximately keep the number of total parameters similar to that used for MVAE-TF.

Dataset	M	B	K	τ, L	λ_{MS}	ν
Lorenz	20	15	15	10	1.0	10
Rössler	20	15	15	10	1.0	15
Lewis-Glass	20	15	15	20	1.0	20

Table 2: Hyperparameter settings for MVAE-TF, GVAE-TF and MS trained on the Lorenz, Rössler and Lewis-Glass model.

Dataset	Method	$D_{stsp}\downarrow$	OPE $\downarrow$	SCC $\downarrow$	OACF $\downarrow$
Lorenz	MVAE-TF	$\mathbf{8.8 \pm 0.59}$	$\mathbf{0.24 \pm 0.015}$	$\mathbf{0.085 \pm 0.02}$	$\mathbf{0.016 \pm 0.04}$
	SVAE	14.7 ± 0.7	0.8 ± 0.03	0.17 ± 0.02	0.23 ± 0.02
	MS	13.8 ± 1.1	X	0.24 ± 0.06	0.15 ± 0.03
Rössler	MVAE-TF	$\mathbf{7.9 \pm 0.8}$	$\mathbf{0.093 \pm 0.007}$	$\mathbf{0.051 \pm 0.009}$	$\mathbf{0.051 \pm 0.009}$
	SVAE	11.5 ± 1.3	0.39 ± 0.02	0.23 ± 0.05	0.18 ± 0.04
	MS	14.1 ± 1.0	X	0.12 ± 0.04	0.14 ± 0.03
Lewis-Glass	MVAE-TF	$\mathbf{0.35 \pm 0.05}$	$\mathbf{0.15 \pm 0.02}$	$\mathbf{0.28 \pm 0.05}$	$\mathbf{0.15 \pm 0.03}$
	SVAE	0.55 ± 0.07	0.29 ± 0.01	0.49 ± 0.04	0.24 ± 0.02
	MS	$\mathbf{0.35 \pm 0.02}$	X	0.51 ± 0.04	0.45 ± 0.03

Table 3: Comparison of dendPLRNN trained by MVAE-TF (proposed method), by a SVAE based on (Kramer et al. 2022), and a multiple-shooting (MS) approach, on 8 ordinal observations with seven ordered categories, produced by the chaotic Lorenz system, Rössler system, and Lewis-Glass model. Values are mean $\pm$ SEM, averaged over 15 trained models. X = value cannot be computed for this model (e.g., because resp. decoder model is not present).

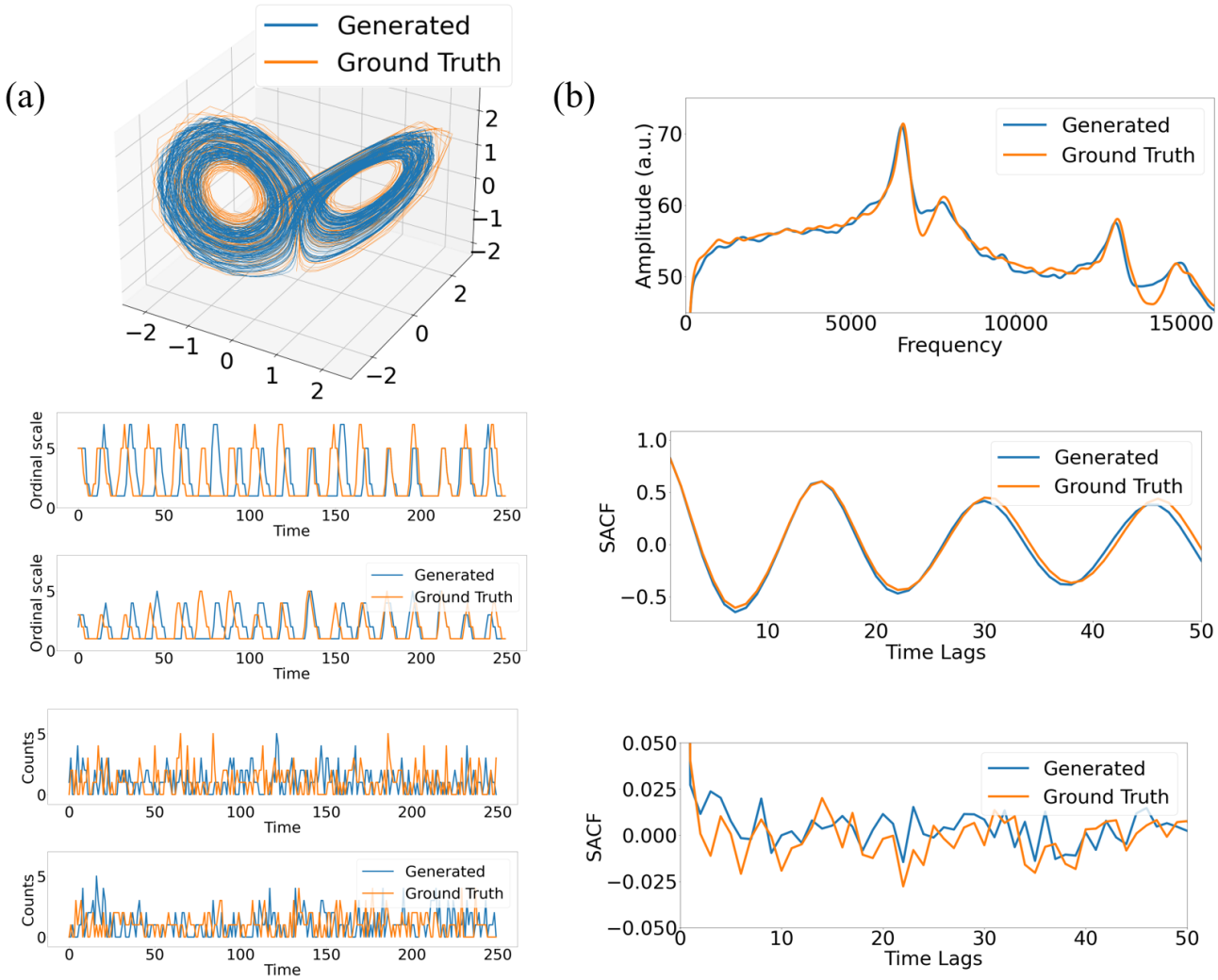

Figure 9: a) Freely generated example trajectories and time series from a dendPLRNN ($M = 20, K = 15, \tau = 10, B = 10$) trained with MVAE-TF jointly on Gaussian, ordinal, and count data sampled from a Lorenz-63 system. b) Example power spectra (Gaussian data) and Spearman autocorrelation functions (ordinal and count data). 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	$D_{stsp}\downarrow$	$D_H\downarrow$	PE $\downarrow$	OPE $\downarrow$	SCC $\downarrow$	OACF $\downarrow$	CACF $\downarrow$
Lorenz	MVAE-TF	$\mathbf{1.1 \pm 0.2}$	$\mathbf{0.16 \pm 0.04}$	$\mathbf{3.9\text{e}{-3} \pm 2.2\text{e}{-4}}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.042 \pm 0.002}$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{4.6\text{e}{-5} \pm 1.4\text{e}{-6}}$
	SVAE	6.7 ± 0.7	0.87 ± 0.05	$3.4\text{e}{-1} \pm 2.1\text{e}{-2}$	0.46 ± 0.03	0.14 ± 0.01	0.13 ± 0.03	$9.5\text{e}{-5} \pm 1.2\text{e}{-5}$
	GVAE-TF	1.94 ± 0.28	0.40 ± 0.08	$2.4\text{e}{-1} \pm 1.2\text{e}{-3}$	X	X	X	X
	BPTT-TF	5.3 ± 0.7	0.45 ± 0.05	$4.4\text{e}{-1} \pm 2.2\text{e}{-2}$	X	X	X	X
	MS	2.44 ± 0.25	0.34 ± 0.03	X	X	0.051 ± 0.04	0.064 ± 0.01	$6.5\text{e}{-5} \pm 3.8\text{e}{-6}$
Rössler	MVAE-TF	$\mathbf{1.01 \pm 0.31}$	$\mathbf{0.20 \pm 0.02}$	$\mathbf{4.8\text{e}{-4} \pm 2.4\text{e}{-5}}$	$\mathbf{0.05 \pm 0.04}$	$\mathbf{0.027 \pm 0.004}$	$\mathbf{0.006 \pm 0.001}$	$\mathbf{4.4\text{e}{-5} \pm 2.3\text{e}{-6}}$
	SVAE	9.7 ± 1.5	0.69 ± 0.05	$1.2\text{e}{-1} \pm 3.4\text{e}{-2}$	0.20 ± 0.06	0.12 ± 0.06	0.10 ± 0.03	$1.1\text{e}{-4} \pm 2.3\text{e}{-5}$
	GVAE-TF	10.1 ± 0.74	0.55 ± 0.06	$3.4\text{e}{-2} \pm 2.3\text{e}{-3}$	X	X	X	X
	BPTT-TF	8.1 ± 1.1	0.64 ± 0.07	$1.8\text{e}{-1} \pm 1.8\text{e}{-3}$	X	X	X	X
	MS	4.21 ± 0.68	0.51 ± 0.03	X	X	0.08 ± 0.04	0.05 ± 0.01	$7.8\text{e}{-5} \pm 8.5\text{e}{-6}$
Lewis-Glass	MVAE-TF	$\mathbf{0.24 \pm 0.16}$	$\mathbf{0.35 \pm 0.03}$	$\mathbf{1.8\text{e}{-3} \pm 8\text{e}{-5}}$	$\mathbf{0.1 \pm 0.01}$	0.12 ± 0.03	$\mathbf{0.04 \pm 0.03}$	$9.3\text{e}{-4} \pm 4.5\text{e}{-4}$
	SVAE	2.8 ± 1.3	0.53 ± 0.05	$6.1\text{e}{-2} \pm 6\text{e}{-3}$	0.23 ± 0.01	0.42 ± 0.04	0.26 ± 0.03	$1.5\text{e}{-2} \pm 8.1\text{e}{-3}$
	GVAE-TF	$\mathbf{0.26 \pm 0.1}$	0.39 ± 0.02	$2.6\text{e}{-3} \pm 5\text{e}{-4}$	X	X	X	X
	BPTT-TF	1.53 ± 0.31	0.41 ± 0.03	$2.4\text{e}{-2} \pm 3\text{e}{-3}$	X	X	X	X
	MS	$\mathbf{0.27 \pm 0.06}$	0.37 ± 0.01	X	X	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.03 \pm 0.02}$	$\mathbf{1.7\text{e}{-4} \pm 7.9\text{e}{-6}}$

Table 4: Comparison of dendPLRNN trained by MVAE-TF (proposed method), by a SVAE based on Kramer et al. (2022), an VAE-TF approach similar to MVAE-TF except that all data modalities were 'Gaussianized' (GVAE-TF), BPTT-TF as in Brenner et al. (2022) using Gaussianized data, and a multiple-shooting (MS) approach. Training was performed on multivariate normal, ordinal, and count data produced by the chaotic Lorenz system, Rössler system, and Lewis-Glass model. Observation noise with 1% of the data variance was added to the Gaussian observations. Values are mean $\pm$ SEM, averaged over 15 trained models. X = value cannot be computed for this model (e.g., because resp. decoder model is not present).