The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Reduced-Rank Linear Dynamical Systems

Qi She,[1,3] **Yuan Gao,**[2] **Kai Xu,**[4,5] **Rosa H. M. Chan**[3]
[1]Princeton Neuroscience Institute, Princeton University
[2]Tencent AI Lab
[3]Department of Electronic Engineering, City University of Hong Kong
[4]Department of Computer Science, Princeton University
[5]School of Computer Science, National University of Defense Technology

Abstract

Linear Dynamical Systems are widely used to study the underlying patterns of multivariate time series. A basic assumption of these models is that high-dimensional time series can be characterized by some underlying, low-dimensional and time-varying latent states. However, existing approaches to LDS modeling mostly learn the latent space with a prescribed dimensionality. When dealing with short-length high-dimensional time series data, such models would be easily overfitted. We propose Reduced-Rank Linear Dynamical Systems (RRLDS), to automatically retrieve the intrinsic dimensionality of the latent space during model learning. Our key observation is that the rank of the dynamics matrix of LDS captures the intrinsic dimensionality, and the variational inference with a reduced-rank regularization finally leads to a concise, structured, and interpretable latent space. To enable our method to handle count-valued data, we introduce the *dispersion-adaptive* distribution to accommodate over-/ equal-/ and under-dispersion nature of such data. Results on both simulated and experimental data demonstrate our model can robustly learn latent space from *short-length*, noisy, *count-valued* data and significantly improve the prediction performance over the state-of-the-art methods.

Introduction

Deciphering the latent structure from high-dimensional time series is one of the fundamental problems of Artificial Intelligence, which has been extensively applied in various fields from social, economics, to biology science (Linderman, Stock, and Adams 2014; She, So, and Chan 2015; She, Chen, and Chan 2016; So et al. 2016; Hein et al. 2016). In such settings, many studies and theories posit that high-dimensional time series are noisy observations of some underlying, low-dimensional, and time-varying signal of interest (Pfau, Pnevmatikakis, and Paninski 2013; Archer et al. 2014; Sussillo et al. 2016). Linear Dynamical Systems (LDS) have been employed to extract a low-dimensional *implicit* network structure from observed multivariate time series data (Archer et al. 2014; Lakshmanan et al. 2015; Linderman et al. 2017), which captures the variability of observations, both spatially and temporally.

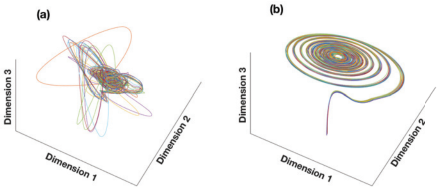


Figure 1: Latent trajectories reconstructed from **(a)** unconstrained dynamics matrix and **(b)** reduced-rank dynamics matrix (different colors indicate different simulated trials). The low-dimensional manifold in **(b)** is smoother and better structured.

However, two main challenges exist when using LDS to retrieve an optimal *implicit* network structure. First, the existing models need a predefined latent dimensionality. In order to ensure the models' capability, it is typically set to be a large value, which leads to the difficulties in modeling the *short-length* high-dimensional time series data due to overfitting. This modeling problem is troublesome since the *short-length* time series data exist in many real-world scenarios. For example, in neuroscience, we cannot obtain long sequences of high-quality neural data in experiments because of (i) the short lifetime of some neurons, (ii) the limited viable time of recording materials and (iii) the micro-movement of recording electrodes during an activity of the animal (Spira and Hai 2013). In the clinical domain, the length of patient clinical data is usually less than 50 because the hospitalization period for most patients is less than two weeks (Banaee, Ahmed, and Loutfi 2013). In economics, the econometric multivariate time series, such as gross domestic product and consumer price index, are measured quarterly or yearly which results in short-length data.

Second, real-world time series data are often count-valued (rather than real-valued). The application of standard LDS, which assumes the observation follows Gaussian distribution, is infeasible (She, So, and Chan 2016). Examples include multiple spike trains recorded from neural populations (Paninski et al. 2010), the data of trades on the S&P 100 index (Linderman and Adams 2014), to name just a few. Extensions on model to handle count nature of the data are necessary. Recently, Poisson linear dynamical system (PLDS) (Buesing et al. 2014) was proposed for count data modeling. Nevertheless, the Poisson assumption implies equal dispersion of the observations, i.e., the conditional mean and variance are equal.

This limits PLDS in characterizing neural spike counts, which are commonly observed to be either over- or under-dispersed (variance greater or less than mean) (Churchland et al. 2010; She, Jelfs, and Chan 2016). Without a proper distribution capturing the dispersion of count data, one cannot learn the variability of the data, thus failing to infer an optimal *implicit* network.

In view of these limitations, we propose a novel solution to infer the *implicit* network from *short-length*, noisy *count* data. We focus on the dynamics matrix of LDS, which represents the influence that one latent node exerts on the subsequent activity of another. In other words, the dynamics matrix is used to govern the node evolution of the *implicit* network. Our key observation is that the rank of the dynamics matrix captures the intrinsic dimensionality of the state space of the nodes. To prevent LDS from overfitting given short-length data, we seek to learn a compact *low-rank dynamics matrix*. Specifically, we compose two different low-rank priors for dynamics matrix, *i.e.*, multivariate Laplacian and nuclear norm, which offer similar performance for retrieving the intrinsic dimensionality and are widely applied to different scenarios(Gao and Yuille 2017; Gao, Ma, and Yuille 2017). Moreover, to facilitate the learning of reduced-rank dynamics matrix from *count* data, we introduce the *dispersion-adaptive ($\mathcal{DA}$)* distribution and develop a novel, flexibly-parameterized observation model.

Figure 1 demonstrates the advantage of reduced-rank dynamics matrix with $\mathcal{DA}$ distribution in recovering low dimensional manifolds from *short-length*, noisy *count-valued* time series data. The observation is 40-dimension (*i.e.* 40D) time series data, which is modeled with a 10D dynamics matrix (same initial states). It shows that our method successfully retrieves three intrinsic dimensionalities from the dynamics matrix, leading to a smoother and better structured manifold indicated by the three dimensional curves, while the method with unconstrained dynamics matrix fails. In summary, our contributions are four-folds:

- We propose to retrieve intrinsic dimensionality of multivariate time series by imposing two *reduced-rank* structures on the dynamics matrix.
- We introduce a count-valued exponential family distribution (called $\mathcal{DA}$ distribution) to capture the dispersion nature of count data, and derive a variety of commonly used distributions as special cases.
- We utilize a latent, *reduced-rank* linear dynamical model to modulate expectation of $\mathcal{DA}$ observation distribution, thus forming a novel linear dynamical system model.
- We develop a Variational Bayes Expectation Maximization (VBEM) algorithm by extending the current state-of-the-art methods to the novel model.

Our framework is evaluated against the baseline methods on both simulated and real-world data. The promising performance demonstrates that our method is able to: (1) automatically reduce redundant dimensions of latent state space, which prevents overfitting with large number of predefined latent states; (2) significantly improve prediction performance over baseline methods for noisy neuronal spiking activities; and (3) robustly and efficiently retrieve intrinsic dimensionality of underlying complex neural systems from two experimental datasets.

Background

Linear Dynamical System (LDS), with well-developed inference and learning methods, is an elegant mathematical framework for modeling and learning multivariate time series (**MTS**). LDS models **real-valued** MTS $\{\mathbf{y}_t \in \mathbb{R}^q\}_{t=1}^T$ using latent states $\{\mathbf{x}_t \in \mathbb{R}^n\}_{t=1}^T$:

$$\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_t|A\mathbf{x}_{t-1}, Q), \quad (1)$$

$$\mathbf{y}_t|\mathbf{x}_t \quad \sim \mathcal{N}(\mathbf{y}_t|C\mathbf{x}_t, R). \quad (2)$$

Eq. 1 represents state dynamics, and Eq. 2 is the observation model. Briefly, $\{\mathbf{x}_t\}$ is evolved via a dynamics matrix $A \in \mathbb{R}^{n\times n}$. Observations $\{\mathbf{y}_t\}$ are generated from $\{\mathbf{x}_t\}$ via a emission matrix $C \in \mathbb{R}^{q\times n}$. These two processes have Gaussian noise with mean $\mathbf{0}$ and covariance matrices Q and R respectively. The initial state $\mathbf{x}_1$ is multivariate Gaussian distribution with mean $\mathbf{x}_0$ and covariance Q_0. The complete set of the LDS parameters is $\Omega = \{A, C, Q, R, \mathbf{x}_0, Q_0\}$.

While in some LDS applications, the model parameters are known a *priori*, in the majority of real-world applications they are unknown, and we learn them from MTS data. This can be done with LDS learning methods such as the Expectation-Maximization (Ghahramani and Hinton 1996) or spectral learning (Buesing, Macke, and Sahani 2012).

Related Work

Various regularization methods have been proposed for both time series modeling and prediction tasks with LDS under different applications (Charles et al. 2011). These can be divided into four categories: (1) state regularization; (2) innovation regularization; (3) combination regularization; and (4) parameter regularization.

State Regularization The latent states $\{\mathbf{x}_t\}_{t=1}^T$ are sparsified at each step of Kalman filter inference. Charles *et al.* (2011) incorporates sparsity constraints to achieve a sparse state estimate $\hat{\mathbf{x}}_t$ at each time stamp. Angelosante *et al.* (2009) treats state sequence $\{\mathbf{x}_t\}_{t=1}^T$ as a state estimate matrix and enforces a row level group lasso on this matrix.

Innovation Regularization The error of state estimation is called "innovation", *i.e.*, $||\hat{\mathbf{x}}_t - A\hat{\mathbf{x}}_{t-1}||$. ℓ_1 regularization is applied on innovation during state estimation, which can help to balance fidelity to the measurements against the sparsity of the innovations (Asif et al. 2011).

Combination Regularization Ghanem and Ahuja (2010) trains a dictionary of LDSs, in which each LDS is learned via one trial training MTS. The final LDS is a weighted combination of these individual LDSs, whose weights are regularized by an ℓ_1 penalty.

Parameter Regularization **($\star$)** Parameter regularization imposes regularization penalties on the parameters of a LDS during the learning process. Boots *et al.* (2007) limited the largest eigenvalue of dynamics matrix within unit circle to avoid unstable latent dynamics. A spectral algorithm has been

Prior Name	Prior Form	Regularization
Multivariate Laplacian	$\propto \exp(-\beta_1 \|\|A_i\|\|_2)$	$\beta_1 \|\|A_i\|\|_2$
Nuclear norm	$\propto \exp(-\beta_2 \|\|A\|\|_*)$	$\beta_2 \|\|A\|\|_*$

Table 1: Prior choices for dynamics matrix

proposed to learn a stable LDS, which is good for simulating and predicting from learned system. Our solution in reduced-rank linear dynamical system belongs to this categories as we introduce a *low-rank dynamics matrix* for recovering the intrinsic dimensionality.

Our method is different from category (1) and (2) because both of them learn a sparse representation for latent states $\{\mathbf{x}_t\}_{t=1}^T$ while they assume all parameters of LDS as a *priori*. The combination regularization in (3) require extra training process since they need to construct a LDSs dictionary from different MTS. For limited time series data in our case, they are unable to solve the overfitting problem and retrieve the intrinsic dimensionality of MTS. While our method belongs to the same category (4) as the stable LDS proposed by Boots *et al.* (2007), we focus on different aspects of the problem. They attempt to achieve stability in LDS, while we aim to find an appropriate state space and prevent overfitting given a small amount of MTS count data.

For learning LDS model from Gaussian observations, standard EM algorithm (Ghahramani and Hinton 1996) can iteratively find the maximum likelihood solution. Subspace Identification (SubspaceID) method compute an asymptotically unbiased solution in closed form by using oblique projection and Singular Value Decomposition (Van Overschee and De Moor 2012). For learning LDS model from count observations, Busing *et al.* proposed Poisson Linear Dynamical Systems (PLDS) and to learn it using spectral learning method (2012) or variational inference (2014). The Poisson assumption of count data, while offering algorithmic conveniences, implies the conditional mean and variance of count data are equal. This property is improper in some analysis of count data, which are observed to be either over- or under-dispersed (Churchland et al. 2010). Thus, it is crucial to develop more general observation distributions to capture over/equal/under-dispersion of count data.

To address these needs, we also employ a count-valued exponential family distribution (*weighted Poisson distribution*), which is superior to current methods in two aspects: (i) adaptive dispersion, where a log-convex/linear/concave weight function will produce the expected over/equal/under-dispersion; (ii) flexible setting, with a variety of previous work are derived as special cases under this distribution.

Methodology

Reduced-Rank Structure

In order to recover the intrinsic dimensionality from MTS datasets through the rank of dynamics matrix A, we shall choose specific priors which can induce the desired low-rank property. We have two choices of inducing a low-rank dynamics matrix: (1) a multivariate Laplacian prior and (2) a nuclear norm prior as shown in Table 1:

(1) Multivariate Laplacian prior It assumes every row in dynamics matrix A is independent of each other and has the multivariate Laplacian density. Also in order to avoid overfitting, we introduce a multivariate Gaussian prior to each element in A , which leads to the ridge regularization. Then, we combine the multivariate Laplacian prior and Gaussian prior to get a new prior $p_{\mathcal{ML}}(A)$, as

$$\log p_{\mathcal{ML}}(A) = -\beta_1 \sum_{i=1}^{n} \|A_i\|_2 - \frac{\beta_2}{2}\|A\|_F^2 + \text{const}, \quad (3)$$

where β_1,β_2 are regularization parameters. $\{A_i\}_{i=1}^n$ indicates rows of A. $\|\cdot\|_2$ and $\|\cdot\|_F$ are ℓ_2 and Frobenius norm.

(2) Nuclear norm prior It can be regarded as a convex relaxation of the number of non-zeros eigenvalues (*i.e.*,the rank) of the dynamics matrix A. We get an alternative prior $p_{\mathcal{NN}}(A)$ by applying nuclear norm density and multivariate Gaussian to dynamics matrix, as

$$\log p_{\mathcal{NN}}(A) = -\beta_3\|A\|_* - \frac{\beta_4}{2}\|A\|_F^2 + \text{const}, \quad (4)$$

where $\|\cdot\|_*$ is nuclear norm. β_3, β_4 are regularization parameters. $\{\beta_i\}_{i=1}^4$ are selected (in all experiments) by the internal cross validation while optimizing model's predictive performance. We impose $p_{\mathcal{ML}}(A)$ and $p_{\mathcal{NN}}(A)$ separately to the learning process, and derive two methods to optimize a low-rank dynamics matrix.

Dispersion-adaptive ($\mathcal{DA}$) Distribution

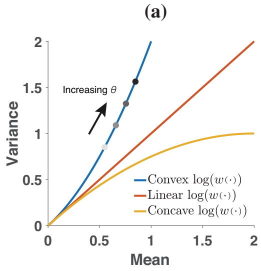
For count-valued observations, we define the $\mathcal{DA}$ distribution as the family of count-valued probability distribution:

$$p_{\mathcal{DA}}(Y = k; \theta, w(\cdot)) = \frac{w(k)\exp(\theta k)}{k!\mathbb{E}[w(Y)]}, k \in \mathbb{N} \quad (5)$$

where $\theta \in \mathbb{R}$ and the function $w(\cdot) : \mathbb{N} \to \mathbb{R}$ parameterizes the distribution, and $\mathbb{E}[w(Y)] = \sum_{k\in\mathbb{N}} \frac{w(k)\exp(\theta k)}{k!}$ is the normalizing constant. It can be viewed as an extension of Poisson distribution with a weight function $w(\cdot)$. This kind of generalizations has been of interest since del Castillo *et al.* (2005), and they proved that: (1) log-concave/linear/convex functions $w(\cdot)$ imply under/equal/over-dispersed distributions; (2) the expectation of any weighted Poisson distribution is monotonically increasing with θ, for a fixed $w(\cdot)$. Figure 2 (a) demonstrates different $w(\cdot)$ functions model the dispersion of count data, and controlling θ can adjust the mean value of $\mathcal{DA}$ distribution. As shown in Figure 2 (b), we derive many of the commonly used count-data distributions as special cases of $\mathcal{DA}$, by restricting the $w(\cdot)$ function and θ to have certain parametric form. Figure 2 shows that $\mathcal{DA}$ offers a rich, flexible exponential family for count data, and allows $w(\cdot)$ and θ to be interpretable for capturing statistics of count-valued data.

Reduced-Rank Linear Dynamical System (RRLDS)

With two reduced-rank structures and $\mathcal{DA}$ distribution in hand, we now couple them with a latent, linear dynamical system. This system, which we call RRLDS, is beneficial for modeling limited count data to retrieve intrinsic



(b)

Model	Typical parametrization	Reparametrization in $\mathcal{DA}$
Bernoulli	$p(y=k) = p^k(1-p)^{1-k}$	$w(k) = 1; k = 0, 1$ $\theta = \text{logit}(p)$
Poisson	$p(y=k) = \frac{\lambda^k \exp(-\lambda)}{k!}$	$w(k) = 1$ $\theta = \log(\lambda)$
Negative binomial	$p(y=k) = \frac{(k+r-1)!}{k!(r-1)!}(1-p)^r p^k$	$w(k) = (k+r-1)!$ $\theta = \log(p)$
COM-Poisson	$p(y=k) = \frac{\lambda^k}{(k!)^v} / \sum_{j=1}^{+\infty} \frac{\lambda^j}{(j!)^v}$	$w(k) = \exp(1-v) + k!$ $\theta = \log(\lambda)$

Figure 2: **(a)** The mean and variance of the $\mathcal{DA}$ distribution with different choices of the function $w(\cdot)$. With a fixed $\log w(\cdot)$, increasing θ can drive mean and variance to be larger (darker dots); **(b)** Common count distributions are special cases of $\mathcal{DA}$ distribution by parameterizing θ and $w(\cdot)$.

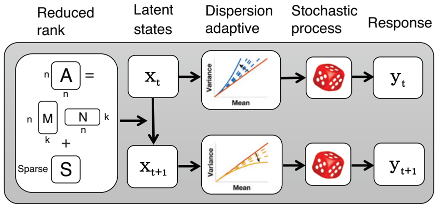


Figure 3: Illustration of the two stages of RRLDS.

dimensionality. We apply it to model time series data (spike counts) recorded from brain neurons, and it is straightforward to extend it to describe and interpret other count-process observations. Denoting $y^i_{t,r}$ as the spike count of neuron $i \in \{1, \dots, q\}$ at time $t \in \{1, \dots, T\}$ on experimental trial $r \in \{1, \dots, R\}$, we assume the spiking activities of neurons are noisy count observations of underlying low-dimensional latent states $\mathbf{x}_{t,r} \in \mathbb{R}^n (n < q)$ (modulating mean value of $\mathcal{DA}$ distribution) and define the $\mathcal{DA}$ observation model as:

$$y^i_{t,r} | \mathbf{x}_{t,r} \sim \mathcal{DA}(c_i^\top \mathbf{x}_{t,r}, w_i(\cdot)). \tag{6}$$

We parametrize $\theta = c_i^\top \mathbf{x}_{t,r}$, where $C = [c_1, \cdots, c_q]^\top \in \mathbb{R}^{q \times n}$ is emission matrix mapping latent space to observation space. $w_i(\cdot)$ is a neuron-specific function capturing the dispersion property of each time series. The evolution of latent state $\mathbf{x}_{t,r}$ is described by a linear first-order process:

$$\begin{aligned} \mathbf{x}_{1,r} &\sim \mathcal{N}(\mathbf{x}_{1,r} | \mathbf{x}_0, Q_0), \\ \mathbf{x}_{t,r} | \mathbf{x}_{t-1,r} &\sim \mathcal{N}(\mathbf{x}_{t,r} | A\mathbf{x}_{t-1,r} + B\mathbf{u}_{t-1,r}, Q). \end{aligned} \tag{7}$$

Here, $\mathbf{x}_0$ and Q_0 are the mean and covariance of the initial state and Q is the covariance of the innovations. External input $\mathbf{u}_{t,r}$ with coupling effects B are considered in the process of latent evolution. For example, in the hippocampal experiments to be presented in the Results section, the rat's position (trajectory) can be regarded as external stimuli, and location-stimulus filter B is obtained under this setting. Meanwhile, reduced-rank structures $p_{\mathcal{ML}}(A)$ and $p_{\mathcal{NN}}(A)$ are imposed on dynamics matrix A.

RRLDS is illustrated in Figure 3 along with two-stage model structure: The first stage includes reduced-rank structures composed on the dynamics matrix A, which governs the evolution of latent states $\mathbf{x}_t$. The second stage maps latent states $\mathbf{x}_t$ onto responses $\mathbf{y}_t$ via $\mathcal{DA}$ observation model, which learns the dispersion property.

Variational Bayes Expectation-Maximization (VBEM) algorithm is adopted for estimating latent states $\mathbf{x}_{1:T,r}$ (E-step) and parameters $\Theta = \{A, B, C, Q, Q_0, \mathbf{x}_0 \{w_i(\cdot)\}_{i=1}^p\}$ (M-step). Since the posterior distribution of $\mathbf{x}_{1:T,r}$ has no analytical solution, an approximation step is implemented via a variational lower bound. Meanwhile, two regularization strategies are proposed for optimizing dynamics matrix.

Inference (E-step)

We need to characterize the full posterior distribution of latent states given the observations $\mathbf{y}_{1:T,r}$ and parameters Θ, such that: $p(\mathbf{x}_{1:T,r} | \mathbf{y}_{1:T,r}, \Theta)$. This distribution is intractable, and usually we make a Gaussian approximation. Denote $\bar{\mathbf{x}}_r = \text{vec}(\mathbf{x}_{1:T,r})$ and $\bar{\mathbf{y}}_r = \text{vec}(\mathbf{y}_{1:T,r})$. For ease of notation in the paper we drop the trial index r. Thus, we use Gaussian approximation as $p(\bar{\mathbf{x}}|\bar{\mathbf{y}}) \approx q(\bar{\mathbf{x}}) = \mathcal{N}(\mu, \Sigma)$. We identify the optimal (μ, Σ) by maximizing a variational Bayesian lower bound (also called "**ELBO**") over the variational parameters μ and Σ as:

$$\mathcal{L}(\mu, \Sigma) = \mathbb{E}_{q(\bar{\mathbf{x}})} \left[\log \left(\frac{p(\bar{\mathbf{x}}|\Theta)}{q(\bar{\mathbf{x}})} \right) \right] + \mathbb{E}_{q(\bar{\mathbf{x}})} \left[\log p\left(\bar{\mathbf{y}} | \bar{\mathbf{x}}, \Theta\right) \right] \tag{8}$$

The first term of Eq. 8 is the negative Kullback-Leibler divergence between the variational distribution and prior distribution, encouraging the variational distribution to be close to the prior. The second term involving the $\mathcal{DA}$ likelihood encourages the variational distribution to explain the observations well. The integrations in the second term are intractable due to non-conjugation. We use the ideas of Khan *et al.* (2013) to derive a further lower bound. First, we expend the second term using $\mathcal{DA}$ likelihood via Eq. 6 as

$$\begin{aligned} &\mathbb{E}_{q(\theta^i_t)} \left[\log p_{\mathcal{DA}} \left(y^i_t | \theta^i_t, w_i(\cdot) \right) \right] \\ &= \mathbb{E}_{q(\theta^i_t)} \left[y^i_t \theta^i_t + \log \frac{w_i(y^i_t)}{y^i_t!} - \log \sum_{k=0}^{K} \frac{\exp(k\theta^i_t) w_i(k)}{k!} \right], \end{aligned} \tag{9}$$

where $\theta^i_t = c_i^\top \mathbf{x}_t$. Denoting $m^i_{t,k} = k\theta^i_t + \log w_i(k) - \log(k!) = kc_i^\top \mathbf{x}_t + \log w_i(k) - \log(k!)$, we can see $m^i_{t,k}$ is a linear transformation of $\mathbf{x}_t$. Under the variational distribution $m^i_{t,k}$ is also normally distributed $m^i_{t,k} \sim \mathcal{N}(h^i_{t,k}, \rho^i_{t,k})$. We have $h^i_{t,k} = kc_i^\top \mu_t + \log w_i(k) - \log k!$, and $\rho^i_{t,k} = k^2 c_i^\top \Sigma_t c_i$, where (μ_t, Σ_t) is the expectation and covariance matrix of $\mathbf{x}_t$ under variational distribution. Then, Eq. 9 is reduced to $\mathbb{E}_{q(m^i_{t,k})}[m^i_{t,k} - \log\sum_k \exp(m^i_{t,k})]$. A further lower bound can then be derived by Jensen's inequality:

$$\begin{aligned}&\mathbb{E}_{q(m^i_{t,k})}[m^i_{t,k} - \log\sum_k \exp(m^i_{t,k})] \qquad (10)\\ &\geq h^i_{t,k} - \log\sum_k \exp(h^i_{t,k} + \rho^i_{t,k}/2) = f(\mathbf{h}^i_t, \rho^i_t).\end{aligned}$$

Combining Eq. 8 and Eq. 10, we can get a tractable variational lower bound, where $\mathcal{L}^*(\mu,\Sigma) \leq \mathcal{L}(\mu,\Sigma)$ as

$$\mathcal{L}^*(\mu,\Sigma) = \mathbb{E}_{q(\bar{\mathbf{x}})}\left[\log\left(\frac{p(\bar{\mathbf{x}}|\Theta)}{q(\bar{\mathbf{x}})}\right)\right] + \sum_{t,i} f(\mathbf{h}^i_t, \rho^i_t). \qquad (11)$$

In the E-step, we maximize the new lower bound $\mathcal{L}^*$ via its dual (Khan et al. 2013). Finally, we have sufficient statistics $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t\mathbf{x}_t^\top]$, $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_{t-1}\mathbf{x}_t^\top]$, and $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t]$, which are necessary to M-step. Details are derived in the supplementary materials.

Learning (M-step)

The M-step requires maximization of $\mathcal{L}^*$ over Θ. This process involves the optimization of three parts: (1) dynamics matrix A; (2) other dynamical system parameters $\{B, Q, Q_0, \mathbf{x}_0\}$; and $\mathcal{DA}$ model parameters $\{C, \{w_i(\cdot)\}^p_i\}$.

Optimization of A In M-step, dynamics matrix A is optimized via maximizing $\mathbb{E}_{\bar{\mathbf{x}}}[\sum_{t=2}^T \log p(\mathbf{x}_t|\mathbf{x}_{t-1})] + \log p(A)$. We already introduced two reduced-rank structure $p_{\mathcal{ML}}(A)$ and $p_{\mathcal{NN}}(A)$. Below we briefly outline two efficient algorithms (*i.e.*, A1 and A2) and our novel contributions.

A1: $p_{\mathcal{ML}}(A)$ Two levels of constraints were applied to A: (1) multivariate Laplacian prior; and (2) multivariate Gaussian prior. The first leads to low-rank structure and the second prevents overfitting. Here the objective function is

$$\min_A g(A) + \beta_1\sum_{i=1}^n \|A_i\|_2 + \frac{\beta_2}{2}\|A\|_F^2, \qquad (12)$$

where $g(A) = \frac{1}{2}\sum_{t=2}^T \mathbb{E}_{\bar{\mathbf{x}}}[\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2_{Q^{-1}}]$ and $\hat{\mathbf{x}}_{t-1} = A\mathbf{x}_{t-1} - B\mathbf{u}_{t-1}$. β_1, β_2 are selected by the internal cross validation. Eq. 12 can be transformed into a quadratic problem with a non-smooth Euclidean norm (details shown in supplementary materials), as

$$\text{argmin}_A \frac{1}{2}a^\top Ha - b^\top a + \beta_1\sum_{i=1}^n \|A_i\|_2. \qquad (13)$$

For clarity $a = \text{vec}(A^\top)$, and $Q^{-1} = LL^\top$. We reformulate

$$\begin{aligned}H &= Q^{-1} \otimes \sum_{t=2}^T \mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top] + \beta_2 I_{n^2},\\ b &= L \otimes \sum_{t=2}^T \mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t\mathbf{x}_{t-1}^\top - B\mathbf{u}_{t-1}\mathbf{x}_{t-1}^\top]^\top \text{vec}(L).\end{aligned}$$

Eq. 13 can be casted into second order cone program (SOCP) and solved using existing SOCP solvers. SOCP always provides solutions with high precision (low duality gap) when the state size remains moderate (<50), which is the case in our experiments. The transformed SOCP is given as

$$\begin{aligned}&\min_{\alpha_0,\cdots,\alpha_n} \alpha_0 + \beta_1\sum_{i=1}^n \alpha_i\\ &\text{s.t.}\quad \alpha_0 \geq \frac{1}{2}a^\top Ha - b^\top a, \quad \alpha_i \geq \|A_i\|_2. \qquad (14)\end{aligned}$$

A2: $p_{\mathcal{NN}}(A)$ We assume that dynamics matrix A has a nuclear norm density and similarly to $p_{\mathcal{ML}}(A)$, we also assume a multivariate Gaussian prior for each element in A. In this case, our objective function becomes:

$$\min_A g(A) + \beta_3\|A\|_* + \frac{\beta_4}{2}\|A\|_F^2. \qquad (15)$$

Denoting $h(A) = g(A) + \frac{\beta_4}{2}\|A\|_F^2$, which is convex and differentiable with respect to A. We can minimize Eq. 15 based on generalized gradient descent algorithm:

$$A^{(j+1)} = \text{prox}_{d_j}\left(A^{(j)} - d_j \bigtriangledown h(A^{(j)})\right), \qquad (16)$$

here $\text{prox}_{d_j}(\cdot)$ is the singular value soft-thresholding operator, and d_j is the step size in iteration j. We select the step size to assure fast convergence rate based on Theorem 1 and proof is in the supplementary material.

Theorem 1 Generalized gradient descent with a fixed step size $d \leq 1/(\|Q^{-1}\|_F \cdot \|\sum_{t=1}^{T-1}\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]\|_F + \beta_3)$ for minimizing Eq. 15 has convergence rate $O(1/J)$, where J is the total number of iterations.

Optimization of $\{B, Q, Q_0, \mathbf{x}_0\}$ With variational distribution $q(\bar{\mathbf{x}})$, the part of the likelihood about Gaussian linear dynamics is quadratic with respect to $\bar{\mathbf{x}}$, and has closed form solutions based on Ghahramani and Hinton (1996).

Optimization of $\{C, w(\cdot)\}$ The part of the likelihood involving $\mathcal{DA}$ observation distribution is not a standard form as in previous work (Ghahramani and Hinton 1996). In our model, considering all experimental trials, we have

$$\begin{aligned}\mathcal{L}_{C,w(\cdot)} &= \sum_{i=1}^q\sum_{t=1}^T\sum_{r=1}^R y^i_{t,r}(c_i^\top\mu_{t,r}) + \log(w_i(y^i_{t,r})) \qquad (17)\\ &- \log(\sum_{k=1}^K \frac{w_i(k)}{k!}\exp(k(c_i^\top\mu_{t,r}) + \frac{1}{2}k^2c_i^\top\Sigma_{t,r}c_i)).\end{aligned}$$

This part is concave and can be optimized efficiently using existing convex optimization techniques. In practice, we initialize our parameters using Laplace-EM algorithm (Buesing et al. 2014), which empirically gives runtime advantages, and produces a sensible optimum.

Above steps provide a fully inference and learning algorithm (VBEM), which is summarized by Algorithm 1.

Results

To demonstrate the generality of $\mathcal{DA}$ and verify our algorithmic implementation, we first test inference and learning method on extensive simulated data. Then we evaluate the prediction performance by comparing with state-of-the-arts over two neuroscience datasets. Finally, we verify that our

Algorithm 1 Framework of inference and learning (VBEM)

```
Input:
  • $\{A, B, C, Q, Q_0, \mathbf{x}_0, w(\cdot)\}$ initialized via Laplace-EM.
  • Observation sequences $y_t^i$ and stimuli sequences $u_t^i$ with $i \in \{1, \dots, q\}$ and $t \in \{1, \dots, T\}$.
Output: $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t]$, $\hat{\Theta} = \{\hat{A}, \hat{B}, \hat{C}, \hat{Q}, \hat{Q}_0, \hat{\mathbf{x}}_0, \{\hat{w}_i(\cdot)\}_{i=1}^p\}$
repeat
    E-step:
    Compute new ELBO : $\mathcal{L}^*(\mu, \Sigma)$ from Eq. 11
    $\{\mu, \Sigma\} \leftarrow$ dual optimization over $\mathcal{L}^*(\mu, \Sigma)$
    $p(\bar{\mathbf{x}}|\bar{\mathbf{y}}, \theta) \leftarrow \mathcal{N}(\mu, \Sigma)$
    Compute $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t \mathbf{x}_t^\top]$, $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_{t-1} \mathbf{x}_t^\top]$, and $\mathbb{E}_{\bar{\mathbf{x}}}[\mathbf{x}_t]$.
    M-step:
    if $p(A) \leftarrow p_{\mathcal{ML}}(A)$ then
        $A \leftarrow$ SOCP solving Eq. 14
    else $p(A) \leftarrow p_{\mathcal{NN}}(A)$
        $d \leftarrow 1/(\|Q^{-1}\|_F \cdot \|\sum_{t=1}^{T-1} \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top]\|_F + \beta_3)$
        Compute gradient of $h(A)$
        $A \leftarrow \text{prox}_d\,(A - d \bigtriangledown h(A))$
    end if
    $\{\hat{B}, \hat{Q}, \hat{Q}_0, \hat{\mathbf{x}}_0\} \leftarrow \text{argmax}_\theta\, \mathbb{E}_{\bar{\mathbf{x}}_k}[\log p(\bar{\mathbf{x}}|\theta)]$.
    $\{\hat{C}, \{\hat{w}_i(\cdot)\}_{i=1}^p\} \leftarrow$ optimization over Eq. 17
until convergence
```

	with Reduced Rank	w/o Reduced Rank
with $\mathcal{DA}$	RRLDS-ML /-NN	DALDS
w/o $\mathcal{DA}$	RRLDS (w/o $\mathcal{DA}$)	alternative LDSs

Table 2: Abbreviations for our method and several baselines. ML stands for Multivariable Laplacian and NN for Nuclear Norm. Alternative LDS methods include vanilla LDS (Ghahramani and Hinton 1996), PLDS (Buesing, Macke, and Sahani 2012), SubspaceID (Van Overschee and De Moor 2012) and StableLDS (Boots, Gordon, and Siddiqi 2007).

method can retrieve the intrinsic dimensionality from those multivariate time series, which is substantially more powerful than the existing works. Table 2 lists the abbreviations of methods compared in the Results part.

Simulated results

Reconstruction of Latent States We evaluate the performance of proposed variational inference for posterior distribution of latent states, given a known set of parameters Θ, observed data $\{\mathbf{y}_{1:T,r}\}_{r=1}^2$ and stimuli $\{u_{1:T,r}\}_{r=1}^2$. The intrinsic dimensionality of simulated data is set to the number of latent states ($\mathbf{x}_{t,r} \in \mathbb{R}^3$) in inference. The true state trajectories $\{\mathbf{x}_{t,r}\}_{r=1}^2$ (grey) and the posterior mean estimates $\{\mathbb{E}[\mathbf{x}_{t,r}]\}_{r=1}^2$ (black) are plotted in Figure 4, with 3 states of 2 trials. It shows that our variational inference method achieves faithful reconstruction of the state trajectories.

Parameter Estimation For parameter estimation, it is inappropriate to perform element-wise comparison between the estimated parameters and ground-truth. This is because dynamics matrices are unique (in terms of producing the same observations) only up to linear transformations. Therefore, we opt to compare the *invariants* of the estimated and true parameters as in Macke *et al.* (2015).

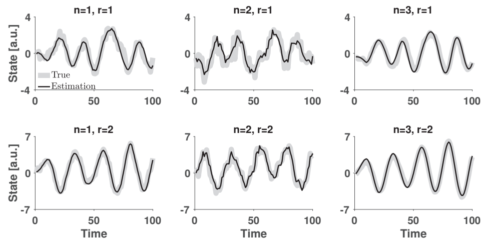


Figure 4: Reconstruction of latent states from multiple count data. Top row: true (grey) and estimated (black) trajectories of 3 latent states in simulated trial #1; Bottom row: the performance in simulated trial #2.

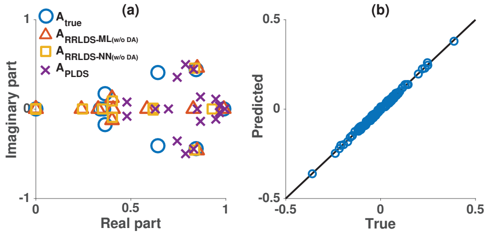


Figure 5: **(a)** The spectrum of estimated dynamics matrices using RRLDS-ML (red triangle), RRLDS-NN (yellow square) and LDS with Poisson observation model (PLDS, purple cross). The true complex eigenvalue spectrum is indicated with blue circles. Results of both RRLDS-ML/-NN methods (w/o DA) are close to true eigenvalues, while PLDS fail to eliminate redundant dimensionality. **(b)** Scatter plot of the elements in stationary covariance matrix of predicted and true count data.

Figure 5 (a) compares the eigenvalue spectrum of the estimated (by RRLDS-ML/-NN (w/o DA), PLDS) and the true dynamics matrices. The true rank of dynamics matrix is 10 (# blue circles), while the number of latent states is initialized to be 20, larger than the rank. The experiment is performed on the simulated data $\mathbf{y}$ with 40 sequences, each of which has 100 bins. It verifies that RRLDS-ML/NN (w/o DA) indeed result in a low-rank estimation of the dynamics matrix, with higher accuracy than PLDS. Specifically, PLDS overfits the data with redundant eigenvalues, instead, our method learns a rank-10 dynamics matrix from the initial 20D state space, leaving the rest of eigenvalues to be 0.

Given the estimated parameters, Figure 5 (b) plots the elements of stationary covariance matrix (Macke, Buesing, and Sahani 2015) for the predicted (using RRLDS-NN) and true count data, also demonstrating the accuracy of our estimation.

Prediction performance We compare the predictive log-likelihood of the learned systems by RRLDS-ML/-NN, DALDS and PLDS. A higher predictive log-likelihood implies better performance. Figure 6 (a) shows the results of four LDSs, each with a dynamics matrix of rank-2, 4, 6, 8,

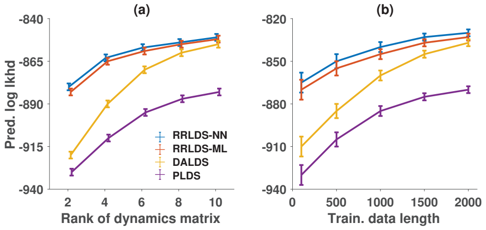


Figure 6: Predictive log-likelihood of **(a)** four learned LDSs with different true ranks of dynamics matrices, and **(b)** four rank-5 LDSs learned with different lengths of training data. The predefined number of latent states is 10 in both **(a)** and **(b)**. RRLDS-ML/-NN significantly (p <0.001, paired t-test) outperform alternatives.

and 10. The number of latent states is 10, and the dimension of simulated observations $\mathbf{y}_{1:T}$ is 40. The length is 500 for training data and 100 for testing data. Ten trials in total are simulated. DALDS outperforms PLDS due to the advantages of dispersion-adaptive distribution. RRLDS-ML/-NN further improve the prediction accuracy by alleviating overfitting with intrinsic dimensionality recovery.

In Figure 6 (b), we plot the predictive log-likelihood over the length of the training data, by fixing the rank to be 5 and the predefined number of states to be 10. Results show that when the training data become long enough, RRLDS-ML/-NN, and DALDS converge to a similar performance. It is because the eigenvalue/eigenvector pairs corresponding to redundant dimensions of dynamics matrix are estimated close to the ground-truth. The decomposition space of dynamics matrix spanned by its eigenvectors is similar for with or without regularization. However, only RRLDS-ML/-NN consistently achieve promising predictions for short-length data, which is more applicable to real-world scenarios.

Neuroscience Data

We also evaluated our method on two experimental hippocampus datasets (Mizuseki et al. 2009). These two datasets contain neuronal spike-count data in the brains of rats performing two different running tasks.

Prediction performance of neural activities Figure 7 shows prediction performances of six LDSs (RRLDS-ML/-NN, PLDS, SubspaceID, Stable LDS, and LDS) with different predefined number of latent states for Task #1. As shown by the predictive log-likelihood, while a single latent state cannot model the system well (the bars to the left), RRLDS-ML/-NN significantly ($p < 0.001$, paired t-test) outperform the alternatives.

Figure 8 compares performances for predicting spike counts of 10 neurons in 100 time bins (Task #2). The color represents the count value predicted at each time step. Results demonstrate that RRLDS can capture more details of spiking activities compared with all baselines. PLDS has better prediction than SubspaceID, StableLDS and LDS, but still loses the temporal precision compared with RRLDS. Specifically, neuron #4 (during time bins 45-65) and neuron #6 (during time bins 80-95) have high and varying firing rate (highlighted in the figure). RRLDS-NN captured it precisely, while others failed. We would like to denote that without Reduced-Rank structures and $\mathcal{DA}$ on LDS, the baseline methods predict spurious spike counts temporally.

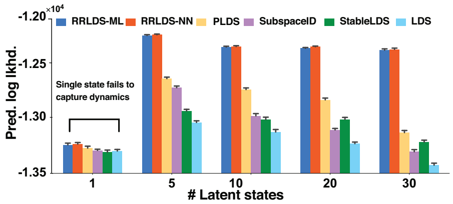


Figure 7: Predictive Log-likelihood of Experimental Spiking Activities in Task #1

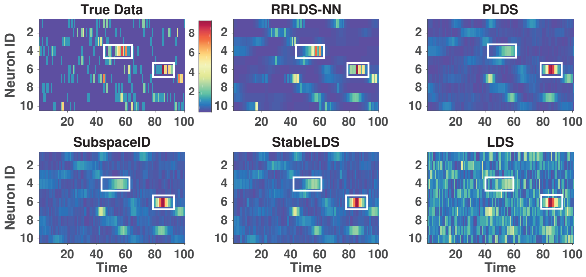


Figure 8: Prediction performance of five models for neurons' spike counts (Task #2). The rows of each subfigure indicate spiking sequence of neurons. The color highlights count values recorded/predicted at each time step.

Retrieval of intrinsic dimensionality We test the retrieval of intrinsic dimensionality for the complex neural system based on the *estimated* rank of dynamics matrix. In Figure 9, each subfigure plots the normalized eigenvalues of the dynamics matrices learned from different experimental trials. It is observed that given the same task, the rank of the optimized dynamics matrix consistently converges to 5 or 6 for Task #1 and 10 or 11 for Task #2, regardless of varying the number of latent states (10, 20 or 30).

This result provides a valuable insight into the internal factors of the neural system: the recorded spiking activities in hippocampus for the two tasks are intrinsically characterized by an underlying low-rank dynamical system.

Conclusion

We have proposed reduced-rank linear dynamical systems to retrieve the intrinsic dimensionality from *short-length*, noisy *count-valued* data. Two reduced-rank structures and a dispersion-adaptive distribution family are introduced and incorporated into our model. Both simulation and experimental results verify the effectiveness of our method in eliminating redundant latent dimensions. Extensions to nonlinear dynamical system are left for future work, which may require higher algorithmic complexity in the learning with prior $p_{\mathcal{ML}/\mathcal{NN}}(A)$ and weighted functions $w(\cdot)$. The applications of this system are not limited in neuroscience. We expect our method can benefit the learning of more concise, structured,

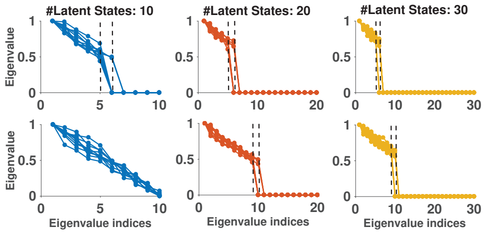


Figure 9: Latent state space recovery from neuroscience data using RRLDS-NN. Top row: Task #1; bottom row: Task #2. Different lines in each subfigure represent different trials. 10, 20, and 30 latent states are selected for testing robustness of RRLDS for retrieving intrinsic dimensionality.

and interpretable patterns from social science and financial data, which are often observed to be *short-length*, noisy and *count-valued*. We implement RRLDS in Matlab(2017a), and our code is available at https://github.com/sheqi/RRLDS

References

Angelosante, D.; Roumeliotis, S. I.; and Giannakis, G. B. 2009. Lasso-kalman smoother for tracking sparse signals. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, 181–185. IEEE.

Archer, E. W.; Koster, U.; Pillow, J. W.; and Macke, J. H. 2014. Low-dimensional models of neural population activity in sensory cortical circuits. In *NIPS*, 343–351.

Asif, M. S.; Charles, A.; Romberg, J.; and Rozell, C. 2011. Estimation and dynamic updating of time-varying signals with sparse variations. In *ICASSP, IEEE International Conference on*, 3908–3911. IEEE.

Banaee, H.; Ahmed, M. U.; and Loutfi, A. 2013. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors* 13(12):17472–17500.

Boots, B.; Gordon, G. J.; and Siddiqi, S. M. 2007. A constraint generation approach to learning stable linear dynamical systems. In *NIPS*, 1329–1336.

Buesing, L.; Machado, T. A.; Cunningham, J. P.; and Paninski, L. 2014. Clustered factor analysis of multineuronal spike data. In *NIPS*, 3500–3508.

Buesing, L.; Macke, J. H.; and Sahani, M. 2012. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *NIPS*, 1682–1690.

Charles, A.; Asif, M. S.; Romberg, J.; and Rozell, C. 2011. Sparsity penalties in dynamical system estimation. In *CISS, 2011 45th Annual Conference on*, 1–6. IEEE.

Churchland, M. M.; Byron, M. Y.; Cunningham, J. P.; Sugrue, L. P.; Cohen, M. R.; Corrado, G. S.; Newsome, W. T.; Clark, A. M.; Hosseini, P.; Scott, B. B.; et al. 2010. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience* 13(3):369–378.

del Castillo, J., and Pérez-Casany, M. 2005. Overdispersed and underdispersed poisson generalizations. *Journal of Statistical Planning and Inference* 134(2):486–500.

Gao, Y., and Yuille, A. L. 2017. Exploiting symmetry and/or manhattan properties for 3d object structure estimation from single and multiple images. *CVPR, 2017*.

Gao, Y.; Ma, J.; and Yuille, A. L. 2017. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Transactions on Image Processing* 26(5):2545–2560.

Ghahramani, Z., and Hinton, G. E. 1996. Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Hein, G.; Morishima, Y.; Leiberg, S.; Sul, S.; and Fehr, E. 2016. The brains functional network architecture reveals human motives. *Science* 351(6277):1074–1078.

Khan, M. E.; Aravkin, A.; Friedlander, M.; and Seeger, M. 2013. Fast dual variational inference for non-conjugate latent gaussian models. In *ICML*, 951–959.

Lakshmanan, K. C.; Sadtler, P. T.; Tyler-Kabara, E. C.; Batista, A. P.; and Byron, M. Y. 2015. Extracting low-dimensional latent structure from time series in the presence of delays. *Neural computation*.

Linderman, S., and Adams, R. 2014. Discovering latent network structure in point process data. In *ICML*, 1413–1421.

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

Linderman, S.; Stock, C. H.; and Adams, R. P. 2014. A framework for studying synaptic plasticity with neural spike train data. In *NIPS*, 2330–2338.

Macke, J. H.; Buesing, L.; and Sahani, M. 2015. Estimating state and parameters in state space models of spike trains. In *NIPS*. 137–159.

Mizuseki, K.; Sirota, A.; Pastalkova, E.; and Buzsáki, G. 2009. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron* 64(2):267–280.

Paninski, L.; Ahmadian, Y.; Ferreira, D. G.; Koyama, S.; Rad, K. R.; Vidne, M.; Vogelstein, J.; and Wu, W. 2010. A new look at state-space models for neural data. *Journal of computational neuroscience* 29(1-2):107–126.

Pfau, D.; Pnevmatikakis, E. A.; and Paninski, L. 2013. Robust learning of low-dimensional dynamics from large neural ensembles. In *NIPS*, 2391–2399.

She, Q.; Chen, G.; and Chan, R. H. 2016. Evaluating the small-world-ness of a sampled network: Functional connectivity of entorhinal-hippocampal circuitry. *Scientific reports* 6.

She, Q.; Jelfs, B.; and Chan, R. H. 2016. Modeling short over-dispersed spike count data: A hierarchical parametric empirical bayes framework. *arXiv preprint arXiv:1605.02869*.

She, Q.; So, W. K.; and Chan, R. H. 2015. Reconstruction of neural network topology using spike train data: Small-world features of hippocampal network. In *EMBC, 2015*, 2506–2509. IEEE.

She, Q.; So, W. K.; and Chan, R. H. 2016. Effective connectivity matrix for neural ensembles. In *EMBC, 2016*, 1612–1615. IEEE.

So, W. K.; Yang, L.; Jelfs, B.; She, Q.; Wong, S. W.; Mak, J. N.; and Chan, R. H. 2016. Cross-frequency information transfer from eeg to emg in grasping. In *EMBC, 2016*, 4531–4534. IEEE.

Spira, M. E., and Hai, A. 2013. Multi-electrode array technologies for neuroscience and cardiology. *Nature nanotechnology* 8(2):83–94.

Sussillo, D.; Jozefowicz, R.; Abbott, L.; and Pandarinath, C. 2016. Latent factor analysis via dynamical systems. *arXiv:1608.06315*.

Van Overschee, P., and De Moor, B. 2012. *Subspace identification for linear systems: Theory Implementation Applications*. Springer Science & Business Media.