Disentangled behavioral representations

Amir Dezfouli[1*], Hassan Ashtiani[2], Omar Ghattas[1,3], Richard Nock[1,4,5], Peter Dayan[6], Cheng Soon Ong[1,4]

[1]Data61, CSIRO, Australia, [2]McMaster University, [3]University of Chicago, [4]Australian National University, [5]University of Sydney, [6]Max Planck Institute for Biological Cybernetics

*akdezfuli@gmail.com

Introduction

- **Individual characteristics in human decision-making** are often quantified by fitting a parametric cognitive model to subjects' behavior and then studying differences between them in the associated parameter space. However, these models often fit behavior more poorl than recurrent neural networks (RNNs), which are more flexible and make fewer assumptions about the underlying decision-making processes.
- **The parameter and latent activity spaces of RNNs** are generally high-dimensional and uninterpretable, making it hard to use them to study individual differences.
- **We show how to benefit from the flexibility of RNNs while representing individual differences in a low-dimensional and interpretable space.** To achieve this, we propose a novel end-to-end learning framework in which an encoder is trained to map the behavior of subjects into a low-dimensional latent space. These low-dimensional representations are used to generate the parameters of individual RNNs corresponding to the decision-making process of each subject.
- We introduce terms into the loss function that ensure that the latent dimensions are informative and disentangled, i.e. encouraged to have distinct effects on behavior. This allows them to align with separate facets of individual differences.
- We illustrate the performance of our framework on synthetic data as well as a dataset including the behavior of patients with psychiatric disorders.

Training losses

The training loss function has three components:

(i) a reconstruction loss which penalizes discrepancies between the predicted and actual input sequence,

(ii) a group-level disentanglement loss which encourages sequences to spread independently across the dimensions of the input sequence,

(iii) a separation loss which favors dimensions of the latent space that have disentangled effects on the behavior generated by the learning networks.

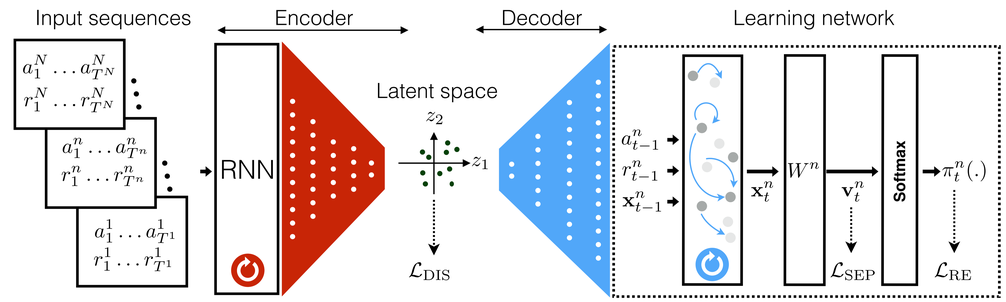
$$\mathcal{L}_{\text{RE}} \equiv -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T^n}\log \pi_t^n(a_t^n; a_{1\dots t-1}^n, r_{1\dots t-1}^n)$$

$$\mathcal{L}_{\text{DIS}} = \lambda_1 \text{MMD}(\hat{q}(\mathbf{z}), p(\mathbf{z})) + \text{KL}(g(\mathbf{z})\|p(\mathbf{z}))$$

$$\mathcal{L}_{\text{SEP}} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial^2}{\partial z_1 \partial z_2}\sum_{t=1}^{T^n} u_t^n\right|$$

$$\mathcal{L} = \mathcal{L}_{\text{RE}} + \lambda_2 \mathcal{L}_{\text{DIS}} + \lambda_3 \mathcal{L}_{\text{SEP}}$$

Model



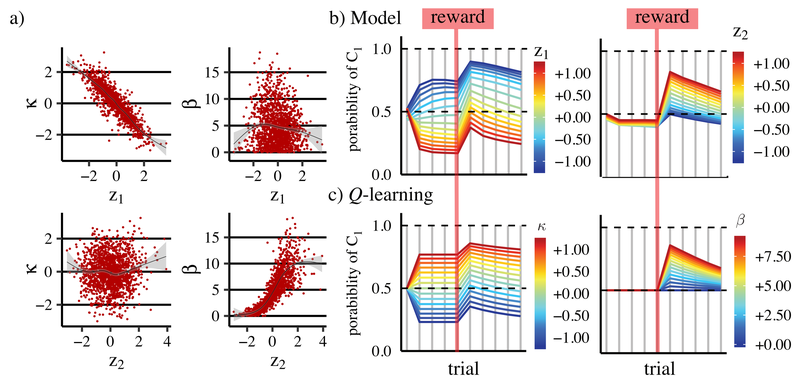
$$\mathbf{z}_{M\times 1}^n \equiv \texttt{enc}(a_{1\dots T}^n, r_{1\dots T}^n; \Theta_{\texttt{enc}}), \quad n = 1\dots N,$$

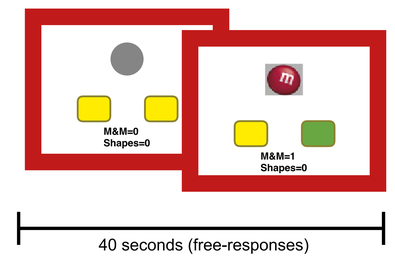
$$\Phi^n \equiv \texttt{dec}(\mathbf{z}^n; \Theta_{\texttt{dec}}).$$

$$\mathbf{x}_t^n \equiv f^n(a_{t-1}^n, r_{t-1}^n, \mathbf{x}_{t-1}^n; \Phi^n)$$

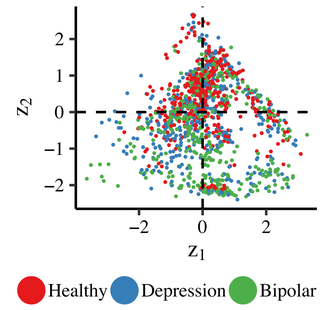
$$\mathbf{v}_t^n \equiv W^n \mathbf{x}_t^n, \quad \pi_t^n(a_t \equiv C_i; a_{1\dots t-1}^n, r_{1\dots t-1}^n) = \frac{e^{\mathbf{v}_t^n[i]}}{\sum_{k=1\dots K} e^{\mathbf{v}_t^n[k]}},$$

Results

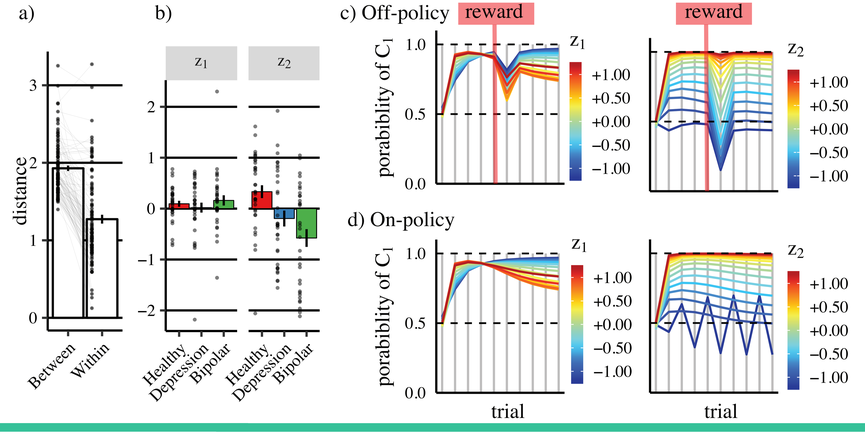




(a)



(b)



Conclusion

Summary:

We proposed a flexible autoencoder-based framework for modelling individual differences in decision-making tasks. The autoencoder maps the sequence of actions and rewards taken and received by a subject to a position in a low-dimensional latent space which is decoded to determine the weights of a `learning network' RNN that characterizes how that subject behaves as a plastic decision-maker. The latent space was disentangled by adding a specific component to the loss function.

References:

Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, Bernard W Balleine. Models that learn how humans learn: the case of decision-making and its disorders.

Acknowledgements:

We are grateful to Bernard W. Balleine for sharing the dataset with us.