

A DISENTANGLED LOW-RANK RNN FRAMEWORK FOR UNCOVERING NEURAL CONNECTIVITY AND DYNAMICS

Anonymous authors

Paper under double-blind review

ABSTRACT

Low-rank recurrent neural networks (lrRNNs) are a class of models that uncover low-dimensional latent dynamics underlying neural population activity. Although their functional connectivity is low-rank, it lacks disentanglement interpretations, making it difficult to assign distinct computational roles to different latent dimensions. To address this, we propose the **Disentangled Recurrent Neural Network (DisRNN)**, a generative lrRNN framework that assumes group-wise independence among latent dynamics while allowing flexible within-group entanglement. These independent latent groups allow latent dynamics to evolve separately, but are internally rich for complex computation. We reformulate the lrRNN under a variational autoencoder (VAE) framework, enabling us to introduce a partial correlation penalty that encourages disentanglement between groups of latent dimensions. Experiments on synthetic, monkey M1, and mouse voltage imaging data show that DisRNN consistently improves the disentanglement and interpretability of learned neural latent trajectories in low-dimensional space and low-rank connectivity over baseline lrRNNs that do not encourage partial disentanglement.

1 INTRODUCTION

Understanding neural dynamics and connectivity from high-dimensional recordings is a central challenge in neuroscience. Existing approaches typically fall into two categories: (1) methods that extract low-dimensional latent dynamics or representations from neural population activity (Yu et al., 2005; Wu et al., 2017; Pandarinath et al., 2018; She & Wu, 2020; Aoi et al., 2020); and (2) models that operate directly in the high-dimensional neural space, enabling inference of connectivity structures (Pillow et al., 2008; Linderman et al., 2016; Roudi et al., 2015; Li et al., 2023). While the former provide insights into neural computation, they offer limited access to interpretable connectivity. The latter, meanwhile, do not directly estimate low-dimensional structure, which often leads to suboptimal discovery of the underlying neural dynamics.

Low-rank recurrent neural networks (lrRNNs) (Mastrogiuseppe & Ostojic, 2018; Beiran et al., 2021; Dubreuil et al., 2022; Schuessler et al., 2020a;b; Valente et al., 2022a) offer a promising middle ground, capturing both structured connectivity and low-dimensional dynamics by constraining the recurrent weights to a low-rank form. Recent methods, such as LINT (Valente et al., 2022b) and other related approaches (Pals et al., 2024), leverage the singular value decomposition (SVD) of low-rank connectivity matrices to extract interpretable latent subspaces. In these models, the left singular vectors define the latent dynamics, enabling both predictive modeling and functional interpretation.

However, SVD yields orthogonal components, which are not necessarily independent. In many experimental settings, it is desirable to identify independent latent subspaces that evolve separately. For example, in a motor decision-making task, Mante et al. (2013) demonstrated that latent trajectories under different task contexts (e.g., color vs. shape) are structured separately, suggesting that neural computations may arise from multiple independent sources.

To address this, techniques such as independent component analysis (ICA) (Hyvärinen & Oja, 2000) and disentangled variational autoencoders (VAEs) (Chen et al., 2018; Kim & Mnih, 2018) have been used to enforce independence. However, these approaches have two key limitations: (1) they are not dynamical models and therefore do not infer connectivity or recurrent dynamics; (2) they typically assume dimension-wise independence in the latent space, so that each latent dimension evolves inde-

pendently, which is often too restrictive. For example, visual neurons have been shown to represent information in a factorized manner, encoding features such as texture and contrast separately (Lee et al., 2025). While contrast might be able to be represented in a one-dimensional latent subspace, texture is inherently more complex and requires a higher-dimensional neural representation for accurate encoding. These reasons highlight that merely applying ICA or disentangled VAE to the neural data may not be accurate and insightful enough for disentangling task-relevant neural subspaces.

To address these limitations, we introduce the **Disentangled Recurrent Neural Network (DisRNN)**—a low-rank RNN framework that captures both neural connectivity and recurrent dynamics while enforcing disentanglement between groups of latent dimensions (i.e., several latent subspaces) (Li et al., 2025). This approach offers three key benefits:

- It is a generative RNN model that enforces group-wise independence among latent dynamics, while allowing flexible interactions within each group.
- It yields interpretable low-rank sub-connectivities associated with each latent group, which can be viewed as distinct neural sub-circuits driving independent sources of task-related neural signals.
- Built on the VAE framework, it is flexible and extensible to complex RNN architectures.

2 BACKGROUND

Low-rank RNN. lrRNN assumes that the observed neural sequence $\mathbf{x}(t) \in \mathbb{R}^N$ from N neurons evolves with

$$\frac{d\mathbf{x}(t)}{dt} = -\mathbf{x}(t) + \mathbf{W} \int_0^\infty \psi(\tau) \sigma(\mathbf{x}(t-\tau)) d\tau + \mathbf{b} + \boldsymbol{\eta}(t) + \boldsymbol{\epsilon}(t), \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the low-rank connectivity matrix, $\mathbf{b} \in \mathbb{R}^N$ is the neuron background intensity vector, $\psi(\tau)$ is a history convolution kernel (with $\int_0^\infty \psi(\tau) d\tau = 1$), σ is a nonlinear activation function (e.g., $\tanh(\cdot)$), $\boldsymbol{\eta}(t) \in \mathbb{R}^N$ are the external inputs, and $\boldsymbol{\epsilon}(t) \in \mathbb{R}^N$ are i.i.d. noise samples (e.g., Gaussian).

In practice, learning such a stochastic differential equation usually relies on its discretized version

$$\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(t-L)} = \mathbf{W} \sum_{l=1}^L \psi_l \sigma(\mathbf{x}^{(t-l)}) + \mathbf{b} + \boldsymbol{\eta}^{(t)} + \boldsymbol{\epsilon}^{(t)}, \quad (2)$$

where $\{\mathbf{x}^{(t)}\}_{t=1}^T$ is the discretized neural sequence in T time bins. $\psi \in \mathbb{R}_{\geq 0}^L$ is the history convolution kernel (with $\sum_{l=1}^L \psi_l = 1$), and the remaining parameters are defined analogously.

To obtain latent dynamics $\mathbf{z}(t)$ given the rank- K weight matrix \mathbf{W} , LINT (Valente et al., 2022b) parameterized $\mathbf{W} = \mathbf{A}\mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{N \times K}$ is the left singular matrix, and $\mathbf{B} \in \mathbb{R}^{K \times N}$ is the transposed and singular-value-scaled right singular matrix. Then they projected $\mathbf{x}(t)$ to $\mathbf{z}(t)$ via \mathbf{B} .

VAE for low-dimensional latent. A variational auto-encoder (VAE) (Kingma, 2013) with linear encoder $q(\mathbf{z}|\mathbf{x})$ and linear decoder $p(\mathbf{x}|\mathbf{z})$ can be viewed as a dimensionality reduction tool similar to probabilistic principal component analysis (PPCA), used to find the low-dimensional neural latent $\mathbf{z} \in \mathbb{R}^K$. The prior $p(\mathbf{z})$ is typically chosen as a standard normal prior.

To fit this VAE model, we optimize the standard evidence lower bound (ELBO) (Blei et al., 2017):

$$\frac{1}{T} \sum_{t=1}^T \text{ELBO}(\mathbf{x}^{(t)}) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{q(\mathbf{z}^{(t)}|\mathbf{x}^{(t)})} \left[\ln p(\mathbf{x}^{(t)}|\mathbf{z}^{(t)}) \right] - \text{KL}(q(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}) || p(\mathbf{z}^{(t)})). \quad (3)$$

However, the inferred latent components z_1, \dots, z_K are generally entangled, and any invertible affine transformation $\mathbf{P} \in \mathbb{R}^{K \times K}$ can form an equivalent solution $\mathbf{P}^{-1}\mathbf{z}$. To see this, assume the encoder is $\mathbf{z} = \mathbf{B}\mathbf{x} + \mathbf{d}$ and the decoder is $\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{c}$. Without noise for simplicity, $\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{c} = \mathbf{A}(\mathbf{B}\mathbf{x} + \mathbf{d}) + \mathbf{c} = \mathbf{A}\mathbf{P}(\mathbf{P}^{-1}\mathbf{B}\mathbf{x} + \mathbf{P}^{-1}\mathbf{d}) + \mathbf{c}$, so the transformed $\mathbf{z}' = \mathbf{P}^{-1}\mathbf{B}\mathbf{x} + \mathbf{P}^{-1}\mathbf{d} = \mathbf{P}^{-1}\mathbf{z}$ is the equivalent latent under the encoder $(\mathbf{P}^{-1}\mathbf{B}, \mathbf{d})$ and decoder $(\mathbf{A}\mathbf{P}, \mathbf{c})$.

3 DISENTANGLED RECURRENT NEURAL NETWORK (DISRNN)

3.1 REFORMULATE LOW-RANK RNN USING VAE

Reformulation. Learning an independent latent structure for lrRNN is not as simple as learning an orthogonal latent structure, which can be obtained via specific decompositions such as SVD. To

address this, we first reformulate the lrRNN using the aforementioned VAE framework by adding nonlinearity and a history convolution $\psi \in \mathbb{R}_{\geq 0}^L$ with $\sum_{l=1}^L \psi_l = 1$ to the original linear encoder:

$$\mathbf{z}^{(t)} | \mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(t-L)} = \mathbf{B} \sum_{l=1}^L \psi_l \sigma(\mathbf{x}^{(t-l)}) + \mathbf{d} + \epsilon_{\text{enc}}^{(t)}, \quad (4)$$

where $\mathbf{B} \in \mathbb{R}^{K \times N}$ and $\mathbf{d} \in \mathbb{R}^K$ are the encoder parameters, and $\epsilon_{\text{enc}}^{(t)}$ are i.i.d. encoder noise samples. We then write the decoder as

$$\mathbf{x}^{(t)} | \mathbf{z}^{(t)} = \mathbf{A} \mathbf{z}^{(t)} + \mathbf{c} + \boldsymbol{\eta}^{(t)} + \epsilon_{\text{dec}}^{(t)}, \quad (5)$$

where $\mathbf{A} \in \mathbb{R}^{N \times K}$ and $\mathbf{c} \in \mathbb{R}^N$ are the decoder parameters, and $\epsilon_{\text{dec}}^{(t)}$ are i.i.d. decoder noise samples. Combining encoder and decoder, we obtain a generative lrRNN form for $\mathbf{x}^{(t)}$:

$$\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(t-L)} = \mathbf{A} \mathbf{B} \sum_{l=1}^L \psi_l \sigma(\mathbf{x}^{(t-l)}) + (\mathbf{A} \mathbf{d} + \mathbf{c}) + \boldsymbol{\eta}^{(t)} + (\mathbf{A} \epsilon_{\text{enc}}^{(t)} + \epsilon_{\text{dec}}^{(t)}). \quad (6)$$

This can be viewed as an L -th order lrRNN (L history time steps dependency), where the rank- K connectivity is $\mathbf{W} := \mathbf{A} \mathbf{B} \in \mathbb{R}^{N \times N}$ and the background intensity vector is $\mathbf{b} := (\mathbf{A} \mathbf{d} + \mathbf{c}) \in \mathbb{R}^N$.

Latent dynamics. A key benefit of expressing lrRNN in the VAE framework is that it allows deriving latent dynamics by replacing $\{\mathbf{x}^{(t-L)}, \dots, \mathbf{x}^{(t-1)}\}$ in Eq.4 with the form in Eq.5:

$$\mathbf{z}^{(t)} | \mathbf{z}^{(t-1)}, \dots, \mathbf{z}^{(t-L)} = \mathbf{B} \sum_{l=1}^L \psi_l \sigma(\mathbf{A} \mathbf{z}^{(t-l)} + \mathbf{c} + \boldsymbol{\eta}^{(t-l)} + \epsilon_{\text{dec}}^{(t-l)}) + \mathbf{d} + \epsilon_{\text{enc}}^{(t)}. \quad (7)$$

In the special case when σ is the identity function, the latent dynamics simplify to

$$\mathbf{z}^{(t)} = \mathbf{B} \mathbf{A} \sum_{l=1}^L \psi_l \mathbf{z}^{(t-l)} + (\mathbf{B} \mathbf{c} + \mathbf{d}) + \left(\mathbf{B} \sum_{l=1}^L \psi_l \boldsymbol{\eta}^{(t-l)} \right) + \left(\mathbf{B} \sum_{l=1}^L \psi_l \epsilon_{\text{dec}}^{(t-l)} + \epsilon_{\text{enc}}^{(t)} \right), \quad (8)$$

where $\mathbf{J} := \mathbf{B} \mathbf{A} \in \mathbb{R}^{K \times K}$ denotes the corresponding latent recurrent connectivity.

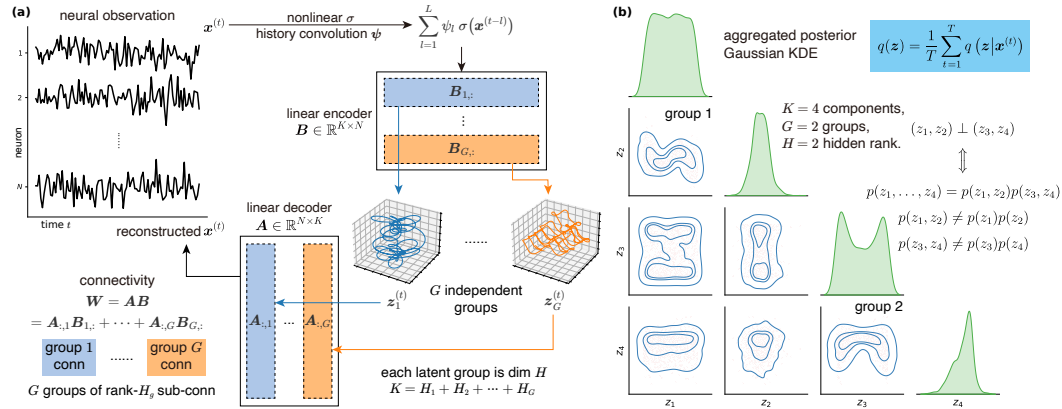


Figure 1: (a): Schematic of the DisRNN, showing independent latent groups and the corresponding low-rank connectivities. (b): Group-wise independence: $(z_1, z_2) \perp (z_3, z_4)$, while within-groups are highly entangled, and components from different groups are marginally independent.

3.2 DISENTANGLED RECURRENT NEURAL NETWORK (DISRNN)

Definition of disentanglement. Another benefit of formulating lrRNN under the VAE framework is that the low-dimensional latent variable $\mathbf{z}^{(t)}$ from the encoder distribution is naturally amenable to disentanglement. We assume latent $\mathbf{z} \in \mathbb{R}^K$ is disentangled by G groups, each with internal group rank H_g , satisfying $K = H_1 + H_2 + \dots + H_G$. For simplicity, we denote the g -th group as $\mathbf{z}_g = (z_{g,1}, z_{g,2}, \dots, z_{g,H_g}) \in \mathbb{R}^{H_g}$, $\forall g \in \{1, \dots, G\}$, so that $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_G)$. Then, the disentanglement is defined as:

$$\prod_{g=1}^G z_g \iff p(\mathbf{z}) = \prod_{g=1}^G p(\mathbf{z}_g), \quad p(\mathbf{z}_g) \neq p(z_{g,1}) \cdots p(z_{g,H_g}), \quad \forall g \in \{1, \dots, G\}. \quad (9)$$

This equation expresses that latent components within a group may exhibit dependencies and may not be further disentangled, while different groups are mutually independent (see Fig. 1(b)).

Inference. We introduce two inference approaches to achieve partial disentanglement: one tailored for linear dynamics (Eq. (8)); while the other is flexible for general non-linear dynamics (Eq. (7)).

Inference via block-diagonal latent structure. To require group-wise independence for the linear case in Eq. (8), a straightforward way is to constrain the latent recurrence matrix $\mathbf{J} = \mathbf{BA}$ to be block-diagonal, so that latent groups do not interact with each other. If latent components within a group are entangled with each other (non-separable), the block will be similar to a Jordan matrix, i.e., the block cannot be diagonalized (see Appendix. A.1 for more details). Therefore, we can either: (1) freely learn \mathbf{A} and \mathbf{J} and solve a least-squares problem $\mathbf{J} = \mathbf{BA}$ to recover \mathbf{B} ; or (2) jointly learn \mathbf{A} and \mathbf{B} while penalizing the off-block-diagonal elements of $\mathbf{J} = \mathbf{BA}$ to promote group-wise independence. We call this approach **block-diagonal RNN (bdRNN)**.

Inference via partial correlation. For the nonlinear case in Eq. (7), there is no \mathbf{J} matrix due to the nonlinear σ function. Thus, we directly deal with Eq. (9). Following Li et al. (2025), we achieve group-wise independence by optimizing the target function

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^T \text{ELBO}(\mathbf{x}^{(t)}) - \beta \cdot \text{KL} \left(q(\mathbf{z}) \parallel \prod_{g=1}^G q(\mathbf{z}_g) \right). \quad (10)$$

The second term in Eq. (10) is the partial correlation (PC), where the aggregated posterior $q(\mathbf{z}) = \frac{1}{T} \sum_{t=1}^T q(\mathbf{z} | \mathbf{x}^{(t)}) = \sum_{t=1}^T q(\mathbf{z} | \mathbf{x}^{(t)}) q(\mathbf{x}^{(t)})$ is defined as in Makhzani et al. (2015). Since each data point is equally contributed, $q(\mathbf{x}^{(t)}) = \frac{1}{T}$ and hence $q(\mathbf{z})$ can be viewed as a Gaussian kernel density estimation over $\{\mathbf{z}^{(t)}\}_{t=1}^T$ in latent space. PC penalizes dependency between latent groups: when $q(\mathbf{z}) = \prod_{g=1}^G q(\mathbf{z}_g)$, $\text{PC} = \text{KL} \left(q(\mathbf{z}) \parallel \prod_{g=1}^G q(\mathbf{z}_g) \right) = 0$; otherwise, $\text{PC} > 0$ and the penalty is scaled by a hyperparameter $\beta > 0$. We call this approach **Disentangled RNN (DisRNN)**.

Sub-circuit connectivity. With DisRNN, we can obtain a partially disentangled latent \mathbf{z} , where each group g has its group-rank H_g . Specifically, we write

$$\mathbf{x}^{(t)} = \mathbf{A}\mathbf{z}^{(t)} + \mathbf{c} + \boldsymbol{\eta}^{(t)} + \boldsymbol{\epsilon}^{(t)} = [\mathbf{A}_{1,:}, \dots, \mathbf{A}_{G,:}] [\mathbf{z}_1^{(t)}, \dots, \mathbf{z}_G^{(t)}]^T + \mathbf{c} + \boldsymbol{\eta}^{(t)} + \boldsymbol{\epsilon}^{(t)}, \quad (11)$$

where $\mathbf{A}_{g,:}$ is the H_g -dimensional embedding of the g -th latent group subspace in the observational space \mathbb{R}^N . Accordingly,

$$\mathbf{W} = \mathbf{AB} = \sum_{g=1}^G \mathbf{A}_{:,g} \mathbf{B}_{g,:} =: \sum_{g=1}^G \mathbf{W}_g, \quad (12)$$

where \mathbf{W}_g is the rank H_g sub-connectivity associated with the g -th group. Each \mathbf{W}_g represents a sub-circuit within the neural population that encodes an independent source of neural activity. With task-related labels for stimuli or choices, for example, we can identify which sub-circuits encode stimulus features and which encode choice signals. This distinction is essential for understanding how different neural circuits support perception, decision-making, and the functional organization of the brain. Fig. 1(a) is a complete schematic of the overall framework.

Benefits. There are three main benefits of DisRNN.

- First, this partial disentanglement term in DisRNN in fact encompasses RNNs of no disentanglement when $G = 1$ (and $\text{PC} \equiv 0$, a standard lrRNN) and full disentanglement when $G = K$ (the component-wise independence in ICA (Hyvärinen & Oja, 2000), β -TCVAE, (Chen et al., 2018) and FactorVAE (Kim & Mnih, 2018), although very restrictive).
- Second, unlike bdRNN, by just setting a large enough group-rank H_g , DisRNN can automatically detect the effective/true group rank H'_g and introduce dummy dimensions $H_g - H'_g$.
- Third, DisRNN is compatible with both linear (Eq. (8)) and nonlinear (Eq. (7)) dynamics, and naturally extends to more general nonlinear dynamical systems of the form $\mathbf{x}^{(t)} = f(\mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(t-L)}) + \boldsymbol{\epsilon}^{(t)}$. Even in the absence of explicit connectivity in such a general nonlinear form, DisRNN still enables learning a disentangled low-dimensional latent representation with a nonlinear encoder. Therefore, DisRNN serves as a more general inference framework than bdRNN.

4 EXPERIMENTS

Methods for comparison. We compare RNN-based methods that emphasize different latent structure representations, without involving their underlying sampling or inference strategies.

- **lrRNN**: The standard low-rank RNN model in Eq. (2).
- **LINT** (Valente et al., 2022b): The lrRNN model with connectivity matrix parameterized by SVD.
- **SMC** (Pals et al., 2024): They proposed a Sequential Monte Carlo (SMC) inference approach to fit the LINT model to data.
- **bdRNN**: Our block-diagonal RNN model, which enforces \mathbf{BA} to be block-diagonal and is theoretically valid only when σ is the identity function.
- **DisRNN**: Our DisRNN model in Eq. (10), which achieves latent disentanglement by penalizing the partial correlation (PC) term.

4.1 SYNTHETIC DATASET

Dataset. To generate the latent, we simulate Lorenz and Thomas’ cyclically symmetric dynamics with $\Delta t = 0.1$ for 2000 steps using RK4 (Dormand & Prince, 1980). Fig. 5 in Appendix A.2.1 visualizes the latent plots. To simulate $\mathbf{x}^{(t)}$, we need to make sure the dataset also satisfies the recurrent relationships in Eq. (6) and Eq. (7). Since $\mathbf{z}^{(t)}$ follows the generative process in Eq. (7), we can get the parameters including \mathbf{A} , \mathbf{B} , and hence generate the observed data $\mathbf{x}^{(t)}$ using the fitted \mathbf{A} and \mathbf{B} with inputs and random Gaussian noises. The observation dimensionality is $N = 20$.

Experimental setup. Since the ground truth contains two independent latent groups of rank 3, we fit bdRNN and DisRNN with $(G, H) = (2, 3)$. To conduct a more comprehensive experiment, we also include two additional intermediate methods: (1) **lrRNN+ICA** that performs a post-hoc ICA on the latent estimated from lrRNN; and (2) **DisRNN-full** that runs DisRNN with $(G, H) = (6, 1)$ (i.e., full latent distanglement via DisRNN). All methods are trained for 5000 epochs using the Adam optimizer (Kingma, 2014) with a learning rate of 10^{-3} and a batch size of 128. The ablation study in Appendix A.2.1 cross-validates $\beta = 20$ and demonstrates the flexibility of (G, H) setting as the second benefit of DisRNN. All methods are run 10 times with different random seeds.

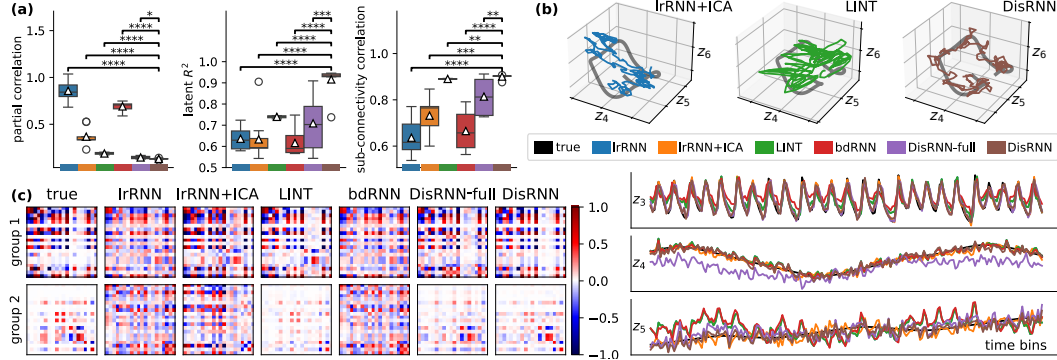


Figure 2: (a): The PC and R^2 of the estimated latent, and the connectivity correlation. The stars indicate the pairwise t -test significance levels. Arrows indicate the higher or lower the better. (b): Group 2 latent trajectories with the true dynamics in 3D plots; and the 1D trajectories from different methods on selected latent components. All latent dimensions are plotted in Fig. 8 in Appendix A.2.1. (c): The learned sub-connectivities from different methods and the ground truth.

Latent evaluations. To evaluate the estimated latent unsupervisedly, we compute the PC of the estimated latent on the test set to check whether different methods uncover desired group structures. Fig. 2(a) shows that DisRNN achieves the lowest PC, indicating that it successfully recovers partially disentangled latent dynamics. Comparing DisRNN-full and lrRNN+ICA, we see that end-to-end training in DisRNN-full can lead to better latent disentanglement. In contrast, the latent space of lrRNN, without any disentanglement constraints, may not be disentangleable, especially when the VAE structure contains non-linearities which makes post-hoc ICA unreliable or even non-decomposable. In other words, the disentanglement objective in an end-to-end model influences both the encoder and decoder during training, whereas imposing disentanglement post hoc after learning the latent representation is ineffective.

Given we know the true latent dynamics $\{z^{(t)} \in \mathbb{R}^K\}_{t=1}^T$ in this synthetic dataset, we can align the estimated latent dynamics to the ground truth. In general, if we have G groups, we can match the estimated latent groups $z_1^{(t)}, \dots, z_G^{(t)}$ to the true latent groups $z_1'^{(t)}, \dots, z_G'^{(t)}$, as illustrated in Fig. 7 in Appendix. Specifically, we create an $\mathbf{R2} \in (-\infty, 1]^{G \times G}$ matrix where $R2_{g_1, g_2}$ is the R^2 score of aligning the estimated latent $z_{g_2}^{(t)}$ to the true latent $(z')_{g_1}^{(t)}$ via an affine transformation, by solving a least squares problem. Then, the best match is obtained by finding a mutually exclusive assignment from true groups g' to estimated groups g that maximizes the total R^2 score. This is essentially a linear sum assignment problem in graph theory (Crouse, 2016). After the assignment, we report the average R^2 over the finally matched pairs.

Results. Although different methods reconstruct similar observation sequences (all methods are approximately 0.85 reconstruction R^2 through their learned recurrent weights), their uncovered latent dynamics differ from each other in terms of disentanglement structure. The latent R^2 in Fig. 2(a) shows that DisRNN recovers the latent dynamics with the highest accuracy, while standard lrRNN and LINT perform noticeably worse. Among all methods, only DisRNN faithfully recovers the partially disentangled structure present in the ground truth. Fig. 2(b) visualize the estimated latent dynamics, confirming that DisRNN’s latent trajectories are better aligned with the ground truth than others. To further explore the estimated latent from methods without group-wise independence assumption, we perform another post-hoc analysis in Fig. 9 in Appendix A.2.1, demonstrating that partitioning latent components from methods without the assumption of group-wise structure has explosively high complexity, which is a theoretical limitation for the baseline methods.

In terms of parameter estimation, although all methods yield similar estimates of the overall recurrent connectivity $\mathbf{W} = \mathbf{AB}$ (with connectivity correlations around 0.85), they find different \mathbf{A} and \mathbf{B} . Particularly, the sub-connectivities discovered by DisRNN match the ground truth the best (Fig. 2(a) and (c)).

4.2 MONKEY M1 DATA

Dataset. We use neural spike train recordings from the macaque M1 cortex (Ma et al., 2023). The dataset consists of firing rate data from 168 trials, 14 time bins, and 154 neurons, recorded while the animal performed a center-out reaching task with eight movement directions.

Experimental setup. Since the ground-truth latent structure is unknown, we explore a range of model configurations based on $K = 2$ and $K = 4$ latent dimensions and analyze their outcomes.

- For $K = 2$, we learn lrRNN and LINT. We also learn a DisRNN with two independent rank-1 latent groups $(G, H) = (2, 1)$. Additionally, since our disentangled RNN supports arbitrary encoder and decoder architectures as illustrated in the inference section, we also try a DisRNN model with an MLP encoder and decoder, denoted as DisRNN-MLP.
- For $K = 4$, we again learn lrRNN, LINT. We include the comparison with SMC, since SMC was also applied to this dataset in Pals et al. (2024). For DisRNN, we learn two independent rank-2 latent groups $(G, H) = (2, 2)$.

The training procedures are similar to the synthetic experiments (see our code for details). After training, we attempt to align the learned latent groups to x -coordinates (horizontal) and y -coordinates (vertical) of the monkey’s hand movement trajectories (the true trajectories in Fig. 3(a)).

Results. For $K = 2$, lrRNN and LINT perform poorly in aligning to the x and y movement separately (Fig. 3). With disentanglement, DisRNN improves the alignment score from about 0.5 to 0.65. However, the trajectories remain visually poor, suggesting that rank-1 latent dynamics are not expressive enough to capture the M1 neural activities responsible for the reaching task.

With more components $K = 4$ in a larger latent space, lrRNN and LINT achieve improved alignment, but still lack any form of latent disentanglement. Although SMC is a more advanced method than LINT due to its sophisticated inference approach, their reported gains in the original paper relied on using direction labels. In our setting, where such labels are not provided, the sophisticated inference alone makes SMC perform even worse than LINT. In contrast, DisRNN with two rank-2 latent groups achieves an alignment R^2 score of about 0.8. The trajectories to all eight directions are significantly better separated than all other configurations. We also tried DisRNN with 4 rank-1 groups. Due to its rigid full disentanglement assumption, it fails to group components meaningfully

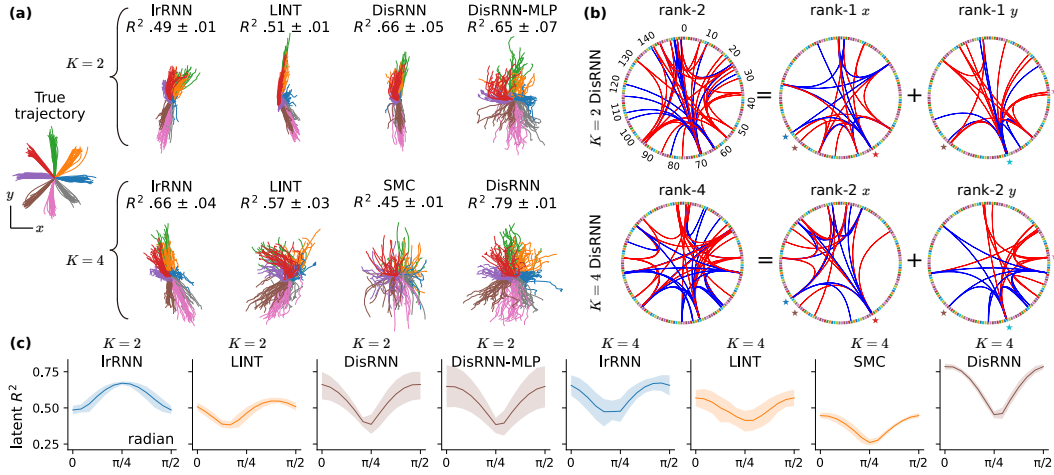


Figure 3: (a): Recovered latent trajectories from different methods v.s. the ground-truth trajectories, and their alignment R^2 scores. (b): The corresponding low-rank connectivity for $K = 2$ DisRNN and $K = 4$ DisRNN. For DisRNN, for example, the rank-4 connectivity can be decomposed into two rank-2 sub-connectivities, one responsible for the dynamics of the x -movement, and the other for the y -movement. (c): The R^2 scores of aligning the estimated latent to rotated coordinate systems.

into x and y -aligned latent subspaces, resulting in a lower alignment score. These results indicate that higher-than-rank-1 is important to model the dynamics of the x or y coordinates effectively.

To understand whether the poor disentanglement in $K = 2$ stems from the linear architecture of the encoder and decoder, we replace the linear encoder and decoder with more complicated MLP encoder and decoder, denoted as DisRNN-MLP. Its alignment R^2 remains lower than that of the DisRNN with two rank-2 groups. In particular, DisRNN-MLP fails to separate trajectories to 45° and 90° (colored by orange, green), and the same for 135° and 180° (colored by red and purple), which remain mixed. Besides, DisRNN-MLP does not support explicit connectivity estimation due to its nonlinear encoder and decoder. This further supports that higher-than-rank-1 latent groups are necessary for describing the x or y movement dynamics, and that simply having a more complicated VAE architecture might not resolve the disentanglement limitations of low-ranked groups.

We also compare the learned sub-connectivity matrices governing the dynamics for both x and y coordinates. Take the $K = 2$ DisRNN and the $K = 4$ DisRNN for example, both have two independent groups, but the $K = 4$ DisRNN has a higher-than-1 within-group rank (Fig. 3(b)). With disentanglement, the connectivity matrix can also be separated into two groups, one for the x coordinate and one for the y coordinate. Comparing the sub-connectivities between $K = 2$ and $K = 4$, they share some similar connections (indicated by the arcs between the stars in Fig. 3(b)). However, the rank-2 sub-connectivities are more expressive, which better support accurate approximation of the corresponding x coordinate and y coordinate latent dynamics. *This enables circuit-level hypotheses, such as one subnetwork supporting horizontal control and another vertical, and reveals how neural populations implement multiple variables through separable low-rank channels rather than a single entangled circuit.* It also explains why ICA-like post hoc separation may fail, since the separation lies in the structured connectivity itself rather than only in latent activity. The learned connectivity matrices from all configurations are shown in Fig. 10 in Appendix A.2.2.

At the beginning of this experiment, we hypothesized that the x and y coordinates represent two underlying true latent groups. However, the eight directions in this experiment are rotationally symmetric, mathematically. This may cause the concern that we can align the disentangled latent groups with respect to an arbitrarily rotated coordinate system (e.g., express the trajectory coordinate under the bases of direction $\frac{\pi}{4}$ and $\frac{3\pi}{4}$), instead of the horizontal/vertical one. To test this, we align the disentangled latent to the trajectories expressed by various rotated coordinate systems, ranging from 0 to $\frac{\pi}{2}$. If there is no disentanglement, the alignment score should be similar across different angles due to the rotational symmetry.

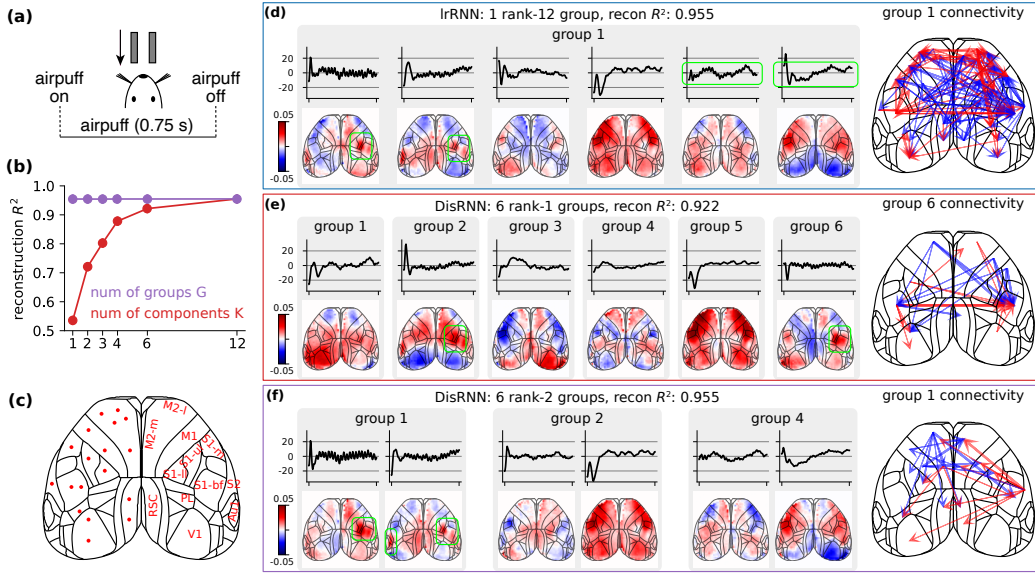


Figure 4: (a): Experiment setup. (b): Reconstruction performances w.r.t. different numbers of components for DisRNN or different numbers of groups for DisRNN. (c): Cortical map with region abbreviations. (d): 1 rank-12 group, a no disentanglement configuration, i.e., a standard low-rank RNN. Left are brain maps $A_{:,g}$ and the corresponding time series $z_g^{(t)}$. Right is the corresponding rank-12 connectivity corresponding to group 1. Similar for (e) and (f).

Fig. 3(c) indicates that for the three DisRNNs, the best alignment occurs under the canonical horizontal-vertical bases (highest latent R^2 at 0 and $\frac{\pi}{2}$), and the alignment R^2 score drops significantly under the rotated bases, especially near $\frac{\pi}{4}$. In contrast, methods without disentanglement (standard IrRNN and LINTs) show flatter and more erratic alignment curves across different rotation angles, lacking a consistent preference for particular bases. These findings are biologically plausible, consistent with prior neurophysiological findings in primate motor cortex (Georgopoulos et al., 1982; Amirkian & Georgopoulos, 2003; Churchland et al., 2012), where neurons are not uniformly tuned to all directions, but more tuned along the cardinal axes, i.e., horizontal (left-right) and vertical (up-down), due to the body symmetry. Especially in these center-out reaching tasks, it is more natural for primates to understand and execute movements under the horizontal-vertical bases (i.e., the Cartesian coordinate system), rather than the polar coordinate system. *This reinforces the idea that M1 does not encode movement uniformly across directions, and certain axes may be overrepresented due to behavioral, biomechanical, habitual, or neural factors.* Taken together, these results support that the separation of x and y movement in our model is not merely an artifact of the model statistics, but may reflect a meaningful subspace underlying the neural data. Results from fitting the model to Poisson spike counts (Appendix A.2.2) similarly demonstrate that disentanglement enhances latent structure separation, yielding better alignment with hand movement trajectories.

4.3 MOUSE DORSAL CORTEX VOLTAGE IMAGING DATA

Dataset and experimental setup. The dataset used in this study is a trial-averaged voltage imaging sequence from a mouse Lu et al. (2023), comprising 150 frames of 50×50 dorsal cortex voltage images recorded during a left-side air puff stimulus lasting 0.75 seconds (Fig. 4(a)). Although DisRNN can in principle handle over-specified group ranks, overly large settings impair efficiency and model quality. Since the true latent structure is unknown, fixing the total latent dimensionality K while varying the group specification (G, H) provides a controlled way to examine how different assumptions affect model performance and the latent interpretation. Thus, we fix $K = 12$ and investigate different numbers of groups $G \in \{1, 2, 3, 4, 6, 12\}$. When $G = 1$, there is no disentanglement, which corresponds to the IrRNN. Additionally, we explore fully disentangled models by varying $K \in \{1, 2, 3, 4, 6, 12\}$ with $G = K$ and $H = 1$. The training procedures are similar to the previous experiments (see code for details).

Results. Fig. 4(b) shows that increasing the total number of latent components in DisRNN improves reconstruction, indicating that a larger latent dimensionality preserves more information in the latent space. The improvement plateaus once the number of components reaches six, suggesting that a latent space of dimension six or higher is sufficient to capture the data structure. In comparison, fixing $K = 12$ maintains consistently high reconstruction performance across different group numbers. Varying the number of groups in this case does not affect observation reconstruction but leads to different degrees of latent disentanglement. This indicates that once sufficient latent capacity is available, the DisRNN framework allows us to explore and control the disentanglement structure via group specification—without compromising reconstruction accuracy.

To evaluate disentanglement, three configurations are selected and compared in Fig. 4(d-f), with a standard cortical region map and annotated ROIs (regions of interest) shown in Fig. 4(c). In each of Figs. 4(d-f), the left panel displays the brain map $\mathbf{A}_{:,g}$ and the corresponding time series $\mathbf{z}_g^{(t)}$ for each group g , while the right panel shows the associated group connectivity \mathbf{W}_g . In Fig. 4(d), learning a standard lrRNN without disentanglement results in mixed latent trajectories, where airpuff-related S1-bf (primary somatosensory-barrel field) oscillations (marked by the green squares) appear across multiple components. In contrast, DisRNN with 6 rank-1 groups ($(K, G) = (6, 1)$) in Fig. 4(e) isolates these oscillations in the right S1-bf and right M2-m (secondary motor-medial), yielding a more localized representation. However, 6 components are less expressive compared to 12 in terms of reconstruction R^2 . To address this, a DisRNN with 6 rank-2 groups ($(K, G) = (6, 2)$) in Fig. 4(f) is explored. In this configuration, airpuff-related oscillations are concentrated within group 1 and are not further separable. To quantify disentanglement, total correlation (TC) is computed between and within groups. $(6, 1)$ yields an average between-group TC of 0.159, whereas $(6, 2)$ achieves a lower between-group TC of 0.110 and a within-group TC of 0.201, indicating better separation between groups while retaining within-group structure.

Regarding connectivity, Fig. 4(d) illustrates that without disentanglement, even low-rank connectivity remains hard to interpret. Under $(6, 1)$ (Fig. 4(e)), the rank-1 sub-connectivity linked to the oscillatory latent (group 6) reveals strong bilateral connections between S1-bf regions and an influence from right M2 to right S1-bf. But with $(6, 2)$ (Fig. 4(f)), the rank-2 connectivity associated with the oscillatory latent (group 1) offers more interesting interpretations. For example, it reveals a strong excitatory connection from S2 (secondary somatosensory) to M2-m. There are also connections from right S2 and left M2 to both sides of RSC (retrosplenial cortex), potentially indicating the formation of episodic memory of receiving the airpuff. Videos of group dynamics are provided in the supplementary materials.

5 CONCLUSION

In this work, we develop a disentangled low-rank RNN (DisRNN) framework that captures both the latent dynamics and connectivity structure of neural systems while relaxing the assumption of full independence among latent components. By combining the expressiveness of low-rank RNN (lrRNN) with a flexible variational inference approach, we enable group-wise disentanglement of latent dynamics via a partial correlation penalty. Compared to lrRNN, fully disentangled models, and SVD-based methods, DisRNN produces latent representations that better match known task variables and reveal an interpretable sub-connectivity structure. Our results across synthetic and real datasets suggest that partial disentanglement is powerful for uncovering modular computation in brain activity.

Despite its advantages, DisRNN also has several limitations. First, the model requires pre-specifying both the number and the size of latent groups. Although our framework can automatically identify and suppress dummy or inactive components within each group, determining the appropriate group structure remains a hyperparameter selection challenge that may depend on domain knowledge or model selection criteria. Second, connectivity interpretability is currently limited to the case where both the encoder and decoder are linear. While this constraint enables direct estimation of functional connectivity between latent variables and observed neural activity, it represents a trade-off between interpretability and representational power. Extending the framework to allow for nonlinear encoders or decoders could improve modeling flexibility, but would make connectivity estimation less straightforward. Finally, the current formulation does not incorporate external task or stimulus inputs, although doing so would be conceptually straightforward. Integrating such inputs offers an exciting direction for future work, potentially enabling the dissection of task-dependent latent dynamics and input-driven connectivity changes.

REFERENCES

- Bagrat Amirikian and Apostolos P Georgopoulos. Modular organization of directionally tuned cells in the motor cortex: is there a short-range order? *Proceedings of the National Academy of Sciences*, 100(21):12474–12479, 2003.
- Mikio C Aoi, Valerio Mante, and Jonathan W Pillow. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature neuroscience*, 23(11):1410–1420, 2020.
- Manuel Beiran, Alexis Dubreuil, Adrian Valente, Francesca Mastrogiuseppe, and Srdjan Ostojic. Shaping dynamics with multiple populations in low-rank recurrent networks. *Neural computation*, 33(6):1572–1615, 2021.
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.
- Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.
- David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.
- John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.
- Alexis Dubreuil, Adrian Valente, Manuel Beiran, Francesca Mastrogiuseppe, and Srdjan Ostojic. The role of population structure in computations through neural dynamics. *Nature neuroscience*, 25(6):783–794, 2022.
- Apostolos P Georgopoulos, John F Kalaska, Roberto Caminiti, and Joe T Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, 2(11):1527–1537, 1982.
- Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
- Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.
- Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Gerick M Lee, Najib J Majaj, CL Rodríguez Deliz, Lynne Kiorpes, and J Anthony Movshon. Emergence of a contrast-invariant representation of naturalistic texture in macaque visual cortex. *bioRxiv*, 2025.
- Chengrui Li, Soon Ho Kim, Chris Rodgers, Hannah Choi, and Anqi Wu. One-hot generalized linear model for switching brain state discovery. *arXiv preprint arXiv:2310.15263*, 2023.
- Chengrui Li, Yunmiao Wang, Yule Wang, Weihai Li, Dieter Jaeger, and Anqi Wu. A revisit of total correlation in disentangled variational auto-encoder with partial disentanglement. *arXiv preprint arXiv:2502.02279*, 2025.
- Scott Linderman, Ryan P Adams, and Jonathan W Pillow. Bayesian latent structure discovery from multi-neuron recordings. *Advances in neural information processing systems*, 29, 2016.

- Xiaoyu Lu, Yunmiao Wang, Zhuohe Liu, Yueyang Gou, Dieter Jaeger, and François St-Pierre. Wide-field imaging of rapid pan-cortical voltage dynamics with an indicator evolved for one-photon microscopy. *Nature Communications*, 14(1):6423, 2023.
- Xuan Ma, Fabio Rizzoglio, Kevin L Bodkin, Eric Perreault, Lee E Miller, and Ann Kennedy. Using adversarial networks to extend brain computer interface decoding accuracy over time. *elife*, 12: e84296, 2023.
- Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.
- Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.
- Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- Matthijs Pals, A Erdem Sağtekin, Felix Pei, Manuel Gloeckler, and Jakob H Macke. Inferring stochastic low-rank recurrent neural networks from neural data. *arXiv preprint arXiv:2406.16749*, 2024.
- Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.
- Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.
- Yasser Roudi, Benjamin Dunn, and John Hertz. Multi-neuronal activity and functional connectivity in cell assemblies. *Current opinion in neurobiology*, 32:38–44, 2015.
- Friedrich Schuessler, Alexis Dubreuil, Francesca Mastrogiuseppe, Srdjan Ostojic, and Omri Barak. Dynamics of random recurrent networks with correlated low-rank structure. *Physical Review Research*, 2(1):013111, 2020a.
- Friedrich Schuessler, Francesca Mastrogiuseppe, Alexis Dubreuil, Srdjan Ostojic, and Omri Barak. The interplay between randomness and structure during learning in rnns. *Advances in neural information processing systems*, 33:13352–13362, 2020b.
- Qi She and Anqi Wu. Neural dynamics discovery via gaussian process recurrent neural networks. In *Uncertainty in Artificial Intelligence*, pp. 454–464. PMLR, 2020.
- Adrian Valente, Srdjan Ostojic, and Jonathan W Pillow. Probing the relationship between latent linear dynamical systems and low-rank recurrent neural network models. *Neural computation*, 34(9):1871–1892, 2022a.
- Adrian Valente, Jonathan W Pillow, and Srdjan Ostojic. Extracting computational mechanisms from neural data using low-rank rnns. *Advances in Neural Information Processing Systems*, 35:24072–24086, 2022b.
- Anqi Wu, Nicholas A Roy, Stephen Keeley, and Jonathan W Pillow. Gaussian process based non-linear latent structure discovery in multivariate spike train data. *Advances in neural information processing systems*, 30, 2017.
- Byron M Yu, Afsheen Afshar, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Extracting dynamical structure embedded in neural activity. *Advances in neural information processing systems*, 18, 2005.

A APPENDIX

A.1 BLOCK-DIAGONAL RNN (BDRNN)

Consider a first-order discrete linear system for simplicity,

$$\mathbf{z}^{(t+1)} = \mathbf{F}\mathbf{z}^{(t)}, \quad \mathbf{F} \in \mathbb{R}^{K \times K}. \quad (13)$$

Since not all matrices can be diagonalized, but all matrices have Jordan normal form, we can always write

$$\mathbf{F} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}, \quad (14)$$

where \mathbf{J} is a Jordan normal form, consisting of several Jordan blocks on its block diagonal. If, without loss of generality, there are G numbers of rank- H latent groups evolving independently, then we have

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{J}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{J}_G \end{bmatrix}, \quad (15)$$

where

$$\mathbf{J}_g = \begin{bmatrix} \lambda_g & 1 & & & \\ & \lambda_g & 1 & & \\ & & \lambda_g & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_g \end{bmatrix} \in \mathbb{R}^{H \times H}, \quad (16)$$

where λ_g is the corresponding eigenvalue, so that within each group, latent components are entangled with each other, non-separable. Then,

$$\mathbf{z}^{(t)} = \mathbf{F}^t \mathbf{z}^{(0)} = \mathbf{P}\mathbf{J}^t \mathbf{P}^{-1} \mathbf{z}^{(0)}. \quad (17)$$

By defining $\mathbf{z}' = \mathbf{P}^{-1} \mathbf{z}$ (similar to the equivalence below Eq. (3)), $\mathbf{z}'^{(t)}$ evolves independently by groups, i.e.,

$$\mathbf{z}'_g(t) = \mathbf{J}_g^t \mathbf{z}'_g(0), \quad \mathbf{J}_g^t = \begin{bmatrix} \lambda_g^t & \binom{t}{1} \lambda_g^{t-1} & \binom{t}{2} \lambda_g^{t-2} & \cdots & \cdots & \binom{t}{H-1} \lambda_g^{t-H+1} \\ & \lambda_g^t & \binom{t}{1} \lambda_g^{t-1} & \cdots & \cdots & \binom{t}{H-2} \lambda_g^{t-H+2} \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & \vdots \\ & & & & \lambda_g^t & \binom{t}{1} \lambda_g^{t-1} \\ & & & & & \lambda_g^t \end{bmatrix}. \quad (18)$$

A.2 SUPPLEMENTARY RESULTS

A.2.1 SYNTHETIC

Data generation. Fig. 5 plots the dataset we created for our synthetic experiment.

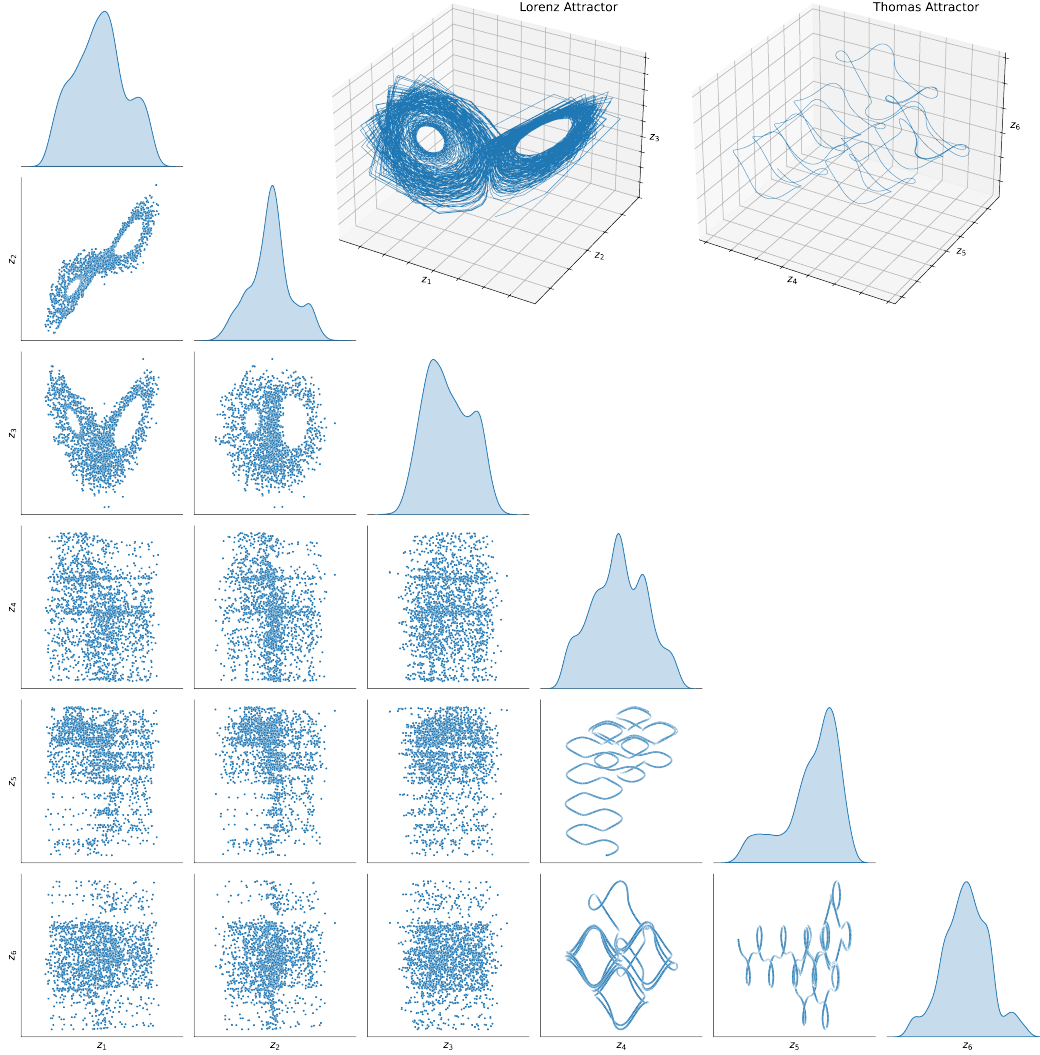


Figure 5: The synthetic latent consists of two independent groups.

Ablation. To analyze the choice of the penalty coefficient β of PC term in Eq. (10), we vary β and plot the cross-validation results in Fig. 6. This supports our choice of $\beta = 20$ in our experiment that has good reconstruction, disentangled latent estimation, and parameter estimation.

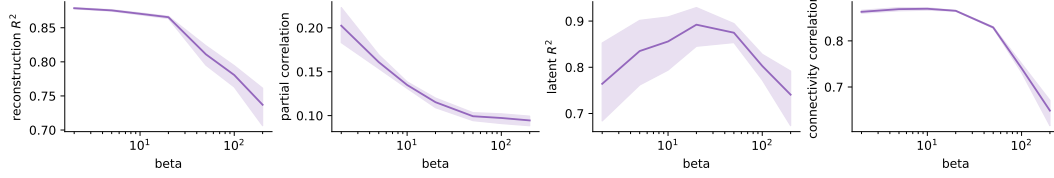


Figure 6: Metrics w.r.t. the PC penalty β in DisRNN on the synthetic dataset.

To verify that DisRNN is flexible in setting the number of groups G and group rank H , we run DisRNN with $(G, H) = (2, 4)$ and $(G, H) = (3, 4)$ against the true $(G, H) = (2, 3)$, and evaluate the latent alignment outcome. The latent R^2 for them are 0.84 ± 0.14 and 0.78 ± 0.09 , respectively. All of them are better than other baseline methods, but worse than the correctly set DisRNN $((G, H) = (2, 3))$. Noticing that for both of them, the internal hidden rank reduces to the effective group rank of 3 automatically. And for $(G, H) = (3, 4)$, there is always a dummy group that aligns badly to both of the true groups. This validates the second benefit of DisRNN that it allows a flexibly set (G, H) for the model to be trained, when we do not know the true (G, H) .

Alignment. Fig. 7 demonstrates the alignment procedure.

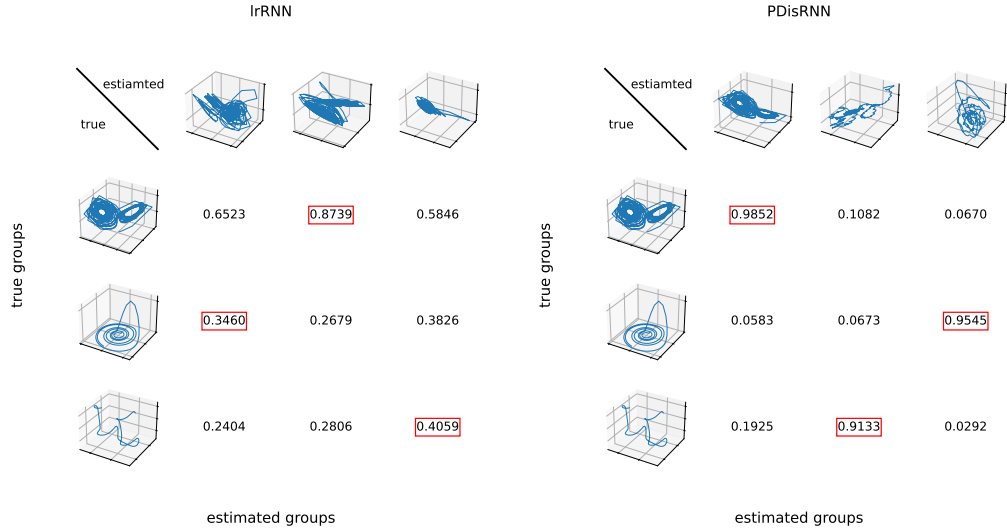


Figure 7: The latent dynamics alignment process. We try to align the estimated latent groups with the ground truth. The best match is marked by the red squared linear least-squares R^2 score.

Latent dynamics. Fig. 8 shows the estimated latent dynamics vs. the ground truth on all six latent components.

All possible permutations. For each method without the assumption of group-wise structure, we try all possible permutations of the latent components that partition the latent components into groups. In general, the complexity is

$$\frac{\prod_{g=1}^G \binom{(G+1-g)H}{H}}{G!} = \frac{K!}{G!(H!)^G}. \quad (19)$$

We instantiate this complexity in Fig. 9. Then, for each random seed, we also plot the latent alignment R^2 for all possible permutations, demonstrating that latent components are entangled with

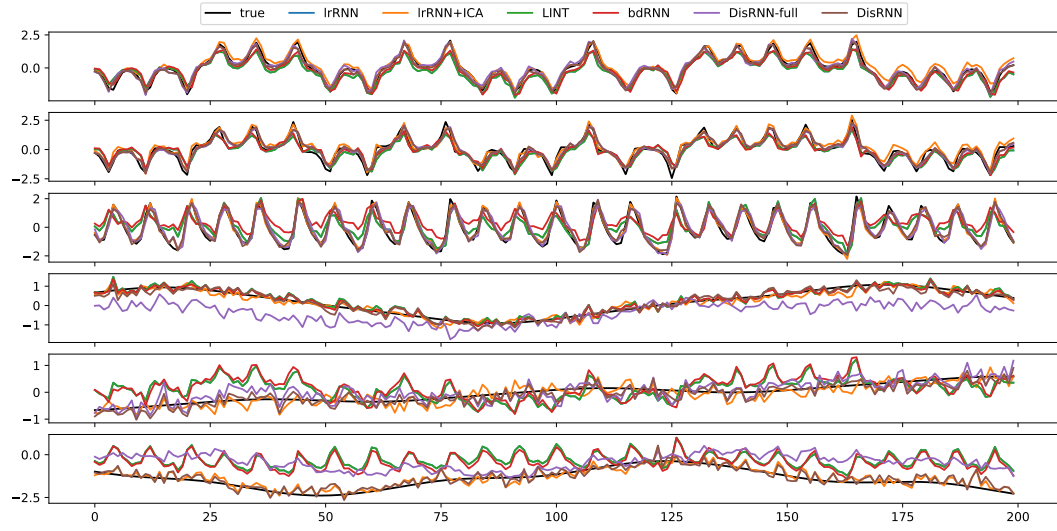


Figure 8: The estimated latent components by different methods on every latent component/dimension.

each other and in random orders. For some methods, the optimal permutation can achieve the same latent R^2 as DisRNN. However, this does not invalidate our DisRNN method, because DisRNN can explicitly disentangle and assign components into groups. When the group specification is only moderately large, exhausting all permutations becomes intractable, which forms a theoretical limitation of baseline methods.

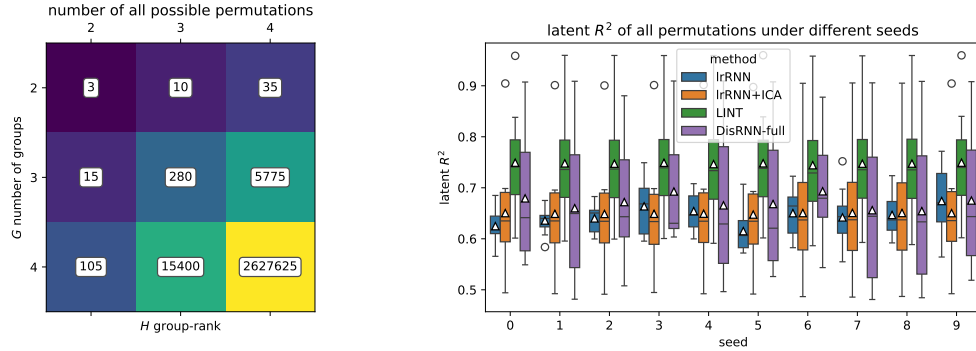


Figure 9: **Left:** Complexity demonstration matrix. **Right:** The latent R^2 of all possible permutations for methods without the assumption of group structure under different random seeds.

A.2.2 MONKEY M1

Connectivity matrices. Fig. 10 shows the learned connectivity matrices from all methods.

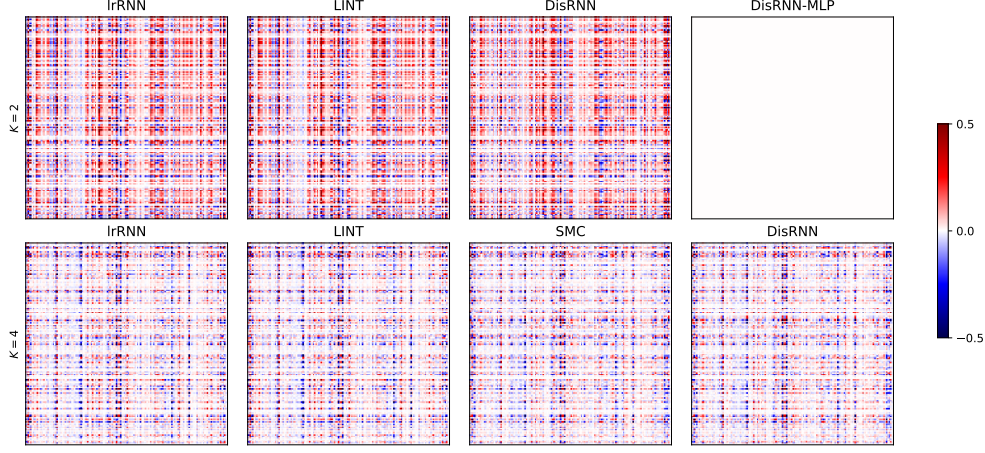


Figure 10: The learned connectivity matrices from all methods in Fig. 3(a), except for $K = 2$ DisRNN-MLP.

Poisson spike count results. To validate that our framework not only works for arbitrary VAE architectures but also works for different observation distributions, we fit different methods to the Poisson spike count data of the same dataset with $K = 2$. The latent alignment R^2 increases from plain IrRNN’s 0.52 ± 0.03 to DisRNN’s 0.76 ± 0.02 to DisRNN-LMP’s 0.81 ± 0.01 . This result is consistent with the result in the main content, confirming that disentanglement is an effective approach for obtaining biologically interpretable latent space structure.

A.3 USE OF LARGE LANGUAGE MODELS

We used large language models (LLMs) solely to aid in writing polish and minor language improvements (e.g., fixing grammar issues, rewriting sentences in a more formal style). They were not used for scientific exploration, conceptualization, experimental design, analysis, or conclusions.