Modeling Neural Activity with Conditionally Linear Dynamical Systems

Victor Geadah [1 2] **Amin Nejatbakhsh** [2] **David Lipshutz** [2 3] **Jonathan W. Pillow** [1 4] **Alex H. Williams** [2 5]

Abstract

Neural population activity exhibits complex, nonlinear dynamics, varying in time, over trials, and across experimental conditions. Here, we develop *Conditionally Linear Dynamical System* (CLDS) models as a general-purpose method to characterize these dynamics. These models use Gaussian Process (GP) priors to capture the nonlinear dependence of circuit dynamics on task and behavioral variables. Conditioned on these covariates, the data is modeled with linear dynamics. This allows for transparent interpretation and tractable Bayesian inference. We find that CLDS models can perform well even in severely data-limited regimes (e.g. one trial per condition) due to their Bayesian formulation and ability to share statistical power across nearby task conditions. In example applications, we apply CLDS to model thalamic neurons that nonlinearly encode heading direction and to model motor cortical neurons during a cued reaching task.

1. Introduction

A central problem in neuroscience is to capture how neural dynamics are affected by external sensory stimuli, task variables, and behavioral covariates. To address this, a longstanding line of research has focused on characterizing neural dynamics through recurrent neural networks (RNNs) and their probabilistic counterparts, state-space models (SSMs; for reviews, see Paninski et al. 2010; Duncker & Sahani 2021; Durstewitz et al. 2023).

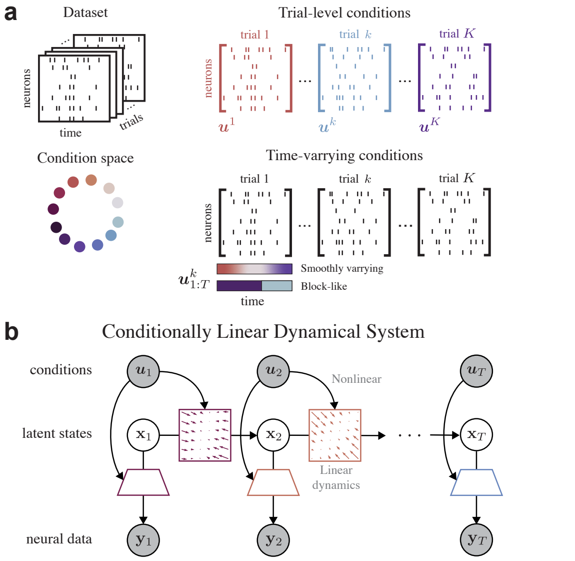

Figure 1. (**a**) Neural dataset consisting of spike trains collected over multiple trials, along with corresponding experimental conditions. (**b**) Conditionally Linear Dynamical Systems are *linear* in state-space dynamics and capture *nonlinear* dependencies over conditions. Shaded nodes are observed, clear nodes are latent.

Early work in this area utilized latent linear dynamical systems (LDS) with Gaussian observation noise. Although these assumptions are restrictive, they are beneficial in two respects. First, they simplify probabilistic inference by enabling Kalman smoothing and expectation maximization (EM)—two classical and highly effective methods (see e.g., Ghahramani & Hinton 1996). Second, they produce models that are mathematically tractable to analyze with well-established tools from linear systems theory (Kailath, 1980). Indeed, many influential results in theoretical neuroscience have come from purely linear models (e.g. Seung 1996; Goldman 2009; Murphy & Miller 2009).

In reality, most neural circuits do not behave like time-invariant linear systems. Thus, more recent work from the machine learning community has cataloged a variety of nonlinear models for neural dynamics. Although these new models often predict held out neural data more accurately than LDS models, they are generally more difficult to fit and more difficult to understand. Thus, there has been a proliferation of competing architectures—e.g. RNNs (e.g. Pandarinath et al. 2018), transformers (e.g. Ye & Pandarinath 2021), and diffusion-based methods (Kapoor et al., 2024)—as well as competing training and probabilistic in-

[1] Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ. [2] Center for Computational Neuroscience, Flatiron Institute, New York City, NY. [3] Department of Neuroscience, Baylor College of Medicine, Houston, TX. [4] Princeton Neuroscience Institute, Princeton, NJ. [5] Center for Neural Science, New York University, New York City, NY. Correspondence to: Victor Geadah <victor.geadah@princeton.edu>, Alex H. Williams <alex.h.williams@nyu.edu>.

Preprint.

ference methods—e.g. generalized teacher forcing (Hess et al., 2023), amortized variational inference (e.g. Pandarinath et al. 2018), and sequential Monte Carlo (e.g. Pals et al. 2024). Choosing among these strategies and scientifically interpreting the outcomes is challenging.

Here we describe *Conditionally Linear Dynamical Systems* (CLDS) as a framework to jointly capture some of the benefits of the classical (i.e. linear) and contemporary (i.e. nonlinear) approaches to modeling neural data. CLDS models parametrize a collection of LDS models that vary smoothly as a function of an observed variable $\boldsymbol{u}_t$ (e.g. measured sensory input or behavior at time t). Assuming the presence of $\boldsymbol{u}_t$ is often a feature—not a bug—of this approach. Indeed, a common goal in neuroscience is to relate measured sensory or behavioral covariates to neural activity. Additional features of the CLDS framework include:

1. CLDS models are locally interpretable. Conditioned on $\boldsymbol{u}_t$, the dynamics are linear and amenable to a number of classical analyses.
2. CLDS models are easy to fit (§2.3). If Gaussian noise is assumed, then exact latent variable inference (via Kalman smoothing) and fast optimization (via closed-form EM) is possible. Under more realistic noise models (e.g. Poisson), the posterior over latent state trajectories is still log concave and amenable to relatively fast and simple inference routines.
3. CLDS models are expressive. As $\boldsymbol{u}_t$ changes, the parameters of the linear system are allowed to change *non-linearly*. Thus, CLDS can model complex dynamical structures such as ring attractors (§3.2), that are impossible for a vanilla LDS to capture.
4. CLDS models are data efficient. To the extent that LDS parameters change smoothly as a function of $\boldsymbol{u}_t$, we can the recover the parameters of the dynamical system with very few trials per condition (§3.4). In fact, CLDS models can interpolate to make accurate predictions on entirely unseen conditions.

Finally, CLDS models have connections to several existing methods (§4). For example, they can be viewed as a dynamical extension of a Wishart process model (Nejatbakhsh et al., 2023) and an extension of Gaussian Process Factor Analysis (GPFA; Yu et al. 2009) with a learnable kernel and readout function that can vary across time and conditions. They are also similar to various forms of switching linear dynamical systems models (Petreska et al., 2011; Nassar et al., 2019; Hu et al., 2024). The key difference is that the "switching" in CLDS models is governed by an observed covariate vector, $\boldsymbol{u}_t$, rather than by a discrete latent process. This makes inference in the CLDS model much more straightforward, albeit at the price of not being fully unsupervised.

2. Methods

Notation We use $\mathbf{f}(\cdot) \sim \mathcal{GP}^N(\boldsymbol{m}(\cdot), k(\cdot,\cdot))$ with mean $\boldsymbol{m} : \mathcal{X} \to \mathbb{R}^N$ and kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, to denote samples $\mathbf{f} : \mathcal{X} \to \mathbb{R}^N$ obtained from stacking independent Gaussian processes into an N-dimensional vector.

2.1. Conditionally Linear Dynamical Systems

Consider an experiment with N neurons recorded over K trials of length T. Our dataset consists of the recorded neural trajectories $\{\mathbf{y}^k_{1:T}\}^K_{k=1}$, with $\mathbf{y}_t \in \mathbb{R}^N$ at time-step $t \in \{1, \ldots, T\}$, along with the corresponding experimental conditions $\{\boldsymbol{u}^k_{1:T}\}^K_{k=1}$, with $\boldsymbol{u}_t$ in the condition space $\mathcal{U}$. By experimental conditions, we refer to available neural data covariates, either experimentally set or collected as measurements. These conditions, see Fig. 1**a**, can vary over time or remain constant (compare §3.3 vs. §3.4). They can also be a step function over time (e.g. animal moving vs. not moving), resulting in a switching-like mechanism between different dynamics.

We model response $\mathbf{y}_t$ as emissions from a latent time-varying linear dynamical system in $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^D$, with the dynamics governed by the conditions $\boldsymbol{u}_t$. Specifically,

$$\mathbf{x}_{t+1} = \mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t + \mathbf{b}(\boldsymbol{u}_t) + \boldsymbol{\epsilon}_t \qquad (1a)$$

$$\mathbf{y}_t = \mathbf{C}(\boldsymbol{u}_t)\mathbf{x}_t + \mathbf{d}(\boldsymbol{u}_t) + \boldsymbol{\omega}_t \qquad (1b)$$

evolving with time steps $t \in \{1, \ldots, T\}$ from initial condition $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{x}_1; \mathbf{m}(\boldsymbol{u}_1), Q_1)$, and where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, Q)$ and $\boldsymbol{\omega}_t \sim \mathcal{N}(0, R)$ are sources of noise, both sampled i.i.d. over time. We assume that the latent variables $\mathbf{x}_t$ follow smooth dynamics defined by time-varying linear matrices $\mathbf{A}(\boldsymbol{u}) \in \mathbb{R}^{D\times D}$ from initial mean $\mathbf{m}(\boldsymbol{u}_1) \in \mathbb{R}^D$, with bias terms $\mathbf{b}(\boldsymbol{u}) \in \mathbb{R}^D$, $\mathbf{d}(\boldsymbol{u}) \in \mathbb{R}^N$, and emissions governed by $\mathbf{C}(\boldsymbol{u}) \in \mathbb{R}^{N\times D}$. See graphical depiction in Fig. 1**b**.

Conditions are typically treated as additive *inputs*, influencing the dynamics in (1a) via additive terms of the form $\mathbf{B}\boldsymbol{u}_t$ for a linear encoding matrix $\mathbf{B} \in \mathbb{R}^{D\times|\mathcal{U}|}$. Instead, the system in (1) parameterizes a *family of linear* systems indexed by a continuous time-varying variable, $\boldsymbol{u}_t$, which is observed. In fact, the system in (1) can be thought of as the linearization in $\mathbf{x}_t$ of a fully nonlinear system in $\mathbf{x}_t$ and $\boldsymbol{u}_t$, under additive noise—we explore this relationship in Appendix §A.2. The mapping of experimental conditions onto linear dynamics, $\boldsymbol{u} \mapsto \{\mathbf{A}(\boldsymbol{u}), \mathbf{b}(\boldsymbol{u}), \mathbf{C}(\boldsymbol{u}), \mathbf{d}(\boldsymbol{u}), \mathbf{m}(\boldsymbol{u})\}$, is allowed to be nonlinear and learnable. Specifically, we place an approximate Gaussian Process (GP) prior on each entry of the parameters through a finite expansion of basis functions, leveraging regular Fourier feature approximations (Hensman et al., 2018). For any $\mathbf{M} \in \{\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{d}, \mathbf{m}\}$, we consider a prior over each i, j-th entry of the form:

$$\mathbf{M}_{ij}(\boldsymbol{u}) = \sum_{\ell=1}^{L} \mathrm{w}_\ell^{(ij)} \phi_\ell(\boldsymbol{u}), \quad \mathrm{w}_\ell^{(ij)} \overset{iid}{\sim} \mathcal{N}(0,1), \qquad (1c)$$

truncated at $L \in \mathbb{N}$ basis functions. Intuitively, each basis function $\phi_\ell : \mathcal{U} \to \mathbb{R}$ is fixed, and the randomness in the prior purely comes from the weights, $\mathrm{w}_\ell^{(ij)}$, which are drawn from a standard normal distribution. When constructed appropriately, the prior in equation (1c) converges to a nonparametric Gaussian process in the limit that $L \to \infty$. In our experiments, the basis functions $\{\phi_\ell\}_{\ell=1}^L$ are chosen and scaled as to approximate a GP prior of the form $\mathbf{M}_{ij}(\cdot) \sim \mathcal{GP}(0, k_u)$ for the squared exponential kernel k_u with variance σ^2 and length-scale κ (Borovitskiy et al., 2020).

We denote $\mathbf{F} = \{\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{d}, \mathbf{m}\}$ as the set of random functions, and analogously the parameter set $\mathbf{F}(\boldsymbol{u}) = \{\mathbf{A}(\boldsymbol{u}), \mathbf{b}(\boldsymbol{u}), \mathbf{C}(\boldsymbol{u}), \mathbf{d}(\boldsymbol{u}), \mathbf{m}(\boldsymbol{u})\}$ for any experimental condition $\boldsymbol{u} \in \mathcal{U}$. The model distribution

$$p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T} \mid \mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{m}, \boldsymbol{u}_{1:T}) = \\ p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T} \mid \mathbf{F}, \boldsymbol{u}_{1:T}) \quad (2)$$

describes a time-varying latent Linear Dynamical System, conditioned on a parameter sequence set at experimental conditions. Therefore, we refer to the model in equations (1) as a *Conditionally Linear Dynamical System* (CLDS). Our CLDS implementation is available at:

`https://github.com/neurostatslab/clds`

2.2. CLDS modeling choices

Practitioners can adapt a CLDS model in several ways to suit different applications and modeling assumptions. First, the GP prior can be tuned to trade off model expressivity for interpretability and learnability. In one extreme, as we let $\kappa \to 0$, the LDS parameters change rapidly, nonlinearly as a function of $\boldsymbol{u}$ and become independent per $\boldsymbol{u}$. In the other extreme, if one takes $\kappa \to \infty$, then the LDS parameters become constant (do not change as a function of $\boldsymbol{u}$) and we recover a time-invariant LDS model with autonomous dynamics. In this regime, we could also modify the GP prior over $\mathbf{b}(\cdot)$ to follow a linear kernel, $k(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{u}^\top \boldsymbol{u}'$, resulting in time-invariant LDS with additive dependence on $\boldsymbol{u}_t$. Thus, CLDS models capture classical linear models as a special case. Moreover, the model's prior can be tuned to capture progressively nonlinear dynamics.

A second source of flexibility is the encoding of experimental covariates, $\boldsymbol{u}$. Recall our notation from section 2.1, that $\boldsymbol{u}_t^k$ represents experimental covariates at time $t \in \{1, \ldots, T\}$ and trial $k \in \{1, \ldots, K\}$. A simple, and broadly applicable, modeling approach would be to set $\boldsymbol{u}_t^k = t$. This achieves a time-varying LDS model in which the GP prior encodes smoothness over time. This is similar in concept to fitting an linear model to data over a sliding time window (see e.g. Costa et al. 2019; Galgali et al. 2023). However, the CLDS formulation of this idea is fully probabilistic, which has several advantages. For example, we will see that one can use a single pass of Kalman smoothing to infer the distribution over the latent state trajectory, $\mathbf{x}_{1:T}^k$, within each trial. It is comparatively non-trivial to average latent state trajectories across multiple LDS models that are independently fit to data in overlapping time windows.

In section 3, we demonstrate more sophisticated examples where $\boldsymbol{u}_t^k$ is specified to track a continuously measured behavioral variable (e.g. heading direction or position of an animal) or follow a stepping or ramping function aligned to discrete task events (e.g. a sensory "go cue" or movement onset). Section 4 discusses further connections between CLDS models and existing state space models.

2.3. Inference

As mentioned earlier, the conditional distribution in eq. (2) has the advantage of describing a latent linear dynamical system (LDS), or linear Gaussian state-space model. As such, we can benefit from analytic tools like Kalman filtering to compute the filtering distributions $p(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \mathbf{F}, \boldsymbol{u}_{1:T})$ and marginal log-likelihood $p(\mathbf{y}_{1:T} \mid \mathbf{F}, \boldsymbol{u}_{1:T})$, and Kalman smoothing to compute the smoothing posterior $p(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T}, \mathbf{F}, \boldsymbol{u}_{1:T})$. We focus on performing maximum-a-posteriori (MAP) inference for these parameters. In principle, it would be a straightforward extension to use variational inference or Markov Chain Monte Carlo to approximate the full posterior over these parameters.

Conditionally Linear Regression As a stepping stone towards our goal of performing MAP inference for $\{\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{d}, \mathbf{m}\}$, consider the model

$$\mathbf{y}_n = \mathbf{M}(\boldsymbol{u}_n)\mathbf{x}_n + \boldsymbol{\epsilon}_n, \quad \boldsymbol{\epsilon}_n \sim \mathcal{N}(0, \Sigma) \quad (3a)$$

$$\mathbf{M}(\cdot) \sim \mathcal{GP}^{D_1 \times D_2}(0, k_u), \quad (3b)$$

given data $\mathbf{y}_n \in \mathbb{R}^{D_1}$, regressors $\mathbf{x}_n \in \mathbb{R}^{D_2}$, conditions $\boldsymbol{u}_n \in \mathcal{U}$, repeats $n \in \{1, \ldots, N\}$, noise covariance $\Sigma \succ 0$, and with an approximate GP prior on $\mathbf{M}$ parametrized as in (1c). We refer to the model (3) as *conditionally linear regression*, and our goal is to perform MAP inference for the weights $\{\mathrm{w}_k^{(ij)}\}_{i,j,k}$ for $\mathbf{M}$.

Our parameterization in eq. (1c) implies that each entry is a dot product, $\mathbf{M}_{ij}(\cdot) = \boldsymbol{w}^{(ij)\top}\boldsymbol{\phi}(\cdot)$, where $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_L(\cdot))^\top \in \mathbb{R}^L$ is our vector of basis-functions evaluations. Therefore,

$$\mathbf{M}(\boldsymbol{u}) = \mathbf{W}^\top(\boldsymbol{\phi}(\boldsymbol{u}) \otimes \boldsymbol{I}_{D_2}) \quad (4)$$

$$\mathbf{M}(\boldsymbol{u})\boldsymbol{X} = \mathbf{W}^\top(\boldsymbol{\phi}(\boldsymbol{u}) \otimes \boldsymbol{X}), \quad (5)$$

for $\boldsymbol{u} \in \mathcal{U}$ and for any vector or matrix $\boldsymbol{X} \in \mathbb{R}^{D_2, \cdots}$ of appropriate dimension, with "⊗" the Kronecker product, and with our weights aggregated into the matrix

$$\mathbf{W}_{j+\ell,i} := \mathrm{w}_\ell^{(ij)}, \quad \mathbf{W} \in \mathbb{R}^{D_2 L \times D_1}. \quad (6)$$

With this, we can rewrite our regression problem as

$$\mathbf{y}_n = \mathbf{M}(\boldsymbol{u}_n)\mathbf{x}_n + \boldsymbol{\epsilon}_n = \mathbf{W}^\top \mathbf{z}_n + \boldsymbol{\epsilon}_n \tag{7}$$

with $\mathbf{z}_n := \boldsymbol{\phi}(\boldsymbol{u}_n) \otimes \mathbf{x}_n \in \mathbb{R}^{D_2 L}$. Thus, we have reformulated our original model into Bayesian linear regression in an expanded feature space. Namely, the MAP estimate of the weights, $\mathbf{W}_{\text{MAP}}$, is given by

$$\arg\max_{\mathbf{W}} \log p(\mathbf{y}_{1:N} \mid \mathbf{W}, \mathbf{x}_{1:N}, \boldsymbol{u}_{1:N}) + \log p(\mathbf{W}) \tag{8}$$

which is a linear regression problem with regularization from the prior $\log p(\mathbf{W}) = -\frac{1}{2}\|\mathbf{W}\|_F^2$ (up to an additive constant). We can analytically solve for the solution (derivations in Appendix §A.1.1), which yields that $\mathbf{W}_{\text{MAP}}$ is the solution to the Sylvester equation

$$\boldsymbol{Z}^\top \boldsymbol{Z}\, \mathbf{W} + \mathbf{W}\Sigma = \boldsymbol{Z}^\top \boldsymbol{Y} \tag{9}$$

with $\boldsymbol{Z} \in \mathbb{R}^{N \times D_2 L}$ our matrix obtained by stacking $\{\mathbf{z}_n\}_{n=1}^N$, and similarly for $\boldsymbol{Y} \in \mathbb{R}^{N \times D_1}$. We see that if $\Sigma = \sigma^2 \boldsymbol{I}_{D_1}$ for some $\sigma > 0$ then we obtain back the familiar looking penalized least squares estimate $\mathbf{W}_{\text{MAP}} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \sigma^2 \boldsymbol{I}_{D_1})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}$.

Expectation Maximization We can leverage the above to perform MAP inference for $\{\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{d}, \mathbf{m}\}$ with the *Expectation-Maximization* (EM) algorithm (Dempster et al., 1977; Ghahramani & Hinton, 1996). In the *E*-step we obtain estimates of the moments of the latents with Kalman-smoothing, which then place us in a setting akin to eq. (3) with sufficient statistics as data and regressors. We can then perform closed-form M-steps with our updates in eq. (9). We provide in Appendix §A.1.2 an example of the associated derivations with these E- and M-steps for the joint update for $\mathbf{A}(\cdot)$ and $\mathbf{b}(\cdot)$. The resulting EM algorithm has several advantages: (1) all E- and M-steps are analytic, (2) the E-step provides us with exact (penalized) marginal log-likelihood calculations, and (3) the algorithm gives monotonic gradient ascent guarantees of the marginal log-likelihood (resp. log posterior) objective.

We initialize the EM algorithm at samples from our GP priors for the parameters. With the EM algorithm we also learn the covariance parameters $\{Q_1, Q, R\}$. Finally, the hyper-parameters $\{L, \kappa, \sigma\}$ from the GP priors are determined through performance on held-out test sets from 80/20 trial splits on all experiments.

Extensions to non-Gaussian likelihoods The closed form EM updates are only applicable when the distribution of $\mathbf{y}_t$ conditioned on $\mathbf{x}_t$ and $\boldsymbol{u}_t$ is Gaussian. This assumption, stated in eq. (1), is common existing methods (e.g. Yu et al. 2009; Williams et al. 2018). However, alternative models are likely a better fit to many neural datasets. For spike count data, past work has utilized Poisson (e.g. Macke et al. 2011) and COM-Poisson (Stevenson, 2016) likelihoods.

Although we assume a Gaussian likelihood in our experiments in §3, we note that inference in CLDS models remains tractable whenever $p(\mathbf{y}_t \mid \mathbf{W}, \mathbf{x}_t, \boldsymbol{u}_t)$ is log concave. This condition satisfied by most likelihood models of interest (e.g. Poisson). Indeed, conditioned on $\boldsymbol{u}_{1:T}$, the log posterior density associated with $\mathbf{x}_{1:T}$ is equal to a sum of concave terms up to an additive constant (Paninski et al., 2010). Thus, we can use standard optimization routines to identify a MAP estimate of $\mathbf{x}_{1:T}$ efficiently with theoretical guarantees. This can then be used to implement an approximate EM algorithm (Macke et al., 2011).

3. Experiments

3.1. Setup

Metrics For a given trajectory $\{\mathbf{y}_{1:T}, \boldsymbol{u}_{1:T}\}$, we denote as *data reconstruction* the mean emission $\mathbb{E}\left[\mathbf{y}_{1:T} \mid \hat{\mathbf{x}}_{1:T}, \mathbf{F}(\boldsymbol{u}_{1:T})\right] = (\mathbf{C}(\boldsymbol{u}_1)\hat{\mathbf{x}}_1, \ldots, \mathbf{C}(\boldsymbol{u}_T)\hat{\mathbf{x}}_T)$ from a the posterior mode $\hat{\mathbf{x}}_{1:T}$, computed with Kalman smoothing, given the observations $\mathbf{y}_{1:T}$ and parameters $\mathbf{F}(\boldsymbol{u}_{1:T})$. As our primary metric, we use *co-smoothing* (Pei et al., 2021) to evaluate the ability of models to predict held-out single-neuron activity. Specifically, for the top 5 neurons with highest variance from the test set, we compute the coefficient of determination R^2 between the true and reconstructed single-neuron firing rate, obtained by performing data reconstruction using only the other neurons.

Composite dynamics The latent dynamical system (1a) depends on the condition $\boldsymbol{u}_t$, which can make visualizations challenging. Building on the idea that CLDS models decompose a nonlinear dynamical system into linearizations governed by $\boldsymbol{u}$ (see Appendix §A.2), we aim to approximate the general nonlinear system by marginalizing out $\boldsymbol{u}_t$, conditioned on $\mathbf{x}_t$. That is,

$$\begin{aligned} \mathbf{x}_{t+1} &= \boldsymbol{g}(\mathbf{x}_t) + \epsilon_t \\ &:= \mathbb{E}_{p(\boldsymbol{u}\mid\mathbf{x}_t)}\left[\mathbf{A}(\boldsymbol{u})\mathbf{x}_t + \mathbf{b}(\boldsymbol{u})\right] + \epsilon_t, \end{aligned} \tag{10}$$

which we define as the *composite dynamical system*. Intuitively, we expect this to provide a good approximation to the underlying nonlinear dynamics when $\boldsymbol{u}_t$ and $\mathbf{x}_t$ tightly co-determine each other—i.e., when the encoding $p(\mathbf{x}_t \mid \boldsymbol{u}_t)$ and decoding $p(\mathbf{u}_t \mid \mathbf{x}_t)$ conditional distributions have low variance (see Appendix §A.2.2). In practice, we estimate the expectations in (10) by computing the empirical average over $\boldsymbol{u}$ per binned $\mathbf{x}_n$, obtained by pooling the $\boldsymbol{u}_n$ associated with the posterior mode $\hat{\mathbf{x}}_n$ given a trajectory $\{\mathbf{y}_{1:T}, \boldsymbol{u}_{1:T}\}$.

Model baselines For model comparison, we use as baselines (1) a standard **LDS** model with additive inputs of the

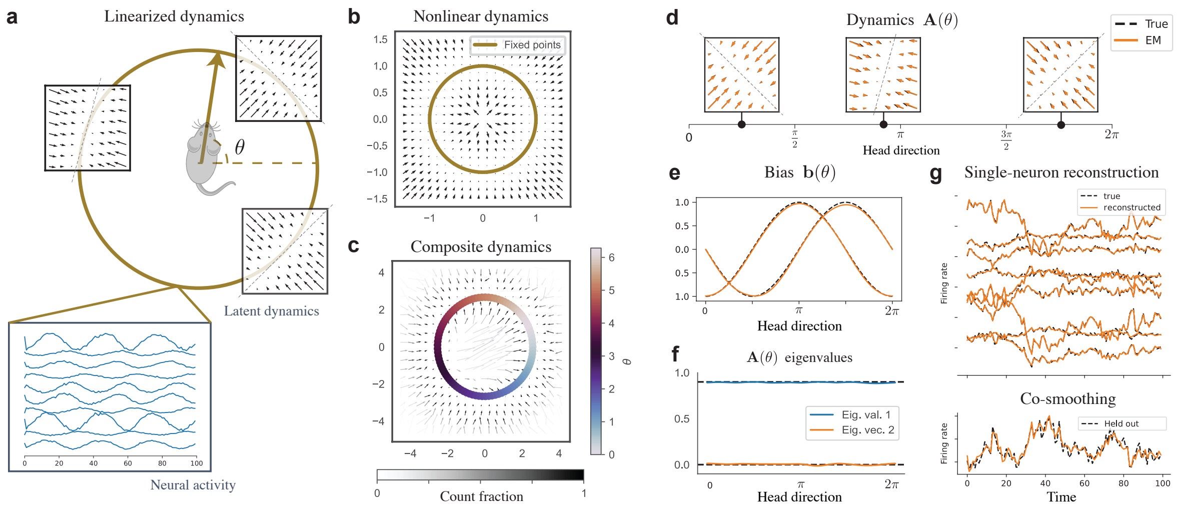

Figure 2. Head direction synthetic experiment. (**a**) Schematic of latent dynamics and neural activity about $\theta \in [0, 2\pi)$, the mouse HD, serving as conditions $\boldsymbol{u} = \theta$ in this task. (**b**) True nonlinear flow-field corresponding to the schematic in **a**, computed considering $p(\theta \mid \mathbf{x}) = \delta_{\angle \mathbf{x}}(\theta)$. (**c**) Recovered composite dynamics by CLDS, see text for computation details. Grey scale indicates $\mathbf{x}_t$ occupancy. The model fixed points (colored) as a function of θ form a perfect ring, overlapping with the true fixed points. (**d**-**e**) Parameter recovery for the dynamics matrix $\mathbf{A}(\theta)$ (**d**) and the bias $\mathbf{b}(\theta)$ (**e**) as functions of head direction θ. (**f**) Recovered eigenvalues of $\mathbf{A}(\theta)$ as a function of θ, true in dashed. (**g**) Co-smoothing reconstruction from the test-set. The firing rate of one neuron is held-out (bottom) while the rest (top) is observed, and we reconstruct accurately the single-neuron firing rates for both the held-in (top) and held-out (bottom) neurons.

form $\boldsymbol{B}\boldsymbol{u}_t$ in the latent dynamics, and (2) the **LFADS** (Pandarinath et al., 2018) model with controller inferred-inputs, with Gaussian observation model to fit directly to the firing rates. See Appendix §B.1 for implementation details.

3.2. Synthetic Head-Direction Ring Attractor

We start by considering a synthetic experiment of head direction neural dynamics. We conceptualize latent dynamics that capture the head direction (HD) of the animal, with attractor dynamics about a HD-dependent fixed point—see schematic in Fig. 2**a**. This synthetic experiment is designed to represent a nonlinear system decomposed as linear systems, per HD serving as the condition. We plot in Fig. 2**b** what the resulting, “composite dynamics” (see §3.1), nonlinear flow-field would be, assuming the latent state encodes the head direction exactly. The generative dynamics are an instance of a CLDS model by construction, so this synthetic example allows us to explore recovery performance.

Concretely, let $\theta_t \in [0, 2\pi)$ denote heading direction at time step t, which we treat as our conditions $\boldsymbol{u}_t := \theta_t$. To build a ring attractor, we parametrize two orthogonal unit vectors

$$\boldsymbol{e}_1(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \quad \boldsymbol{e}_2(\theta) = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}, \tag{11}$$

that describe the position on the ring and the tangent vector respectively. We design (i.e. impose) that the system converges to a stable fixed point at $\boldsymbol{e}_1(\theta)$, and at head direction θ we approximately integrate speed input along the subspace spanned by $\boldsymbol{e}_2(\theta)$. To do this, we define $\boldsymbol{A}(\theta)$ to be a rank-one matrix that defines a leaky line attractor, with attracting (i.e. contracting) dynamics along the orthogonal $\boldsymbol{e}_1(\theta)$. For a hyperparameter $0 < \epsilon < 1$ define:

$$\boldsymbol{A}(\theta) := (1 - \epsilon)\boldsymbol{e}_2(\theta)\boldsymbol{e}_2(\theta)^\top, \quad \boldsymbol{b}(\theta) := \boldsymbol{e}_1(\theta). \tag{12}$$

Completing the model description, we assume that the firing rate of individual neurons is given by a linear readout. For neuron i, the firing rate at time t is:

$$\mathbf{y}_{t,i} = \boldsymbol{C}_{i,:}(\theta_t)^\top \mathbf{x}_t + \boldsymbol{\omega}_t \tag{13}$$

where $\boldsymbol{C}_{i,:}$ is sinusoidal bump tuning curve function (see App. §B.2). Note that we set $\mathbf{d}(\boldsymbol{u}_t) = \mathbf{0}$. Finally, we sampled trials of length $T = 100$ with $M = 10$ neurons, generating the evolution of the heading direction as a random walk, $\theta_t \sim \mathcal{N}(\theta_{t-1}, 0.5^2)$, and initialize at the origin $\mathbf{x}_1 \sim \mathcal{N}(0, 1)$.

We report our recovery results in Fig. 2, fixing the decoding matrix $\mathbf{C}(\cdot)$ to a known value as to avoid non-identifiability considerations. We refer to Appendix §B.2 for recovery plots of $\mathbf{C}$ when fitted. First, we observed that we can generally recover the nonlinear flow-field, plotting in Fig. 2**c** the composite dynamics obtained from the posterior trajectories. This paints an activity-based portrait of the dynamics, and our ability to accurately estimate the flow-field around a given point under this method depends on how many posterior samples pass by it. We thus indicate, by gray-scale shading the flow-field arrows, the fraction of posterior samples pooled per bin as a fraction of the highest bin count.

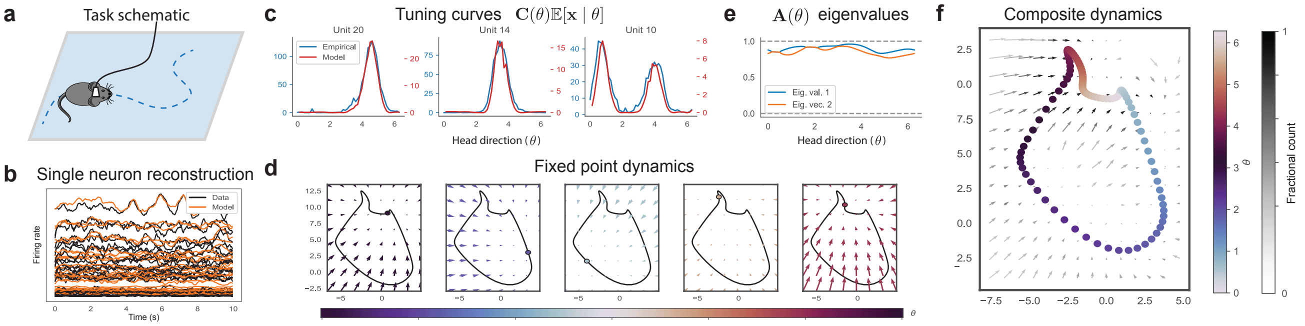

Figure 3. (**a**) Schematic of mouse foraging in an open environment. We have access to $\theta_t \in [0, 2\pi)$ the mouse HD in time t, which we use as conditions $\boldsymbol{u}_t$ just like Fig. 2. (**b**) Model reconstruction on the whole dataset recovers the true data. We plot single-neuron traces, averaged over 10s trials. (**c**) Model tuning curves over head direction θ, obtained as $\mathbf{C}(\theta)\mathbb{E}[\mathbf{x} \mid \theta]$, recover the empirical tuning curves. Plotted for the top three units in firing rate norm. (**d**) Dynamics around each fixed point in $\mathbf{x}$-state space as a function of head direction θ, with the solid-line representing the complete set of fixed points. (**e**) Eigenvalues and angles of eigenvectors of $\mathbf{A}(\theta)$ as a function of θ. (**f**) Composite dynamics in $\mathbf{x}_t$-space, with overlaid colored model fixed points as a function of θ.

Second, for a more parameter-based account of the recovery, we plot in Fig. 2**d**-**e** the varying biases $\mathbf{b}(\theta)$ and dynamics matrices $\mathbf{A}(\theta)$ as functions of the head direction θ—we recovered with high-fidelity the true parameters. This recovery translated into the properties of the dynamics such as the recovered eigenvalues of $\mathbf{A}(\theta)$ in Fig. 2**f**. Finally, we observed that the test data single-neuron reconstruction (Fig. 2**g**) recovers the true observations, and the model was able to accurately ($R^2 = 0.86$) reconstruct a held out neuron from this test-set through co-smoothing.

3.3. Mouse Head-Direction Circuit Dynamics

Next, we turned to the analysis of antero-dorsal thalamic nucleus (ADn) recordings from Peyrache et al. (2015) of the mouse HD system in mice foraging in an open environment (Fig. 3**a**). We considered neural activity from the “wake” period, binned in 50ms time-bins, then processed to firing rates and separated into 10s trials. As with the synthetic experiment of the previous section, we treat the recorded head-direction $\theta_t \in [0, 2\pi)$ as conditions $\boldsymbol{u}_t = \theta_t$.

We recovered single-neuron firing rates with high accuracy (Fig. 3**b**) through data reconstruction. We further validated our fit by computing the empirical tuning curves, which our model recovered almost exactly (Fig. 3**c**). The model tuning curves are given by

$$\mathbb{E}\left[\mathbf{y}_i \mid \theta\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{y}_i \mid \mathbf{x}, \theta\right]\right] = \boldsymbol{C}_{i,:}(\theta)^\top \mathbb{E}\left[\mathbf{x} \mid \theta\right], \quad (14)$$

which follows from the law of total expectation. The later quantity $\mathbb{E}[\mathbf{x} \mid \theta]$ represents the expected encoding of the conditions θ, which we estimate by averaging posterior trajectories, obtained with Kalman smoothing, over (binned) θ given corresponding firing rates.

Finally, we analyzed the learned latent dynamics. Like the synthetic example, we identified a ring attractor structure (Fig. 3**d**). Now, unlike the synthetic example, we observed that this ring attractor is composed of HD-dependent fixed points as opposed to line attractors, as per the eigenvalues of $\mathbf{A}(\theta)$ in Fig. 3**e**.

3.4. Macaque Center-Out Reaching Task

Finally, we analyzed neural recordings of dorsal premotor cortex (PMd) in macaques performing center-out reaching task (Fig. 4**a**) from N. Even-Chen, B. Sheffer et al. (2019). In contrast to the previous experimental conditions, we consider here two-dimensional conditions $\boldsymbol{u}_t^k = (\theta^k, z_t)$, where $\theta^k \in [0, 2\pi)$ is the instructed reach angle, constant per trial k, and $z_t \in \{0, 1\}$ indicates the task reach condition (see Fig. 4**b**) set at 0 during the delay and 1 at 100ms past the go-cue, at the onset of the movement-related firing rate ramp Fig. 4**a**-(right). Discrete-valued conditions, such as the reach onset $z_t \in \{0, 1\}$, are considered as supported on a continuous interval. The correlation between such discrete points is determined by the length-scale parameter κ, which we’ve set to $\kappa = 0.5$ from a hyperparameter search. More details on hyperparameters and data-preprocessing can be found in Appendix §B. Finally, we use a fixed emission matrix $\boldsymbol{C}$ and let the latent dynamics capture the dependency on experimental conditions through $\mathbf{A}(\boldsymbol{u})$ and $\mathbf{b}(\boldsymbol{u})$.

We found the latent dynamics to encode the conditions through attracting fixed-points during both the delay and reach periods. We show in Fig. 4**c** the projection of the $D = 5$ latent dimensions along the 3-dimensional subspace most aligned (i.e. best decoding) with the experimental conditions, following similar analyses from (N. Even-Chen, B. Sheffer et al., 2019)—we observed clear aligned rings of fixed points from delay to reach. In CLDS models, we obtain the fixed points by simply solving for $\boldsymbol{x}^*(\boldsymbol{u})$ satisfying $(\boldsymbol{I} - \mathbf{A}(\boldsymbol{u}))\boldsymbol{x}^* = \mathbf{b}(\boldsymbol{u})$ for any $\boldsymbol{u}$, in contrast to numerical fixed-point methods usually employed (Sussillo & Barak, 2013).

We performed co-smoothing (see §3.1) to evaluate the model. We recovered with good accuracy single held-out neurons from the validation set excluded from training (Fig. 4**d**). We then compared the performance of the CLDS against the LDS and LFADS models, exploring further how each fares in low-data regimes. We report in Fig. 4**e** the co-smoothing R^2 per model, computed as a function of the number of trials used in each reach-angle θ^k, averaged over 5 random seeds. We found that the CLDS outperformed both models consistently, with the highest difference at 1 training trial per condition. While the LFADS model showed progressively better performance that did not plateau yet, it nonetheless underperformed in these low data regimes.

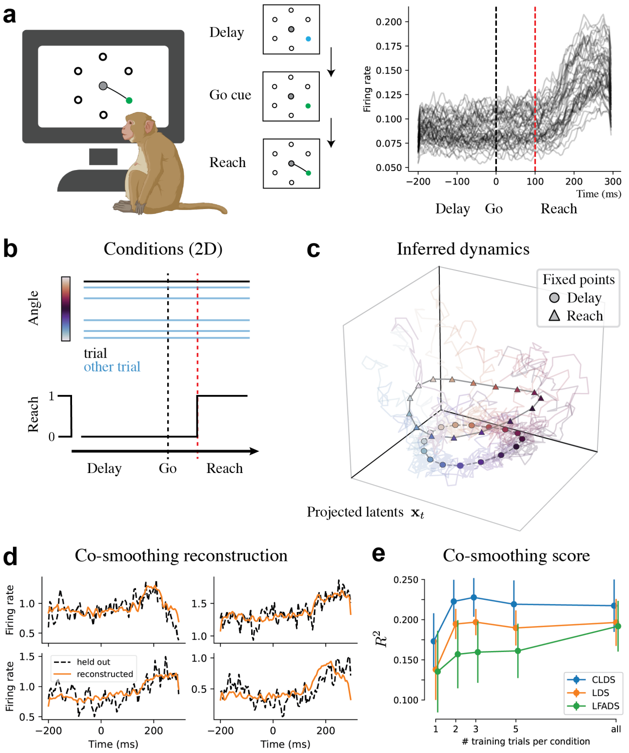

Figure 4. Macaque reaching experiment. (**a**) Task schematic (left) and population-averaged firing rates per trial (right). (**b**) 2D conditions, with trial orientation θ, and reach variable $z_t \in \{0, 1\}$ switching at ramp onset. (**c**) 3D projection of the 5 dimensional latents used, projected as to align best with condition decoding. We show the model fixed-points per reach angle θ and reach condition z, plotted over posterior mean trajectories per trial. (**d**) Co-smoothing reconstruction of single held-out neurons from the test-set. (**e**) Co-smoothing R^2 per model as a function of the number of trials used per reach angle during training. Error bars indicated std. around the mean over 5 random initialization seeds.

4. Related work

Wishart Process Models CLDS models capture the dependence of neural responses $\mathbf{y}_{1:T}$ on continuous experimental conditions $\boldsymbol{u}_{1:T}$. Nejatbakhsh et al. (2023) investigated a very similar problem, focusing on single-trial responses $\mathbf{y}_k$ to continuous experimental conditions $\boldsymbol{u}_k \in \mathcal{U}$. They use a conditional Gaussian model for responses $\mathbf{y}_k$ given conditions $\boldsymbol{u}_k$

$$\mathbf{y}_k \mid \boldsymbol{u}_k \sim \mathcal{N}(\mathbf{y}; \mu(\boldsymbol{u}_k), \Sigma(\boldsymbol{u}_k)) \tag{15}$$

and they place Gaussian process and Wishart process (Wilson & Ghahramani, 2011) priors on the mean and covariance functions. Concretely, they posit that

$$\mu(\cdot) \sim \mathcal{GP}^M(0, k_\mu), \tag{16a}$$

$$\Sigma(\boldsymbol{u}) = \mathbf{U}(\boldsymbol{u})\mathbf{U}(\boldsymbol{u})^\top + \Lambda(\boldsymbol{u}) \tag{16b}$$

with $\mathbf{U}(\cdot) \sim \mathcal{GP}^{M\times p}(0, k_\Sigma)$ and $\Lambda(\cdot) \sim \mathcal{GP}^M(0, k_\Lambda)$ for chosen kernel functions $\{k_\mu, k_\Sigma, k_\Lambda\}$. The hyper-parameter $p \in \mathbb{N}$ determines the low-rank structure of Σ.

For a single time step $T = 1$, $t = 1$, our system in (1) reads

$$\begin{aligned} p(\mathbf{x}_1 \mid \boldsymbol{u}_1) &= \mathcal{N}(\mathbf{x}_1; \mathbf{m}(\boldsymbol{u}_1), Q_1) \\ p(\mathbf{y}_1 \mid \mathbf{x}_1, \boldsymbol{u}_1) &= \mathcal{N}\left(\mathbf{y}_1; \mathbf{C}(\boldsymbol{u}_1)\mathbf{x}_1 + \mathbf{d}(\boldsymbol{u}_1), R\right), \end{aligned}$$

Assuming a degenerate prior that $\mathbf{m}(\boldsymbol{u}_1) = \mathbf{0}$, the marginal distribution of $\mathbf{y}_1$ conditioned on $\boldsymbol{u}_1$ equals

$$\mathcal{N}\left(\mathbf{d}(\boldsymbol{u}_1), \mathbf{C}(\boldsymbol{u}_1)Q_1\mathbf{C}(\boldsymbol{u}_1)^\top + R\right). \tag{17}$$

which can be compared with (15) and (16). We observe that the models are essentially equivalent with $\mu(\boldsymbol{u}) = \mathbf{d}(\boldsymbol{u})$, and with the CLDS emission matrix $\mathbf{C}(\boldsymbol{u})$ serving as the Wishart process prior decomposition matrix $\mathbf{U}(\boldsymbol{u})$, right-scaled by $Q_1^{1/2} \in \mathbb{R}^{D\times D}$. This makes the parameter $p = D$ now bear meaning as the dimensionality of the latents $\mathbf{x} \in \mathbb{R}^D$. Thus, we can view CLDS models as a direct extension of Wishart process models that capture condition-dependent dynamics across multiple time steps.

Markovian GPs Latent GP models (Lawrence, 2007; Wang et al., 2005), such as the foundational model of GPFA by Yu et al. (2009), are widely used in neuroscience. Here we show that one can view MAP inference in a CLDS model as optimizing a kernel that defines a latent GP prior. While GPFA is not a dynamical system model, Yu et al. (2009), as well as Turner & Sahani (2007), detail how the the stationary dynamics of an AR-1 process (i.e. linear dynamical system) can be expressed as draws from a GP. More generally, all stationary, real-valued, and finitely differentiable GPs admit a representation in terms of linear state space models (Dowling et al., 2021; 2023). The main departure with our work is that we additionally place a GP prior on the *parameters* (coefficients) of an LDS, allowing the dynamics to vary across conditions and across time. For a fixed set of LDS parameters, $\boldsymbol{F}$, and experimental covariates, $\boldsymbol{u}_{1:T}$, the distribution of latent states in a CLDS are jointly Gaussian. Thus, a set of LDS parameters induces a (generally non-stationary) GP prior on the latent trajectories. In this view, the GP prior we place over the parameters of the LDS can be seen as a hyperprior over the latent dynamical process.

Switching Dynamical Systems A second class of relevant models generalizing the LDS are *Switching LDS* models (SLDS; Murphy 1998; Pavlovic et al. 2000; Petreska et al. 2011). SLDS models consist of a discrete latent state z_t with Markov chain dynamics dictating the dynamics matrix $\boldsymbol{A}^{(\mathrm{z}_t)}$. This switching behavior can be mimicked in our setting if the condition space is discrete (see, e.g., §3.4). We can take the relationship a step further by embedding the discrete process in the continuous dynamics parameter space of $\boldsymbol{A}^{(\mathrm{z}_t)} \in \mathbb{R}^{D\times D}$. Under this lens and in a similar line of thinking as with Markovian GPs, we show in Appendix §A.3 how a one-dimensional SLDS model with latent dynamics

$$p(\mathrm{z}_{t+1} = i \mid \mathrm{z}_t = j) = P_{ij}, \quad \mathrm{x}_{t+1} = a^{(\mathrm{z}_t)}\mathrm{x}_t + \epsilon_t$$

is equivalent, up to the first two moments of the stochastic process $\mathrm{a} := a^{\mathrm{z}}$, to a CLDS model with

$$\mathrm{a}(\cdot) \sim \mathcal{GP}\left(\pi^\top \boldsymbol{a}, \boldsymbol{a}^\top \left(P^{|t_j - t_i|}\mathrm{diag}(\pi) - \pi\pi^\top\right)\boldsymbol{a}\right),$$

over time conditions $\boldsymbol{u}_t = t$, for $\boldsymbol{a}$ the vector of values taken by $a^{(z_i)}$ and π the stationary distribution of the z_t discrete state process. Finally, in a similar vein, Geadah et al. (2024) consider the discrete states z_t dictating the dynamics to live on a continuous support. However, they do not leverage this continuity in the parameters $\mathbf{A}(\cdot)$ themselves.

The *recurrent SLDS* (rSLDS) model (Linderman et al., 2017) takes an important departure from the SLDS by leveraging the continuous latent states $\mathbf{x}_t$ to guide the discrete state transitions. Smith et al. (2021) use this dependency but turn to the linearization of nonlinear systems, using $\mathbf{x}$-space fixed points as guide for the linear dynamics. In contrast, we linearize based on observed external conditions.

Smoothly varying dynamical systems models Switching LDS models can be contrasted with models that smoothly interpolate between dynamical parameter regimes. The simplest example of this would be time-varying linear models (e.g. Costa et al. 2019); CLDS models are a generalization of this idea that comes with several advantages (see §2.2). Work by Costacurta et al. (2022) introduced an autoregressive model with a dynamics matrix that is subject to an approximately continuous and latent time warping factor. Unlike CLDS, this model does not infer a low-dimensional latent dynamical space. More similar to CLDS models is recent work by Hu et al. (2024). They relax the discrete switching in rSLDS models to allow smoothly varying soft mixtures of linear dynamics. Again, CLDS models achieve a similar effect but utilize observed experimental covariates to infer these dynamical transitions. Thus, CLDS models can loosely be seen as supervised analogs to these models.

5. Conclusion

In this work, we revisited and extended classical linear-Gaussian state space models of neural circuit dynamics. Our results suggest that these models can be competitive with modern methods when the dynamical parameters vary smoothly as a function of experimentally measured covariates. Like classical linear models, CLDS models are easy to fit and interpret. Our main technical contribution was to introduce an approximate GP prior over system parameters and show that this leads to closed form inference and parameter updates under a Gaussian noise model.

As their name implies, CLDS models assume conditionally linear latent dynamics, and this assumption brings some limitations. First, these models rely on observing a time series of experimental covariates, $\boldsymbol{u}_{1:T}^k$. We expect performance to suffer if the covariates are corrupted for portions of time, such as during forecasting or with partial observations. Second, the model assumes linear dynamics conditioned on $\boldsymbol{u}_t^k$. We believe this is a good approximation in many settings of interest, particularly when there is strong tuning to sensory or behavioral variables—i.e. when the value of $\boldsymbol{u}_t^k$ can be used to accurately predict the position of $\boldsymbol{x}_t^k$. We expect (see §A.2) CLDS models to struggle in other settings where external measurements are only loosely correlated with the position of $\mathbf{x}_t^k$ (e.g. cognitive tasks with long periods of internal deliberation). In these scenarios, we expect that modern approaches that leverage deep learning (e.g. LFADS) will outperform CLDS models when given access to large amounts of data. Nevertheless, neural recordings are often trial-limited in practice (Williams & Linderman, 2021). We therefore view CLDS models as a broadly applicable modeling tool for many neuroscience applications.

Future work could extend CLDS models to overcome these limitations, such as handling partially observed covariates, $\boldsymbol{u}_t^k$. Since CLDS models can be viewed as a dynamical extension of Wishart process models (see §4), future work could also apply this method to infer across-time noise correlations (reviewed in Panzeri et al., 2022), in addition to classical across-trial noise correlations. Nejatbakhsh et al. (2024) show how across-time correlations can be used to quantify similarity in dynamical systems—a topic that has recently attracted strong interest (Ostrow et al., 2023). CLDS models are a potentially attractive framework for tackling the unsolved challenge of estimating this high-dimensional correlation structure in trial-limited regimes.

Acknowledgments

VG was supported by doctoral scholarships from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de recherche du Québec – Nature et technologies (FRQNT). JWP was supported by grants from the Simons Collaboration on the Global Brain (SCGB AWD543027), the NIH BRAIN initiative (9R01DA056404-04), an NIH R01 (NIH 1R01EY033064), and a U19 NIH-NINDS BRAIN Initiative Award (U19NS104648). AHW was supported by the NIH BRAIN initiative (1RF1MH133778).

Impact Statement

The modeling presented here aims to provide a methodology to enhance our understanding of neural computation. Analysis of electrophysiological data can have long-term implications for the treatment and understanding of medical treatment and neurological disorders across different species. However, these considerations are far removed for the preliminary analyses and theoretical modeling presented, and we foresee no immediate societal consequences of this work.

References

Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. Matérn gaussian processes on riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Costa, A. C., Ahamed, T., and Stephens, G. J. Adaptive, locally linear models of complex dynamics. *Proceedings of the National Academy of Sciences*, 116(5):1501–1510, 2019.

Costacurta, J., Duncker, L., Sheffer, B., Gillis, W., Weinreb, C., Markowitz, J., Datta, S. R., Williams, A., and Linderman, S. Distinguishing discrete and continuous behavioral variability using warped autoregressive hmms. *Advances in neural information processing systems*, 35: 23838–23850, 2022.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):1–22, September 1977.

Dowling, M., Sokół, P., and Park, I. M. Hida-mat\'ern kernel. *arXiv preprint arXiv:2107.07098*, 2021.

Dowling, M., Zhao, Y., and Park, I. M. Linear time GPs for inferring latent trajectories from neural spike trains. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Duncker, L. and Sahani, M. Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. *Current opinion in neurobiology*, 70:163–170, 2021.

Durstewitz, D., Koppe, G., and Thurm, M. I. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nature Reviews Neuroscience*, 24(11):693–710, Nov 2023.

Galgali, A. R., Sahani, M., and Mante, V. Residual dynamics resolves recurrent contributions to neural computation. *Nature Neuroscience*, 26(2):326–338, 2023.

Geadah, V., Laboratory, I. B., and Pillow, J. W. Parsing neural dynamics with infinite recurrent switching linear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024.

Ghahramani, Z. and Hinton, G. Parameter estimation for linear dynamical systems. Technical report, University of Toronto, 1996. Tech. Rep. CRG-TR-96-2. Available as https://mlg.eng.cam.ac.uk/zoubin/course04/tr-96-2.pdf.

Goldman, M. S. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

Hensman, J., Durrande, N., and Solin, A. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Hess, F., Monfared, Z., Brenner, M., and Durstewitz, D. Generalized teacher forcing for learning chaotic dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Hu, A., Zoltowski, D., Nair, A., Anderson, D., Duncker, L., and Linderman, S. Modeling latent neural dynamics with gaussian process switching linear dynamical systems, 2024.

Kailath, T. *Linear systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

Kapoor, J., Schulz, A., Vetter, J., Pei, F. C., Gao, R., and Macke, J. H. Latent diffusion for neural spiking data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lawrence, N. D. Learning for larger datasets with the gaussian process latent variable model. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 243–250. PMLR, 2007.

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., and Sahani, M. Empirical models of spiking in neural populations. *Advances in neural information processing systems*, 24, 2011.

Murphy, B. K. and Miller, K. D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

Murphy, K. Learning switching kalman filter models. Tech Report 98-10, Compaq Cambridge Research Lab, Cambridge, MA, 1998.

N. Even-Chen, B. Sheffer, Vyas, S., Ryu, S. I., and Shenoy, K. V. Structure and variability of delay activity in premotor cortex. *PLOS Computational Biology*, 15(2): e1006808, February 2019.

Nassar, J., Linderman, S., Bugallo, M., and Park, I. M. Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. In *International Conference on Learning Representations*, 2019.

Nejatbakhsh, A., Garon, I., and Williams, A. Estimating noise correlations across continuous conditions with wishart processes. In *Advances in Neural Information Processing Systems*, 2023.

Nejatbakhsh, A., Geadah, V., Williams, A. H., and Lipshutz, D. Comparing noisy neural population dynamics using optimal transport distances. *arXiv preprint arXiv:2412.14421*, 2024.

Ostrow, M., Eisen, A., Kozachkov, L., and Fiete, I. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In *Advances in Neural Information Processing Systems*, 2023.

Pals, M., Sağtekin, A. E., Pei, F. C., Gloeckler, M., and Macke, J. H. Inferring stochastic low-rank recurrent neural networks from neural data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, September 2018.

Paninski, L., Ahmadian, Y., Ferreira, D. G., Koyama, S., Rahnama Rad, K., Vidne, M., Vogelstein, J., and Wu, W. A new look at state-space models for neural data. *Journal of computational neuroscience*, 29:107–126, 2010.

Panzeri, S., Moroni, M., Safaai, H., and Harvey, C. D. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience*, 23(9):551–567, 2022.

Pavlovic, V., Rehg, J. M., and MacCormick, J. Learning switching linear models of human motion. In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Pei, F., Ye, J., Zoltowski, D. M., Wu, A., Chowdhury, R. H., Sohn, H., O'Doherty, J. E., Shenoy, K. V., Kaufman, M. T., Churchland, M., Jazayeri, M., Miller, L. E., Pillow, J., Park, I. M., Dyer, E. L., and Pandarinath, C. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2021.

Petreska, B., Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S., Shenoy, K. V., and Sahani, M. Dynamical segmentation of single trials from population neural data. In *Advances in Neural Information Processing Systems*, 2011.

Peyrache, A., Lacroix, M. M., Petersen, P. C., and Buzsáki, G. Internally organized mechanisms of the head direction sense. *Nature Neuroscience*, 18(4):569–575, March 2015.

Seung, H. S. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23): 13339–13344, 1996.

Smith, J., Linderman, S., and Sussillo, D. Reverse engineering recurrent neural networks with jacobian switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, 2021.

Stevenson, I. H. Flexible models for spike count data with both over-and under-dispersion. *Journal of computational neuroscience*, 41:29–43, 2016.

Sussillo, D. and Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

Turner, R. and Sahani, M. A maximum-likelihood interpretation for slow feature analysis. *Neural Computation*, 19 (4):1022–1038, April 2007.

Wang, J., Hertzmann, A., and Fleet, D. J. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Williams, A. H. and Linderman, S. W. Statistical neuroscience in the single trial limit. *Current Opinion in Neurobiology*, 70:193–205, 2021.

Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., Schnitzer, M., Kolda, T. G., and Ganguli, S. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115, 2018.

Wilson, A. G. and Ghahramani, Z. Generalised wishart processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 736–744, 2011.

Ye, J. and Pandarinath, C. Representation learning for neural population activity with neural data transformers. *arXiv preprint arXiv:2108.01210*, 2021.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102 (1):614–635, July 2009.

A. Modeling

A.1. Expectation-Maximization Steps

A.1.1. Least squares derivation

Recall

$$\mathbf{M}(\boldsymbol{u})\boldsymbol{X} = \mathbf{W}^\top(\boldsymbol{\phi}(\boldsymbol{u}) \otimes \boldsymbol{X}), \qquad \mathbf{W} \in \mathbb{R}^{D_2 L \times D_1}$$

for $\mathbf{M}(\boldsymbol{u}) \in \mathbb{R}^{D_1 \times D_2}$, $\boldsymbol{X} \in \mathbb{R}^{D_2 \times M}$. In particular, $\mathbf{M}(\boldsymbol{u}_n)\mathbf{x}_n = \mathbf{W}^\top \mathbf{z}_n$. What follows are standard least-squares derivations for matrix coefficients with matrix regularization, which we include for completeness.

Our posterior objective reads as

$$\begin{aligned}\log p(\mathbf{W} \mid \mathbf{y}_{1:N}, \mathbf{x}_{1:N}, \boldsymbol{u}_{1:N}) &\propto \log p(\mathbf{y}_{1:N} \mid \mathbf{W}, \mathbf{x}_{1:N}, \boldsymbol{u}_{1:N}) + \log p(\mathbf{W}) \\ &= \sum_{n=1}^{N} \log p(\mathbf{y}_n \mid \mathbf{W}, \mathbf{x}_n, \boldsymbol{u}_n) + \log p(\mathbf{W}).\end{aligned}$$

We have

$$\begin{aligned}\log p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{u}_t) &= -\tfrac{1}{2}(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n)^\top \Sigma^{-1}(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n) - c \\ &= -\tfrac{1}{2}\mathrm{Tr}\left[(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n)^\top \Sigma^{-1}(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n)\right] - c \\ &= -\tfrac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n)(\mathbf{y}_n - \mathbf{W}^\top \mathbf{z}_n)^\top\right] - c \\ &= -\tfrac{1}{2}\left(\mathrm{Tr}\left[\Sigma^{-1}\mathbf{y}_n\mathbf{y}_n^\top\right] - 2\mathrm{Tr}\left[\Sigma^{-1}\mathbf{W}^\top \mathbf{z}_n \mathbf{y}_n^\top\right] + \mathrm{Tr}\left[\Sigma^{-1}\mathbf{W}^\top \mathbf{z}_n \mathbf{z}_n^\top \mathbf{W}\right]\right) - c.\end{aligned}$$

with the normalizing constant $c = \frac{1}{2}\log|2\pi\Sigma|$.

To optimize this expression with respect to $\mathbf{W}$, we consider the zeros of the derivative

$$\begin{aligned}\frac{\partial}{\partial \mathbf{W}} \log p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{u}_n) &= \frac{\partial}{\partial \mathbf{W}}\mathrm{Tr}\left[\mathbf{W}^\top \mathbf{z}_n \mathbf{y}_n^\top \Sigma^{-1}\right] - \frac{1}{2}\frac{\partial}{\partial \mathbf{W}}\mathrm{Tr}\left[\Sigma^{-1}\mathbf{W}^\top \mathbf{z}_n \mathbf{z}_n^\top \mathbf{W}\right] \\ &= \mathbf{z}_n \mathbf{y}_n^\top \Sigma^{-1} - \mathbf{z}_n \mathbf{z}_n^\top \mathbf{W}\Sigma^{-1},\end{aligned}$$

and

$$\log p(\mathbf{W}) = -\tfrac{1}{2}\left\|\mathbf{W}\right\|_F^2 \quad \Longrightarrow \quad \tfrac{\partial}{\partial \mathbf{W}}\log p(\mathbf{W}) = -\mathbf{W}.$$

Taken together, we thus have that the stationary point of the posterior satisfies

$$\sum_{n=1}^{N}\left(\mathbf{z}_n \mathbf{y}_n^\top \Sigma^{-1} - \mathbf{z}_n \mathbf{z}_n^\top \mathbf{W}\Sigma^{-1}\right) - \mathbf{W} = 0.$$

Define $Y \in \mathbb{R}^{N \times D_1}$, $Z \in \mathbb{R}^{N \times D_2 L}$ by row-wise stacking $\mathbf{y}_n$ and $\mathbf{z}_n$ respectively, and note that $\sum_n \mathbf{z}_n \mathbf{y}_n^\top = Z^\top Y$. We get

$$\begin{aligned}Z^\top Y \Sigma^{-1} - Z^\top Z \mathbf{W}\Sigma^{-1} - \mathbf{W} = 0 \\ \Longrightarrow \quad Z^\top Z\mathbf{W} + \mathbf{W}\Sigma = Z^\top Y\end{aligned}$$

as desired.

A.1.2. Joint dynamics and bias EM update in weight-space

Here we detail how one EM update for the parameters governing $\{\mathbf{A}, \mathbf{b}\}$ is carried out. Using the function-space weights $\mathbf{W}_A \in \mathbb{R}^{DL \times D}$ and $\mathbf{W}_b \in \mathbb{R}^{L \times D}$, the dynamics read

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t + \mathbf{b}(\boldsymbol{u}_t) + \epsilon_t \\ &= \mathbf{W}_A^\top \underbrace{(\boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbf{x}_t)}_{\mathbf{z}_t} + \mathbf{W}_b^\top \boldsymbol{\phi}(\boldsymbol{u}_t) + \epsilon_t\end{aligned}$$

Define $\mathbf{z}_t = \boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbf{x}_t \in \mathbb{R}^{LD}$, and note

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_t}\left[\mathbf{z}_t\right] &= \boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbb{E}_{\mathbf{x}_t}\left[x_t\right] \\
\mathbb{E}_{\mathbf{x}_t,\mathbf{x}_{t+1}}\left[\mathbf{z}_t\mathbf{x}_{t+1}^\top\right] &= \boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbb{E}_{\mathbf{x}_t,\mathbf{x}_{t+1}}\left[\mathbf{x}_t\mathbf{x}_{t+1}^\top\right]
\end{aligned}$$

The quantity of interest for the M-step from the complete data log-likelihood is

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T-1} \log p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{F}, \boldsymbol{u}_t)\right] \\
&= -\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T-1} \left(\mathbf{x}_{t+1} - (\mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t + \mathbf{b}(\boldsymbol{u}_t))\right)^\top Q^{-1}\left(\mathbf{x}_{t+1} - (\mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t + \mathbf{b}(\boldsymbol{u}_t))\right)\right] \\
&= -\frac{1}{2}\mathbb{E}\Bigg[\sum_{t=1}^{T-1} \mathbf{x}_{t+1}^\top Q^{-1}\mathbf{x}_{t+1} - 2\mathbf{x}_{t+1}^\top Q^{-1}\mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t - 2\mathbf{x}_{t+1}^\top Q^{-1}\mathbf{b}(\boldsymbol{u}_t) \\
&\qquad\qquad + \mathbf{x}_t^\top \mathbf{A}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{A}(\boldsymbol{u}_t)\mathbf{x}_t + 2\mathbf{x}_t^\top \mathbf{A}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{b}(\boldsymbol{u}_t) + \mathbf{b}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{b}(\boldsymbol{u}_t)\Bigg] \\
&= -\frac{1}{2}\mathrm{Tr}\Bigg[\sum_{t=1}^{T-1} Q^{-1}\mathbb{E}[\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\top] - 2Q^{-1}\mathbf{A}(\boldsymbol{u}_t)\mathbb{E}[\mathbf{x}_t\mathbf{x}_{t+1}^\top] - 2Q^{-1}\mathbf{b}(\boldsymbol{u}_t)\mathbb{E}[\mathbf{x}_{t+1}^\top] \\
&\qquad\qquad + \mathbf{A}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{A}(\boldsymbol{u}_t)\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] + 2\mathbf{A}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{b}(\boldsymbol{u}_t)\mathbb{E}[\mathbf{x}_t^\top] + \mathbf{b}(\boldsymbol{u}_t)^\top Q^{-1}\mathbf{b}(\boldsymbol{u}_t)\Bigg]
\end{aligned}$$

Which with the function-space weights reads as

$$\begin{aligned}
\mathcal{L} = -\frac{1}{2}\mathrm{Tr}\Bigg[&\sum_{t=1}^{T-1} Q^{-1}\mathbb{E}[\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\top] - 2Q^{-1}\mathbf{W}_A^\top(\boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_{t+1}^\top]) - 2Q^{-1}\mathbf{W}_b^\top(\boldsymbol{\phi}(\boldsymbol{u}_t)\mathbb{E}[\mathbf{x}_{t+1}^\top]) \\
&+ \mathbf{W}_A Q^{-1}\mathbf{W}_A^\top(\boldsymbol{\phi}(\boldsymbol{u}_t)\boldsymbol{\phi}(\boldsymbol{u}_t)^\top \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]) + 2\mathbf{W}_A Q^{-1}\mathbf{W}_b^\top(\boldsymbol{\phi}(\boldsymbol{u}_t)\boldsymbol{\phi}(\boldsymbol{u}_t)^\top \otimes \mathbb{E}[\mathbf{x}_t^\top]) \\
&+ \mathbf{W}_b Q^{-1}\mathbf{W}_b^\top\boldsymbol{\phi}(\boldsymbol{u}_t)\boldsymbol{\phi}(\boldsymbol{u}_t)^\top\Bigg]
\end{aligned}$$

The partial derivatives satisfy, denoting $\boldsymbol{\phi}_t = \boldsymbol{\phi}(\boldsymbol{u}_t)$,

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_A} &= \sum_{t=1}^{T-1}(\boldsymbol{\phi}_t \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_{t+1}^\top])Q^{-1} - (\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top])\mathbf{W}_A Q^{-1} - (\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top \otimes \mathbb{E}[\mathbf{x}_t]^\top)^\top\mathbf{W}_b Q^{-1} \\
&=: N_\Delta Q^{-1} - N_{(1,T-1)}\mathbf{W}_A Q^{-1} - (\Phi^\top Z)^\top\mathbf{W}_b Q^{-1} \qquad (18) \\
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_b} &= \sum_{t=1}^{T-1}\boldsymbol{\phi}_t\mathbb{E}[\mathbf{x}_{t+1}^\top]Q^{-1} - (\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top \otimes \mathbb{E}[\mathbf{x}_t^\top])\mathbf{W}_A Q^{-1} + \boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\mathbf{W}_b Q^{-1} - \mathbf{W}_b \\
&=: \Phi^\top X Q^{-1} - \Phi^\top Z\mathbf{W}_A Q^{-1} + \Phi^\top\Phi\mathbf{W}_b Q^{-1} \qquad (19)
\end{aligned}$$

where we've defined the matrices $\Phi \in \mathbb{R}^{(T-1)\times L}$, $Z \in \mathbb{R}^{(T-1)\times DL}$, $X \in \mathbb{R}^{(T-1)\times D}$ obtained by stacking $\boldsymbol{\phi}(\boldsymbol{u}_t)$, $\boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbb{E}[\mathbf{x}_t]$ and $\mathbb{E}[\mathbf{x}_{t+1}]$ respectively for $t \in \{1, \ldots, T-1\}$, and the sufficient statistics

$$N_{(t_1,t_2)} = \sum_{t_1}^{t_2}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top], \quad N_\Delta = \sum_{t=1}^{T-1}\boldsymbol{\phi}(\boldsymbol{u}_t) \otimes \mathbb{E}[\mathbf{x}_t\mathbf{x}_{t+1}^\top].$$

which are all defined during our E-step. These statistics are in contrast to the "typical" sufficient stats without the weight-space parametrization

$$M_{(t_1,t_2)} = \sum_{t_1}^{t_2}\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top], \quad M_\Delta = \sum_{t=1}^{T-1}\mathbb{E}[\mathbf{x}_t\mathbf{x}_{t+1}^\top].$$

Incorporating the Gaussian prior on $\mathbf{W}_A$ and $\mathbf{W}_b$ and equating the two partial derivatives in (18-19) to 0 to obtain the stationary points, we obtain the system of equations

$$\begin{bmatrix}\mathbf{W}_A\\ \mathbf{W}_b\end{bmatrix} Q + \begin{bmatrix}N_{(1,T-1)} & Z^\top \Phi\\ \Phi^\top Z & -\Phi^\top \Phi\end{bmatrix}\begin{bmatrix}\mathbf{W}_A\\ \mathbf{W}_b\end{bmatrix} = \begin{bmatrix}N_\Delta\\ \Phi^\top X\end{bmatrix} \tag{20}$$

which solving for $\mathbf{W}_A$ and $\mathbf{W}_b$ jointly amounts to solving our M-step.

A.2. Nonlinear dynamics: linearization and composite dynamics

Consider input-driven nonlinear dynamics in $\mathbf{x}_t \in \mathbb{R}^D$,

$$\mathbf{x}_{t+1} = \boldsymbol{f}(\mathbf{x}_t, \mathbf{u}_t) + \epsilon_t \tag{21}$$

governed by $\boldsymbol{f} : \mathbb{R}^D \times \mathcal{U} \to \mathbb{R}^D$, and where ϵ_t is zero-mean Gaussian noise. We assume that $\boldsymbol{f}$ has continuous and bounded second-order partial derivatives in both $\mathbf{x}$ and $\mathbf{u}$. Below we treat the state and input variables $(\mathbf{x}_t, \mathbf{u}_t)$ as random variables that are jointly drawn from some unspecified distribution.

A.2.1. CLDS CONDITIONAL APPROXIMATION ERROR

Let $f_1, \dots, f_D$ denote the output dimensions of $\boldsymbol{f}$; that is, $\boldsymbol{f}(\mathbf{x}_t, \mathbf{u}_t) = \begin{bmatrix} f_1(\mathbf{x}_t, \mathbf{u}_t) & \dots & f_D(\mathbf{x}_t, \mathbf{u}_t)\end{bmatrix}^\top$. In a first step to relate our CLDS dynamics to $\boldsymbol{f}$, consider the first-order Taylor expansion for each output dimension in the first argument $\boldsymbol{x}$ about $\boldsymbol{a} \in \mathbb{R}^D$. For $i \in \{1, \dots, D\}$, this is

$$f_i(\mathbf{x}_t, \mathbf{u}_t) = f_i(\boldsymbol{a}, \mathbf{u}_t) + \nabla_{\boldsymbol{x}} f_i(\boldsymbol{a}, \mathbf{u}_t)^\top (\mathbf{x}_t - \boldsymbol{a}) + \mathcal{E}_i \tag{22}$$

where $\nabla_{\boldsymbol{x}} f_i$ denotes the vector-valued gradient of f_i with respect to it's first argument $\boldsymbol{x}$ and $\mathcal{E}_i$ is the residual of the Taylor approximation. The Lagrange remainder form of Taylor's theorem tells us that this residual can be expressed as:

$$\mathcal{E}_i = (\mathbf{x}_t - \boldsymbol{a})^\top \nabla^2_{\boldsymbol{x}} f_i(\zeta, \mathbf{u}_t)(\mathbf{x}_t - \boldsymbol{a}) \tag{23}$$

for some $\zeta \in \mathbb{R}^D$, where $\nabla^2_{\boldsymbol{x}} f_i$ is the matrix-valued Hessian of f_i with respect to its first argument. We can upper bound the absolute value of this remainder using the Cauchy-Schwartz inequality and a standard operator norm inequality. Specifically, for $i \in \{1, \dots, D\}$, we have

$$|\mathcal{E}_i| \le \left\|\nabla^2_{\boldsymbol{x}} f_i(\zeta, \mathbf{u}_t)\right\|_2 \|\mathbf{x}_t - \boldsymbol{a}\|_2^2 \tag{24}$$

where $\left\|\nabla^2_{\boldsymbol{x}} f_i(\zeta, \mathbf{u}_t)\right\|_2$ denotes the maximal singular value (operator norm) of the matrix $\nabla^2_{\boldsymbol{x}} f_i(\zeta, \mathbf{u}_t) \in \mathbb{R}^{D \times D}$. We assume that this operator norm is upper bounded globally by a constant $L_i > 0$,

$$\left\|\nabla^2_{\boldsymbol{x}} f_i(\mathbf{x}, \mathbf{u})\right\|_2 \le L_i \qquad \forall\, \mathbf{x}, \mathbf{u} \in \mathbb{R}^D. \tag{25}$$

Intuitively, this assumption implies that the second-order derivatives of $\boldsymbol{f}$ with respect to $\mathbf{x}$ are not too large, meaning that the accuracy of the first-order Taylor approximation degrades in proportion to the magnitude of curvature in $\boldsymbol{f}$.

Returning to equation (24), we proceed by taking conditional expectations with respect to $\mathbf{x}_t$ given $\mathbf{u}_t$ on both sides of the inequality. This yields an upper bound on the expected approximation error,

$$\mathbb{E}\left[|\mathcal{E}_i| \mid \mathbf{u}_t\right] \le L_i \cdot \mathbb{E}\left[\|\mathbf{x}_t - \boldsymbol{a}\|_2^2 \,\middle|\, \mathbf{u}_t\right]. \tag{26}$$

This upper bound is minimized by choosing $\boldsymbol{a} = \mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t]$. Plugging in this choice, we observe that

$$\mathbb{E}\left[|\mathcal{E}_i| \mid \mathbf{u}_t\right] \le L_i \cdot \mathrm{Tr}\left[\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t]\right] \tag{27}$$

where $\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t] = \mathbb{E}\left[(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t])(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t])^\top \,\middle|\, \mathbf{u}_t\right]$ is the conditional covariance of $\mathbf{x}_t$ given $\mathbf{u}_t$. Finally, we can sum these upper bounds over $i = \{1, \dots, D\}$ to bound the total expected approximation error as

$$\sum_i \mathbb{E}\left[|\mathcal{E}_i| \mid \mathbf{u}_t\right] \le L \cdot \mathrm{Tr}\left[\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t]\right] \tag{28}$$

where we have defined $L = \sum_i L_i$ as a global constant bounding the second-order smoothness of $\boldsymbol{f}$ across all dimensions.

Returning to equation (22) and plugging in the optimal choice of $\mathbf{a} = \mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t]$, we obtain the following CLDS approximation to the nonlinear dynamics

$$\boldsymbol{h}(\mathbf{x}_t, \mathbf{u}_t) \coloneqq \boldsymbol{f}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t], \mathbf{u}_t) + \nabla_{\boldsymbol{x}} \boldsymbol{f}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t], \mathbf{u}_t)(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t]) = \boldsymbol{A}(\mathbf{u}_t)\mathbf{x}_t + \boldsymbol{b}(\mathbf{u}_t) \tag{29}$$

where $\nabla_{\boldsymbol{x}} \boldsymbol{f}$ is the matrix-valued Jacobian, $\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}^{D \times D}$, with respect to the first argument of $\boldsymbol{f}$ and we have re-arranged the terms and defined

$$\boldsymbol{A}(\mathbf{u}_t) \coloneqq \nabla_{\boldsymbol{x}} \boldsymbol{f}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t], \mathbf{u}_t), \quad \boldsymbol{b}(\mathbf{u}_t) \coloneqq \boldsymbol{f}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t], \mathbf{u}_t) - \nabla_{\boldsymbol{x}} \boldsymbol{f}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t], \mathbf{u}_t)\mathbb{E}[\mathbf{x}_t \mid \mathbf{u}_t]. \tag{30}$$

For each value of $\mathbf{u}_t$, the quality of this approximation is guaranteed by equation (28) to be small if the second-order derivatives of $\boldsymbol{f}$ with respect to $\mathbf{x}$ are small and if the conditional variance of $\mathbf{x}_t$ given $\mathbf{u}_t$ is small. We note that this analysis of approximation error and does not account for the additional estimation error we incur when learning the functions $\boldsymbol{A}(\mathbf{u})$ and $\mathbf{b}(\mathbf{u})$ from noisy and limited data. Nonetheless, this analysis tells us that we expect the CLDS to perform well in circumstances where the underlying dynamics are smooth and the conditional distributions of $\mathbf{x}_t$ given $\mathbf{u}_t$ have low variance.

A.2.2. Composite dynamics

When considering the *composite dynamics* in (10) (Section §3.1), we are interested in approximating the input-driven nonlinear dynamics given by equation (21) with autonomous nonlinear dynamics, governed by some function $\boldsymbol{g}$ such that

$$\boldsymbol{f}(\mathbf{x}_t, \mathbf{u}_t) \approx \boldsymbol{g}(\mathbf{x}_t). \tag{31}$$

We can evaluate the quality of this approximation using the conditional expectation of the squared error

$$\mathbb{E}\left[\|\boldsymbol{f}(\mathbf{x}_t, \mathbf{u}_t) - \boldsymbol{g}(\mathbf{x}_t)\|_2^2 \;\middle|\; \mathbf{x}_t\right] \tag{32}$$

where the conditional expectation is taken over $\mathbf{u}_t$ given $\mathbf{x}_t$. This approximation error is minimized by choosing

$$\boldsymbol{g}(\mathbf{x}_t) = \mathbb{E}[\boldsymbol{f}(\mathbf{x}_t, \mathbf{u}_t) \mid \mathbf{x}_t] = \mathbb{E}[\mathbf{x}_{t+1} \mid \mathbf{x}_t] \tag{33}$$

However, learning $\boldsymbol{f}$ from limited data is challenging, so we replace this with our CLDS model to achieve the composite dynamical system

$$\boldsymbol{g}(\mathbf{x}_t) \approx \mathbb{E}[\boldsymbol{h}(\mathbf{x}_t, \mathbf{u}_t) \mid \mathbf{x}_t] \tag{34}$$

where $\boldsymbol{h}(\mathbf{x}, \mathbf{u}) = \boldsymbol{A}(\mathbf{u})\mathbf{x} + \boldsymbol{b}(\mathbf{u})$ as in equation (29). Now we analyze the quality of this approximation. Consider the residual along dimension $i \in \{1, \ldots, D\}$,

$$\mathcal{R}_i = f_i(\mathbf{x}_t, \mathbf{u}_t) - \mathbb{E}[h_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) \mid \mathbf{x}_t] \tag{35}$$

where the expectation in the second term is taken over $\tilde{\mathbf{u}}_t$, which is drawn from the conditional distribution of $\mathbf{u}_t$ given $\mathbf{x}_t$. That is, $\mathcal{R}_i$ is a random variable that depends on a joint sample of $(\mathbf{x}_t, \mathbf{u}_t)$ from the stationary distribution, and $\tilde{\mathbf{u}}_t$ is a dummy variable that is integrated out during the calculation of $\mathcal{R}_i$. To proceed, we note that

$$\begin{aligned} \mathcal{R}_i &= \mathbb{E}[f_i(\mathbf{x}_t, \mathbf{u}_t) - h_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) \mid \mathbf{x}_t] & (36) \\ &= \mathbb{E}[f_i(\mathbf{x}_t, \mathbf{u}_t) - f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) + f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) - h_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) \mid \mathbf{x}_t] & (37) \end{aligned}$$

and, applying Jensen's inequality and the triangle inequality, we conclude

$$\begin{aligned} |\mathcal{R}_i| &\leq \mathbb{E}_{\tilde{\mathbf{u}}_t|\mathbf{x}_t}\big[\,\big|f_i(\mathbf{x}_t, \mathbf{u}_t) - f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) + f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) - h_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t)\big|\,\big] & (38) \\ &\leq \mathbb{E}_{\tilde{\mathbf{u}}_t|\mathbf{x}_t}\big[\,\big|f_i(\mathbf{x}_t, \mathbf{u}_t) - f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t)\big|\,\big] + \mathbb{E}_{\tilde{\mathbf{u}}_t|\mathbf{x}_t}\big[\,\big|f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) - h_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t)\big|\,\big] & (39) \end{aligned}$$

where we have introduced a minor change in notation, using $\mathbb{E}_{\tilde{\mathbf{u}}_t|\mathbf{x}_t}$ to denote the conditional expectation of $\tilde{\mathbf{u}}_t$, given $\mathbf{x}_t$. Now we take expectations on both sides of this inequality with respect to the remaining random variables, $\mathbf{x}_t$ and $\mathbf{u}_t$, which are sampled from some stationary distribution associated with the dynamical system. Using the law of total expectation, we obtain

$$\mathbb{E}|\mathcal{R}_i| \leq \mathbb{E}_{\mathbf{x}_t}\left[\mathbb{E}_{\mathbf{u}_t, \tilde{\mathbf{u}}_t|\mathbf{x}_t}\big[\,\big|f_i(\mathbf{x}_t, \mathbf{u}_t) - f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t)\big|\,\big]\right] + \mathbb{E}_{\mathbf{u}_t}\left[\mathbb{E}_{\mathbf{x}_t|\mathbf{u}_t}\big[\,\big|f_i(\mathbf{x}_t, \mathbf{u}_t) - h_i(\mathbf{x}_t, \mathbf{u}_t)\big|\,\big]\right] \tag{40}$$

On the right hand side, the first term takes the conditional expectation over $\mathbf{u}_t$ and $\tilde{\mathbf{u}}_t$, followed by an expectation over $\mathbf{x}_t$. The second term reverses this order, taking the conditional expectation over $\mathbf{x}_t$, followed by an expectation over $\mathbf{u}_t$ (since this is identically distributed to $\tilde{\mathbf{u}}_t$, we drop the tilde).

To upper bound the first term, we introduce an assumption that f_i is Lipschitz continuous in its second argument. That is, there exists a constant $C_i > 0$ such that

$$|f_i(\mathbf{x}_t, \mathbf{u}) - f_i(\mathbf{x}_t, \mathbf{u}')| \leq C_i \|\mathbf{u} - \mathbf{u}'\|_2 \qquad \forall \mathbf{u}, \mathbf{u}' \in \mathcal{U}. \tag{41}$$

In conjunction with Jensen's inequality, this Lipschitz assumption implies the following upper bound:

$$\begin{aligned}
\mathbb{E}_{\mathbf{u}_t, \tilde{\mathbf{u}}_t | \mathbf{x}_t} [\, \big| f_i(\mathbf{x}_t, \mathbf{u}_t) - f_i(\mathbf{x}_t, \tilde{\mathbf{u}}_t) \big| \,] &\leq C_i \cdot \mathbb{E}_{\mathbf{u}_t, \tilde{\mathbf{u}}_t | \mathbf{x}_t} [\, \big\| \mathbf{u}_t - \tilde{\mathbf{u}}_t \big\|_2 \,] && (42) \\
&\leq C_i \sqrt{\mathbb{E}_{\mathbf{u}_t, \tilde{\mathbf{u}}_t | \mathbf{x}_t} \big\| \mathbf{u}_t - \tilde{\mathbf{u}}_t \big\|_2^2} && (43) \\
&= C_i \sqrt{2 \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{u}_t \mid \mathbf{x}_t]]} && (44)
\end{aligned}$$

It remains to upper bound the second term in equation (40). A direct application of the results in §A.2.1 yields the bound

$$\mathbb{E}_{\mathbf{x}_t | \mathbf{u}_t} [\, \big| f_i(\mathbf{x}_t, \mathbf{u}_t) - h_i(\mathbf{x}_t, \mathbf{u}_t) \big| \,] \leq L_i \cdot \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t]] \tag{45}$$

where L_i, defined in (25), is a constant bounding the second derivatives of f_i with respect to $\mathbf{x}_t$. Putting these pieces together we conclude that

$$\mathbb{E}|\mathcal{R}_i| \leq C_i \cdot \mathbb{E}_{\mathbf{x}_t} \sqrt{2 \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{u}_t \mid \mathbf{x}_t]]} + L_i \cdot \mathbb{E}_{\mathbf{u}_t} \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t]] \tag{46}$$

And so an upper bound on the total absolute error of the composite dynamics is given by

$$\sum_{i=1}^{D} \mathbb{E}|\mathcal{R}_i| \leq C \cdot \mathbb{E}_{\mathbf{x}_t} \sqrt{2 \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{u}_t \mid \mathbf{x}_t]]} + L \cdot \mathbb{E}_{\mathbf{u}_t} \operatorname{Tr}[\mathbb{C}\mathrm{ov}[\mathbf{x}_t \mid \mathbf{u}_t]] \tag{47}$$

where we have defined $C = \sum_i C_i$ and $L = \sum_i L_i$.

In summary, we have shown that the approximation error of the composite dynamical system, defined in equation (10), is bounded by a sum of two terms. The first term approaches zero in the limit that the conditional covariance of $\mathbf{u}_t$ given $\mathbf{x}_t$ goes to zero, while the second term approaches zero in the limit that the conditional covariance of $\mathbf{x}_t$ given $\mathbf{u}_t$ goes to zero. Thus, the composite dynamics have the potential to provide an accurate depiction of the true nonlinear dynamical system if $\mathbf{x}_t$ and $\mathbf{u}_t$ are close to being in one-to-one correspondence with each other.

We note that in practice when computing the composite dynamics in (10), we make the simplifying assumption that $p(\mathbf{u}_t | \mathbf{x}_t = \boldsymbol{x})$ does not depend on t (i.e. we assume that this decoding distribution is stationary).

A.3. Correspondence between CLDS and Switching LDS

In this section, we explore the relationship between the linear time-variant dynamics of (1a) with $\mathbf{A}_{ij} \overset{iid}{\sim} \mathcal{GP}(0, k_t)$, and the dynamics of a switching linear dynamical system. While the latter has parameters evolving over a discrete set, we can nonetheless explore how to think of this discrete support as embedded within $\mathbb{R}^{D \times D}$, and seek to match the moments of these two processes.

The first thing to note is that by drawing the entries of $\mathbf{A}_{ij}$ i.i.d., we can gain insight by considering a single process $\mathrm{a}(\cdot) \sim \mathcal{GP}(0, k_t)$, and a SLDS with one-dimensional dynamics. Thus, we consider the SLDS model

$$p(\mathrm{z}_{t+1} = i \mid \mathrm{z}_t = j) = P_{ij} \tag{48a}$$

$$\mathrm{x}_{t+1} = \mathrm{a}^{(\mathrm{z}_t)} \mathrm{x}_t + \epsilon_t \tag{48b}$$

with discrete states z_t governing dynamics in x_t, with transition matrix P and dynamics $\mathrm{a}^{(z)}$ for $z \in \mathcal{Z}$. Denote $\boldsymbol{z} = (z_1, \ldots, z_K)^\top \in \mathbb{R}^{|\mathcal{Z}|}$ for the vector of the $|\mathcal{Z}| = K$ values that can be taken by the stochastic process z, and similarly $\boldsymbol{a} = (a^{(z_1)}, \ldots, a^{(z_K)})^\top$. We consider π the stationary distribution for the z process. Finally, note that the map $z_n \to a^{(z_n)}$ is deterministic, one-to-one and onto, such that we can think of the Markov chain in z_t as having support over $\boldsymbol{a}$. Let $\mathrm{a}_t := a^{(\mathrm{z}_t)}$, and denote $a_i = a^{(z_i)}$.

Let us now explore the moments of the stochastic process a_t to determine its relationship with the CLDS. Assume z_t has reached stationarity, such that $p(\mathrm{z}_t) = p(\mathrm{a}_t) = \pi$. Then, first,

$$\mathbb{E}[\mathrm{a}] = \pi^\top \boldsymbol{a} \tag{49}$$

and then we have the cross-correlation

$$\begin{aligned}
\mathbb{E}\left[\mathrm{a}_t \mathrm{a}_{t+n}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{a}_t \mathrm{a}_{t+n} \mid \mathrm{a}_t\right]\right] \\
&= \sum_{j=1}^{K} \mathbb{E}\left[\mathrm{a}_t \mathrm{a}_{t+n} \mid \mathrm{a}_t = a_j\right] p(\mathrm{a}_t = a_j) \\
&= \sum_{j=1}^{K} \sum_{i=1}^{K} a_j \left(p(\mathrm{a}_{t+n} = z_i \mid \mathrm{a}_t = a_j)\right) p(\mathrm{a}_t = a_j) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} a_i a_j P_{ij}^n \pi_j \\
&= \boldsymbol{a}^\top P^n \mathrm{diag}(\pi) \boldsymbol{a}.
\end{aligned} \tag{50}$$

Denote $\Pi = \mathrm{diag}(\pi)$. We get the desired covariance

$$\mathrm{cov}(\mathrm{a}_t, \mathrm{a}_{t+n}) = \mathbb{E}\left[\mathrm{a}_t \mathrm{a}_{t+n}\right] - \mathbb{E}\left[\mathrm{a}_t\right] \mathbb{E}\left[\mathrm{a}_{t+n}\right] = \boldsymbol{a}^\top \left(P^n \Pi - \pi\pi^\top\right) \boldsymbol{a} \tag{51}$$

yielding our kernel form

$$\mathrm{cov}(\mathrm{a}(t_i), \mathrm{a}(t_j)) = k_t(t_i, t_j) = \boldsymbol{a}^\top \left(P^{|t_j - t_i|} \Pi - \pi\pi^\top\right) \boldsymbol{a} \tag{52}$$

Hence, in all, our approximation of the stochastic process a_t over $\mathbb{R}$ up to the first two moments is

$$\mathrm{a}(\cdot) \sim \mathcal{GP}\left(\pi^\top \boldsymbol{a}, \boldsymbol{a}^\top \left(P^{|t_j - t_i|} \Pi - \pi\pi^\top\right) \boldsymbol{a}\right) \tag{53}$$

which in particular can be made zero-mean by consider values $\boldsymbol{a}$ such that $\pi^\top \boldsymbol{a} = 0$. This establishes the form of the GP prior over $\mathrm{a}(\cdot)$ for which the corresponding CLDS best matches the SLDS.

B. Experiments

B.1. Task-agnostic model implementations

We initialize observation matrices for all models ($\boldsymbol{C}$ in (C)LDS models, log-rate decoder weights in LFADS) as the PCA principal axes in $\mathbf{y}$-space—that is, the top D right singular vectors of the data—for each task.

CLDS In all experiments, we assume that $\mathbf{d}(\boldsymbol{u}_t) = \mathbf{0}$, which forces the predicted firing rates, conditioned on $\boldsymbol{u}_t$, to lie in a D-dimensional space spanned by the columns of $\mathbf{C}(\boldsymbol{u}_t)$.

LFADS We use the Jax implementation of LFADS available at https://github.com/google-research/computation-thru-dynamics/tree/master/lfads_tutorial. We choose the factor dimension to be the same as the latent dimension D of the (C)LDS models and the inferred inputs to be of dimension $|\mathcal{U}|$, both following the (C)LDS models on any given experiment. The other components of the architecture are held fixed across all experiments:

- Encoder, controller and generator have hidden-states of dimension 32;
- The inferred inputs are modeled as having an auto-correlation of 1.0 and a variance of 0.1;
- We train for 1000 epochs, with an initial learning rate of 0.5 with exponential decay at rate 0.995, along with a KL warm-up coming in at 500 steps.

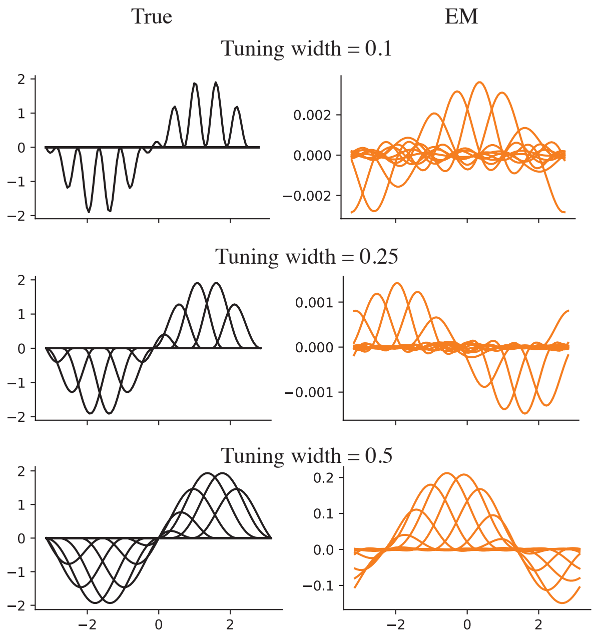

Figure 5. Recovery of $\mathbf{C}$. Rows indicate varying level of tuning curve width γ for $\mathbf{C}_{i,:}(\boldsymbol{u})$. Recovery becomes more challenging for smaller width since it requires a higher and higher number of bases L to approximate the true tuning bump.

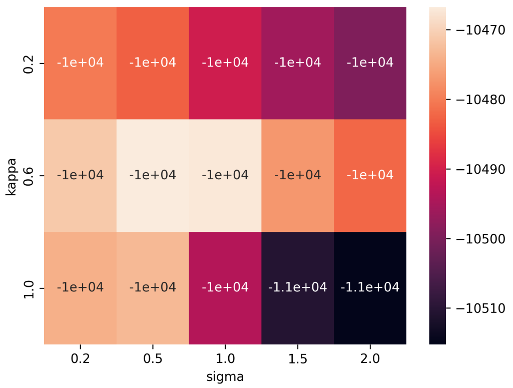

Figure 6. Hyperparameter search, CLDS marginal log-likelihood on a held-out validation dataset set for $\kappa \in \{0.2, 0.6, 1.0\}$ and $\sigma \in \{0.2, 0.5, 1.0, 1.5, 2.0\}$. Maximum attained at $\{\kappa, \sigma\} = \{0.6, 0.5\}$.

B.2. Synthetic experiment and parameter recovery

For the peaks ξ_i spanning regularly the interval $[-\pi, \pi)$ and widths γ, the tuning curves in the rows of $\mathbf{C}$ for the synthetic experiment are defined as

$$\boldsymbol{C}_{i,:}(\boldsymbol{u}) = \begin{cases} \left(1 + \cos\left(\frac{\boldsymbol{u} - \xi_i}{\gamma}\right)\right) \boldsymbol{u}^\top & \text{if } \boldsymbol{u} \in (\xi_i - \gamma\pi, \xi_i + \gamma\pi) \\ 0 & \text{else} \end{cases} \tag{54}$$

We plot the its recovery with our inference procedure in Fig. 5, up to the invertible transform non-identifiability inherent to LDS models.

B.3. Pre-processing

Mice Head-Direction We considered neural activity from the "wake" period, binned in 50ms time-bins, then processed to firing rates and separated into 10s trials.

Macaque center-out reaching We analyzed neural recordings of dorsal premotor cortex (PMd) in macaques performing center-out reaching task from N. Even-Chen, B. Sheffer et al. (2019). The experiments supported 3 different reach radii, and we selected only the middle reach radius at 8cm. We were left with only the angular direction as the reach condition, over $N = 16$ possible reach angles. We aligned all trials around the go-cue, selecting 200 ms before the go-cue and 300ms after. We binned the data into 5ms bins, and performed Gaussian kernel smoothing with a standard deviation of 0.5 over bins.

B.4. Hyper-parameters

Throughout all experiments, we've set $L = 5$ to balance expressivity and number of parameters. To select the other hyper-parameters of GP prior length-scale κ and scale σ, we evaluated the various models on a held-out validation set. We plot in Fig. 6 our search over $\{\kappa, \sigma\}$ on the macaque center-out reaching data.