Latent variable sequence identification for cognitive models with recurrent neural networks

Ti-Fen Pan (tfpan@berkeley.edu)
Helen Wills Neuroscience Institute,
University of California, Berkeley, 2121 Berkeley Way Berkeley, CA 94704 US

Bill Thompson (wdt@berkeley.edu)
Department of Psychology,
University of California, Berkeley, 2121 Berkeley Way Berkeley, CA 94704 US

Anne Collins (annecollins@berkeley.edu)
Department of Psychology,
Helen Wills Neuroscience Institute,
University of California, Berkeley, 2121 Berkeley Way Berkeley, CA 94704 US

Abstract

Extracting time-varying latent variables from computational cognitive models is a key step in model-based neural analysis, which aims to understand the neural correlates of cognitive processes. To derive latent variables, researchers typically fit computational models with likelihood-dependent techniques such as Maximum Likelihood Estimation. However, many relevant cognitive models have intractable likelihood, limiting our ability to use these models for analyses. Here, we present an approach to learn a direct mapping between time-series experimental data and the targeted latent variable space using recurrent neural networks trained to recover latent variable sequences in synthetic data. The results show that our approach reaches high accuracy in inferring latent variables in both tractable and intractable models. Furthermore, the approach is generalizable across different computational models and can identify both continuous and discrete latent spaces. Overall, our work suggests that using neural networks trained on synthetic data to analyze experimental data is a promising way to access a broader class of cognitive models in model-based neural analyses.

Keywords: artificial neural network; latent variable identification; intractable likelihood; model-based neural analysis

Introduction

Computational cognitive models are widely used to relate derived time-varying latent variables to neural data (Cohen et al., 2017). Model variables provide a quantitative, trial-by-trial predictor of neural activity, allowing researchers to explore the underlying computational processes and individual differences (Katahira & Toyama, 2021). For instance, Reward Prediction Errors (RPEs) extracted from a reinforcement learning model have been found to correlate with BOLD activity in the ventral striatum, as well as phasic activity of dopamine neurons (O'Doherty, Hampton, & Kim, 2007).

Extracting time-varying latent variables from experimental data typically necessitates two steps: model fitting to identify the best-fitting parameters; and running the computational model with the best-fitting parameters to obtain the latent variable sequences. In the first model fitting step, likelihood-dependent methods such as Maximum Likelihood Estimation (MLE) or Maximum a Posteriori (MAP) are commonly used. However, these methods fall short for models with intractable likelihood (Rmus, Pan, Xia, & Collins, 2023). When dealing with models with intractable likelihood, researchers usually have to develop complex and customized statistical approaches that are not generalizable to broader computational models (Ashwood et al., 2022).

Most computational models with intractable likelihoods can be simulated. Recently, a variety of simulation-based methods (Busetto Alberto et al., 2013) have taken advantage of this attribute to overcome the hurdle in likelihood computation. Specifically, methods leveraging artificial neural networks (ANN) to estimate posterior probability (Radev, Mertens, Voss, Ardizzone, & Köthe, 2020) or generative parameters (Lenzi, Bessac, Rudi, & Stein, 2023) have successfully enabled parameter recovery across a wide range of computational models. However, these simulation-based methods are primarily concerned with parameter recovery or model identification. Time-varying latent variables extraction in likelihood intractable models is still under-explored (Schumacher, Bürkner, Voss, Köthe, & Radev, 2023). Here, we propose an ANN-based method for learning a direct mapping between a sequence of observable variables and the targeted latent variable space, using simulations to train our ANN. We show that our resulting tool is useful in a variety of computational models and can identify both discrete and continuous latent model variables.

Methods

The proposed method consists of two phases: training and inference (Fig. 1). During the training phase, we first create a synthetic dataset using the targeted computational model and parameter priors. This synthetic dataset includes the model behavior (similar in structure to what participants' observable behavior would be), and model latent variables (which are unobservable in participants' data). An ANN is trained using model-simulated observable behavior as input and a series of model-derived latent variables as output. During the inference phase, the trained ANN is supplied the experimental data as input, resulting in a sequence of inferred unobservable latent variables.

Note that our technique does not require large data sets and can be applied at the individual level with standard number of trials. This is because the neural network training is done not on real data-sets, but on synthetic data-sets, of which we can have an arbitrarily large number as needed for training.

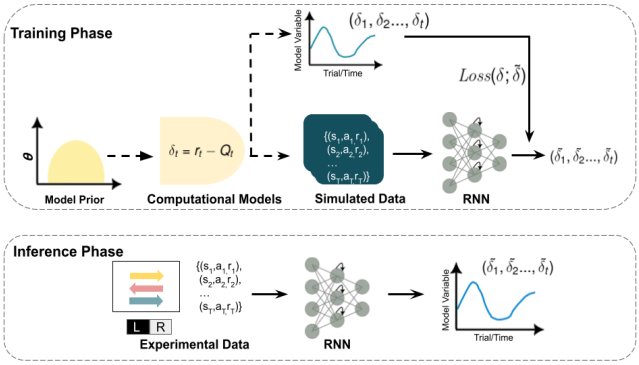

Figure 1: Overview of our proposed ANN-based method

Network Architecture

The basic building block of our neural networks is a recurrent neural network (RNN) (Funahashi & Nakamura, 1993) followed by Multilayer Perceptrons (MLPs). The RNN can be

seen as a time-series representation learner that generates an embedding for each time point. Our RNN is based on 193 bidirectional Gated Recurrent Units (GRU) (Cho et al., 2014). Bidirectionality enables the network to learn embeddings from both past and future history (Schuster & Paliwal, 1997). Following the RNN, MLPs learn the direct mapping between time-series embeddings and the targeted variable space. MLPs consisted of two hidden layers with 95 and 48 units, respectively. We use the rectified linear unit (ReLU) activation function in all MLP layers.

The proposed approach's adaptability allows us to identify both discrete and continuous latent variables with minor changes to the final activation and loss functions. For discrete variables, we used a softmax activation function in the output layer, with a cross-entropy loss as the objective function. For continuous variables, we used linear activation with mean-squared error loss.

Results

All results in this study were evaluated against previously unseen testing data. Overall, our approach performed well in both likelihood tractable and intractable models.

Likelihood Tractable Models We first simulated 5000 participants with 500 trials per participant using a 4-parameter reinforcement learning model (4-P RL) on a two-armed bandit with probabilistic reversal task (Zou, Muñoz Lopez, Johnson, & Collins, 2022). This synthetic data was used to train an ANN model to predict chosen Q-value sequences. The trained model and MLE were evaluated against an additional 1000 simulated participants. We obtained RPEs by subtracting Q-values from the received rewards at each time point (Fig. 2A). We calculated the Root Mean Squared Error (RMSE) between true and estimated Q-values in all trials and averaged it across participants. We found the ANN reaches similar average RMSE (0.041) to MLE (0.042) in likelihood tractable models (Fig. 2B).

Likelihood Intractable Models We tested our approach using two likelihood intractable models: a model based on the generalized linear model and hidden Markov models (GLM-HMM) in a perceptual decision making task (Ashwood et al., 2022), which provides a benchmark method in a likelihood intractable context, and a Hierarchical reinforcement learning model (HRL) in a novel dynamic decision making task (Rmus et al., 2023), in which no benchmark method exists. Both models' training data included 3000 simulated participants and 720 trials per participant. All evaluations across methods were conducted on an additional 1000 simulated participants.

The GLM-HMM model's synthetic dataset was based on a mouse binary perceptual decision-making task. The model was trained to predict three discrete latent states: engaged, biased-left, and biased-right. To assess performance, we used a balanced accuracy score that is the macro average of recall per state label and avoids score inflation in an imbalanced dataset. We compared our approach to the approximate expectation–maximization (EM) algorithm used in Ashwood et al. (2022). The results showed the our method increased identification accuracy by 4.8% in average(Fig. 2C).

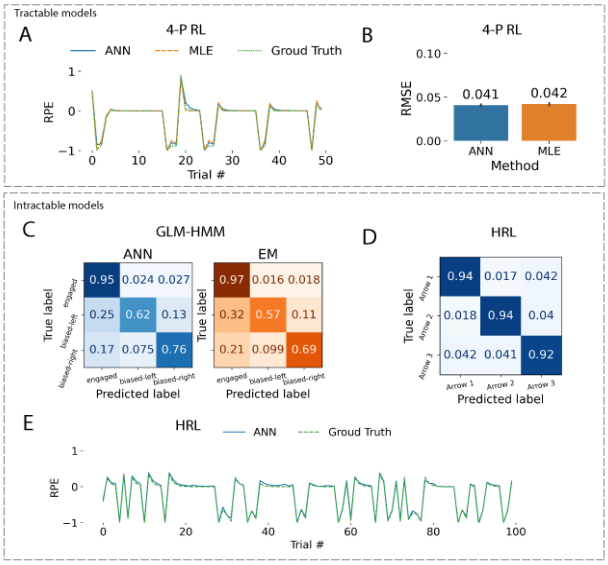

Figure 2: Latent variable identification results across models. A) Derived RPE from one simulated participant in 4P-RL B) Average RMSE of Q-values in 4P-RL C) Latent state identification in GLM-HMM D) Latent cue/arrow identification in HRL E) Derived RPE from one simulated participant in HRL

Finally, we tested our method with the HRL model on a hierarchical reinforcement learning task. In this task, participants are shown three arrows, each pointing either right or left. Participants learned which arrow to follow and press right or left for rewards. The HRL model tracks the Q-values of arrows and decides which one to follow based on Q-values. Because the arrow chosen by the participant is non-observable (only the right or left choice is), this model likelihood is intractable (see (Rmus et al., 2023) for further task details). We trained the model to identify the arrow that simulated participants covertly follow and its corresponding Q-values. Our model reaches 93% accuracy in latent discrete cue identification (arrow selection; Fig. 2D) and the RMSE across agents is 0.119 in Q-values identification (Fig. 2E).

Conclusion

In this work, we show that our method performs well even when the likelihood is intractable. Our method is adaptable to both discrete and continuous latent variable identification, as well as generalizable across different computational models. To evaluate our method further, our ongoing work includes real data fitting and robustness tests (e.g. misspecified parameter priors). In conclusion, breaking down the barrier of intractable likelihood and recovering the latent dynamics of computational models will provide researchers with new insights into previously inaccessible relations between behavioral and neural data.

Acknowledgments

This work was supported by NIH R21MH132974.

References

Ashwood, Z. C., Roy, N. A., Stone, I. R., Laboratory, I. B., Urai, A. E., Churchland, A. K., ... Pillow, J. W. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, *25*(2), 201–212.

Busetto Alberto, G., Numminen, E., Corander, J., Foll, M., Dessimoz, C., et al. (2013). Approximate bayesian computation. *PLoS computational biology*, *9*(1).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., ... others (2017). Computational approaches to fmri analysis. *Nature neuroscience*, *20*(3), 304–313.

Dezfouli, A., Ashtiani, H., Ghattas, O., Nock, R., Dayan, P., & Ong, C. S. (2019). Disentangled behavioural representations. *Advances in neural information processing systems*, *32*.

Funahashi, K.-i., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, *6*(6), 801–806.

Katahira, K., & Toyama, A. (2021). Revisiting the importance of model fitting for model-based fmri: It does matter in computational psychiatry. *PLoS computational biology*, *17*(2), e1008738.

Lenzi, A., Bessac, J., Rudi, J., & Stein, M. L. (2023). Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis*, *185*, 107762.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of sciences*, *1104*(1), 35–53.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, *33*(4), 1452–1466.

Rmus, M., Pan, T.-F., Xia, L., & Collins, A. G. (2023). Artificial neural networks for model identification and parameter estimation in computational cognitive models. *Biorxiv*.

Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for bayesian estimation of dynamic cognitive models. *Scientific Reports*, *13*(1), 13778.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

van Opheusden, B., Acerbi, L., & Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLoS computational biology*, *16*(12), e1008483.

Zou, A. R., Muñoz Lopez, D. E., Johnson, S. L., & Collins, A. G. (2022). Impulsivity relates to multi-trial choice strategy in probabilistic reversal learning. *Frontiers in Psychiatry*, *13*, 800290.