Moderational Learning

A Machine-learning Framework to Discover Models and Individual Differences

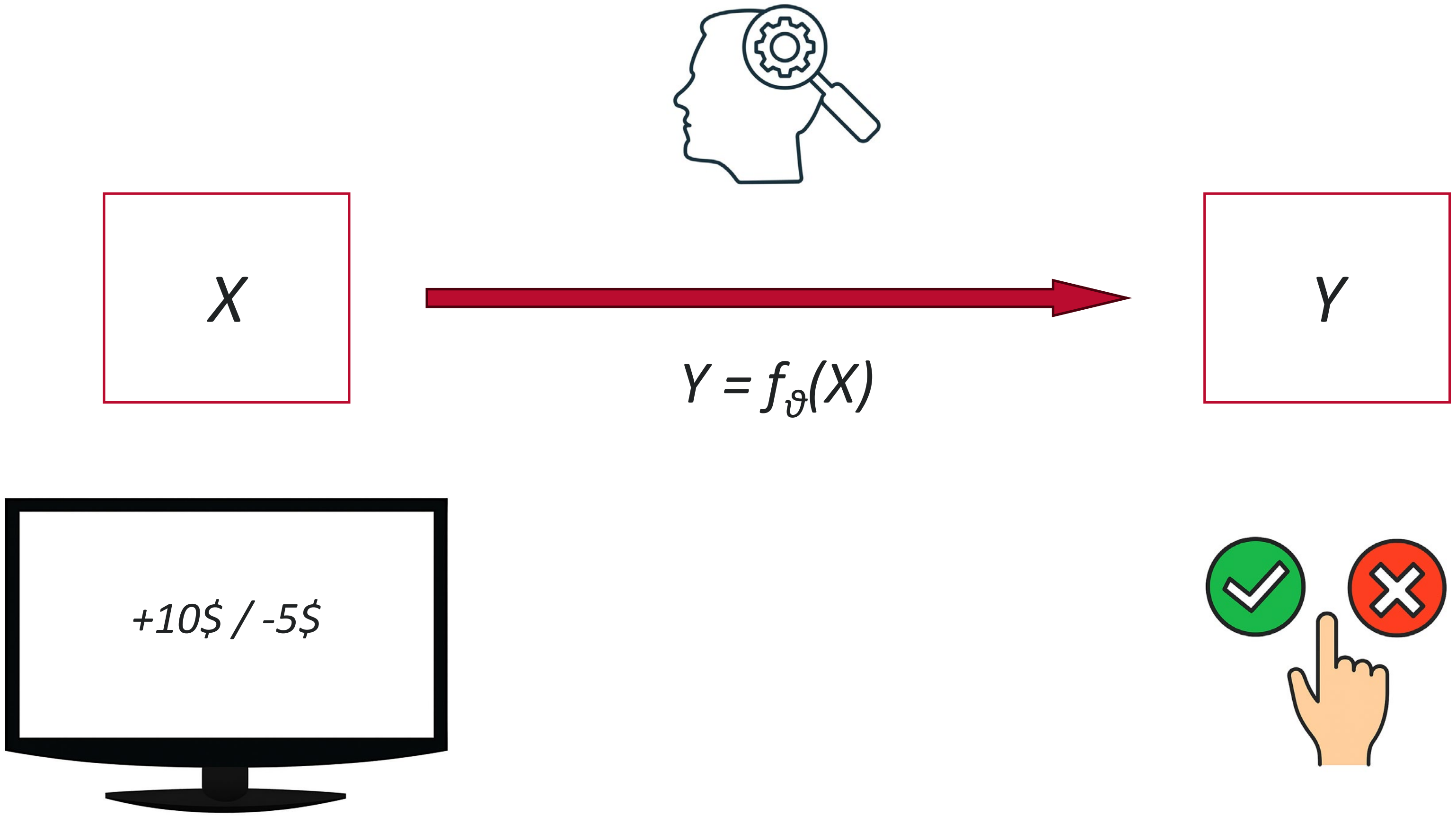
Reporter: Bingsong (Benson) Zhao

Supervisor: Peter Kvam

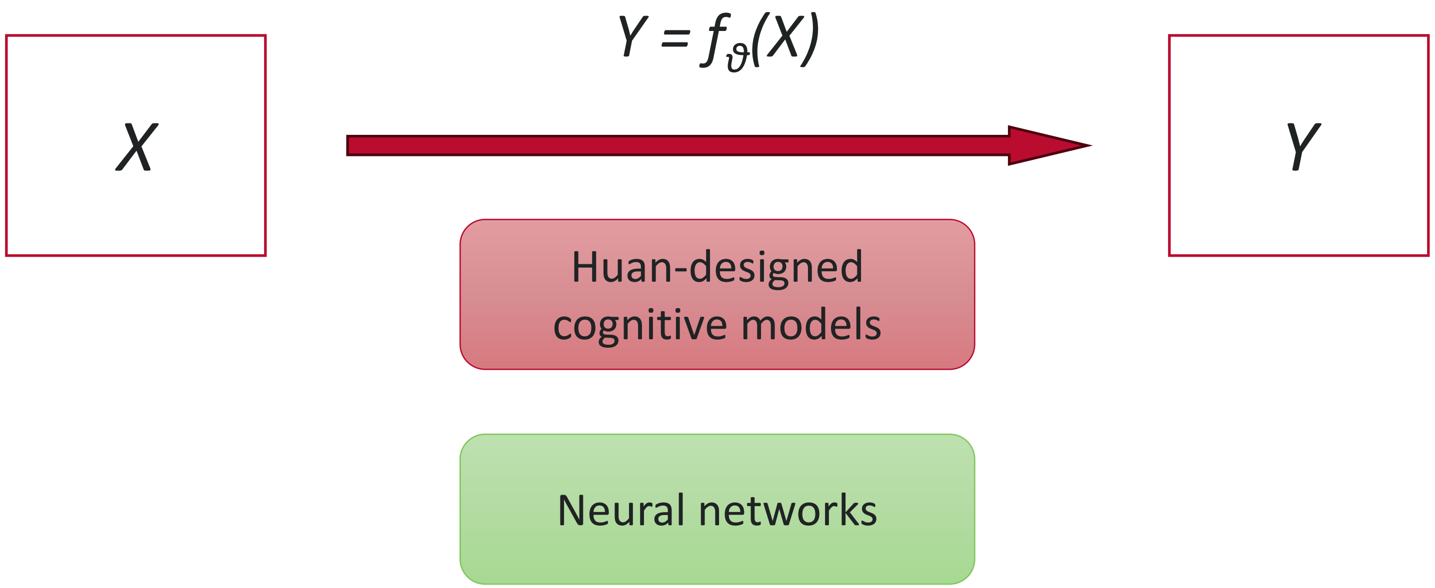
Collaborators: Stefan Radev

Konstantina Sokratous

From cognitive modeling to moderational learning



From cognitive modeling to moderational learning



Using machine learning to model human behavior

Group-level learning

Train a neural network with data merged for all subjects.

Failed to capture individual differences and heterogeneity.

Subject-level learning

Train separate neural networks for each subject.

Failed to capture cross-subject generative processes; Limited data points.

Common limitations

- Limited interpretability.
- Fail to represent individual differences in a common, meaningful space.
- Difficult generating data for new subjects.

Variational auto-encoders (VAE) for latent variable learning

Variational inference

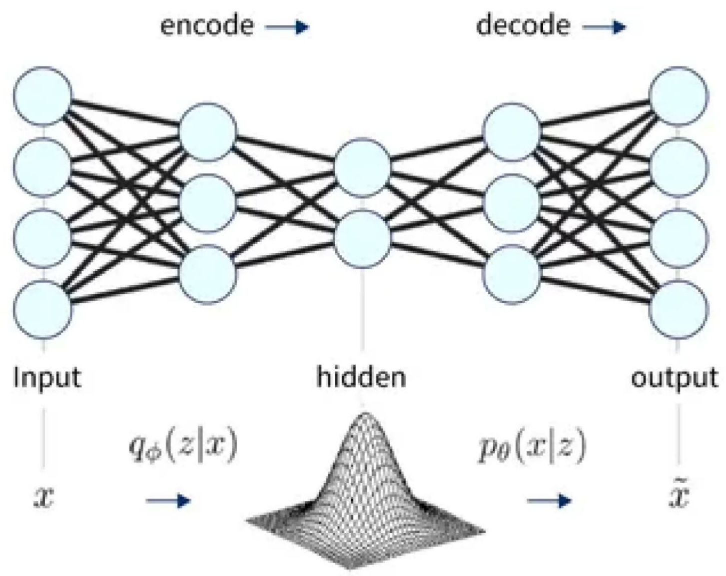
Approximates an intractable latent posterior using a parameterized, tractable distribution, and optimizes its parameters

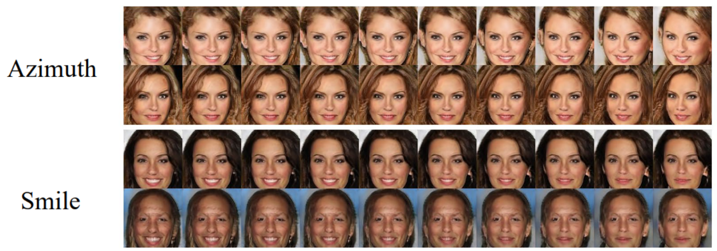
Change-of-variables principle

Simple Gaussian variables can be transformed into arbitrarily complex distributions

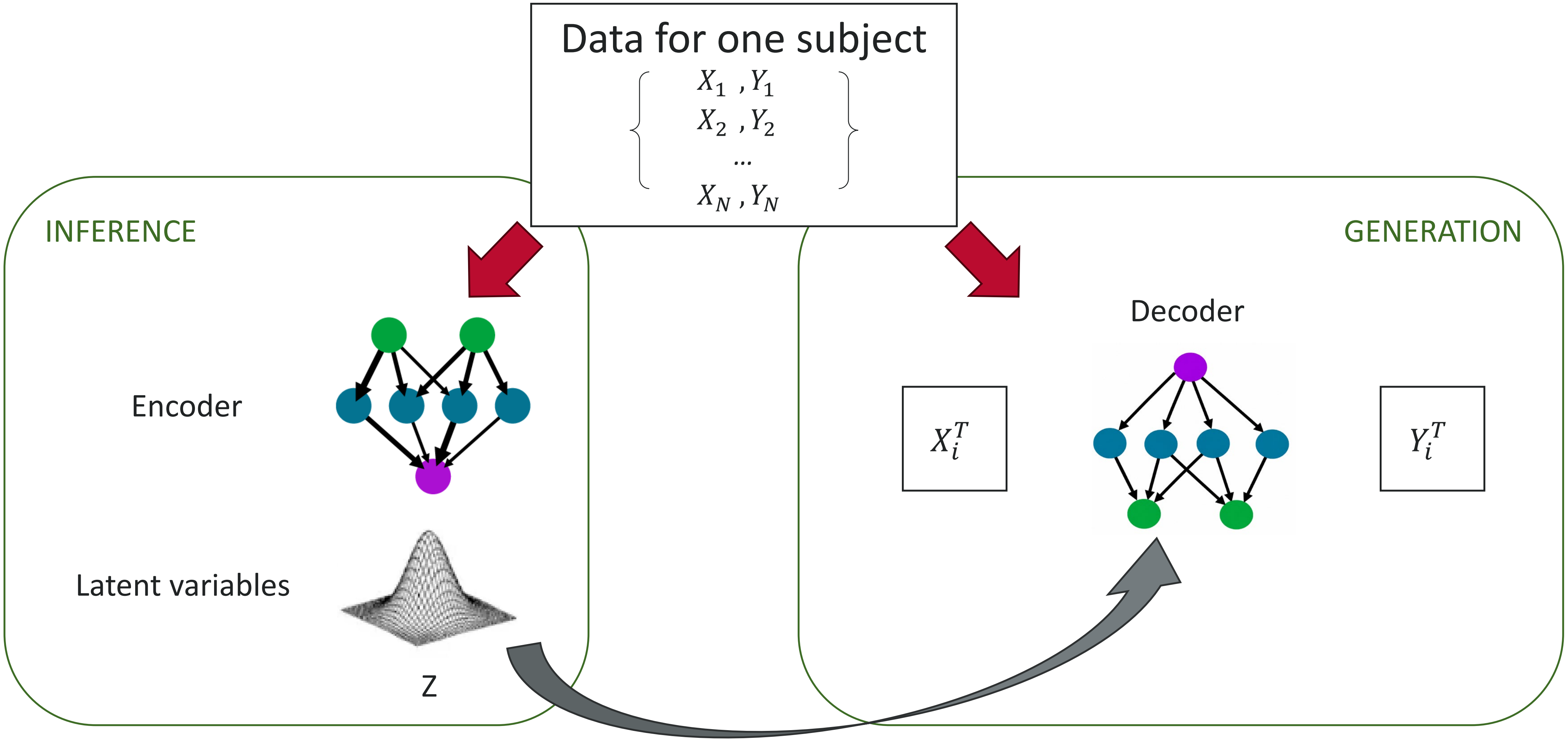
Universal function approximation

Neural networks can approximate any continuous function

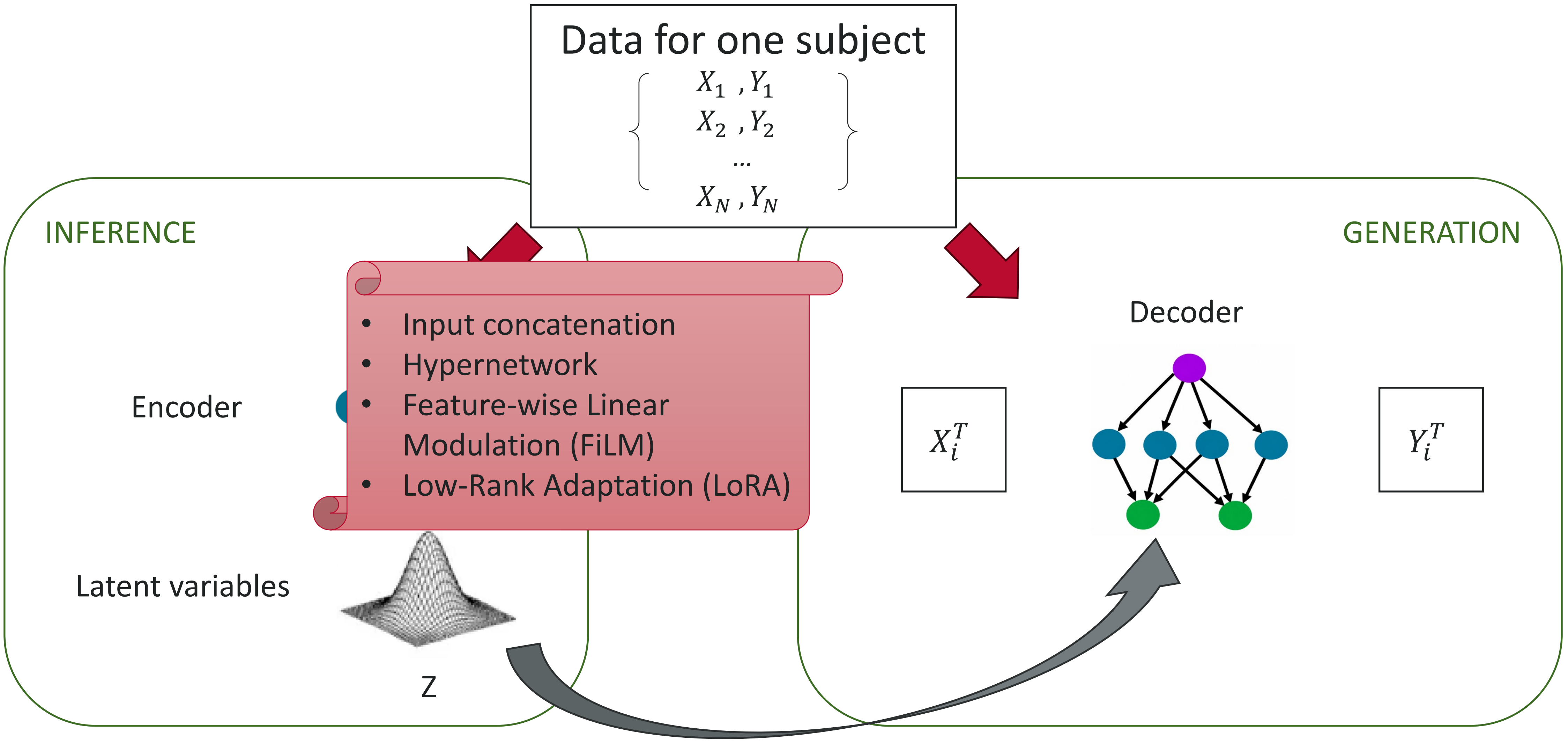




The moderational learning framework



The moderational learning framework



Research map

i.i.d. trial data: point-level reconstruction — Zhao et al., *PsyArXiv* (2026)

i.i.d. trial data: distributional reconstruction — Zhao et al., *In prep*;

Non-i.i.d. trial data — Zhao et al., *In prep*; Dezfouli et al. (2019), *NeurIPS*

Multi-modal/multi-view data — Ongoing

Time-varying and time-invariant latent variables — Ongoing

…

i.i.d. trial data: point-level reconstruction

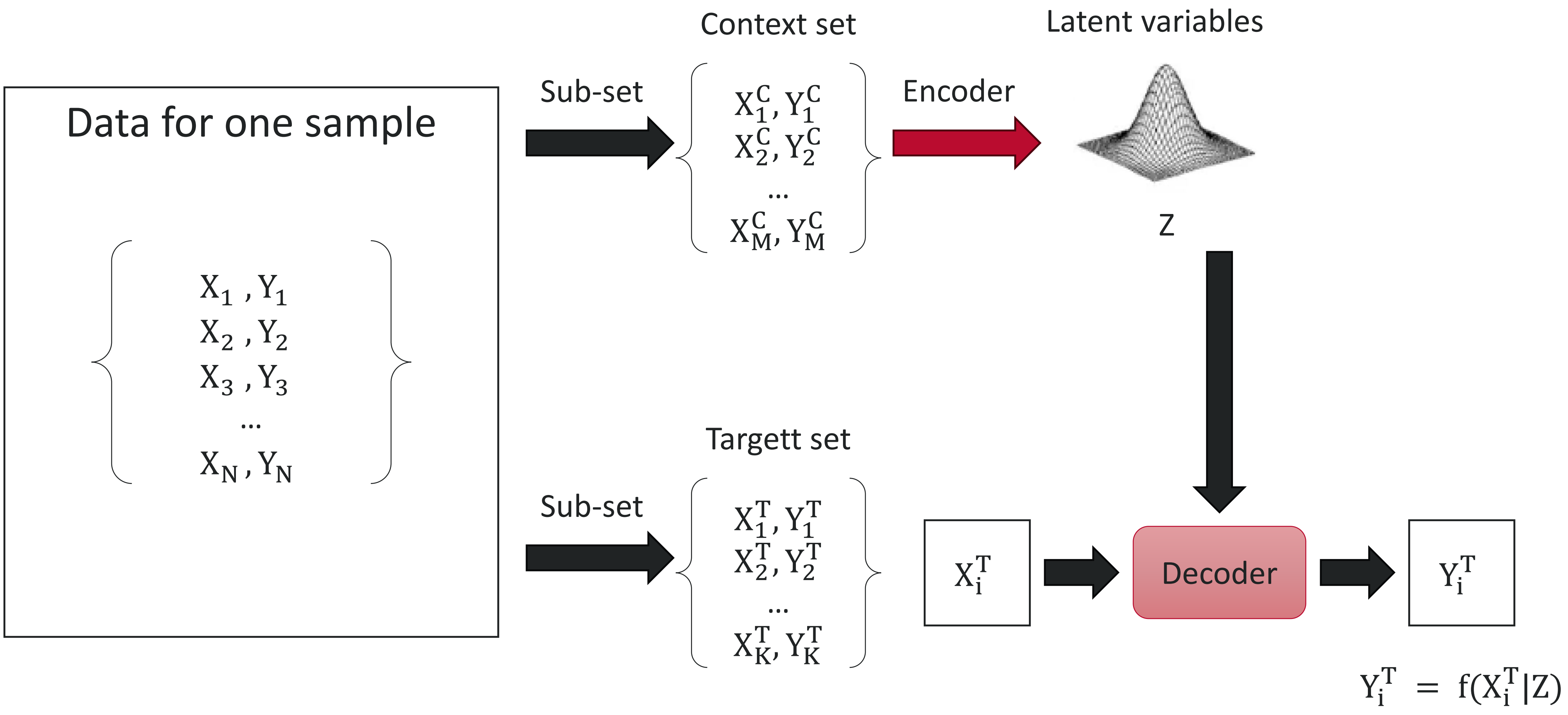
Discovering individual differences and cognitive models for decision-making tasks

Background

psyarxiv/b865r

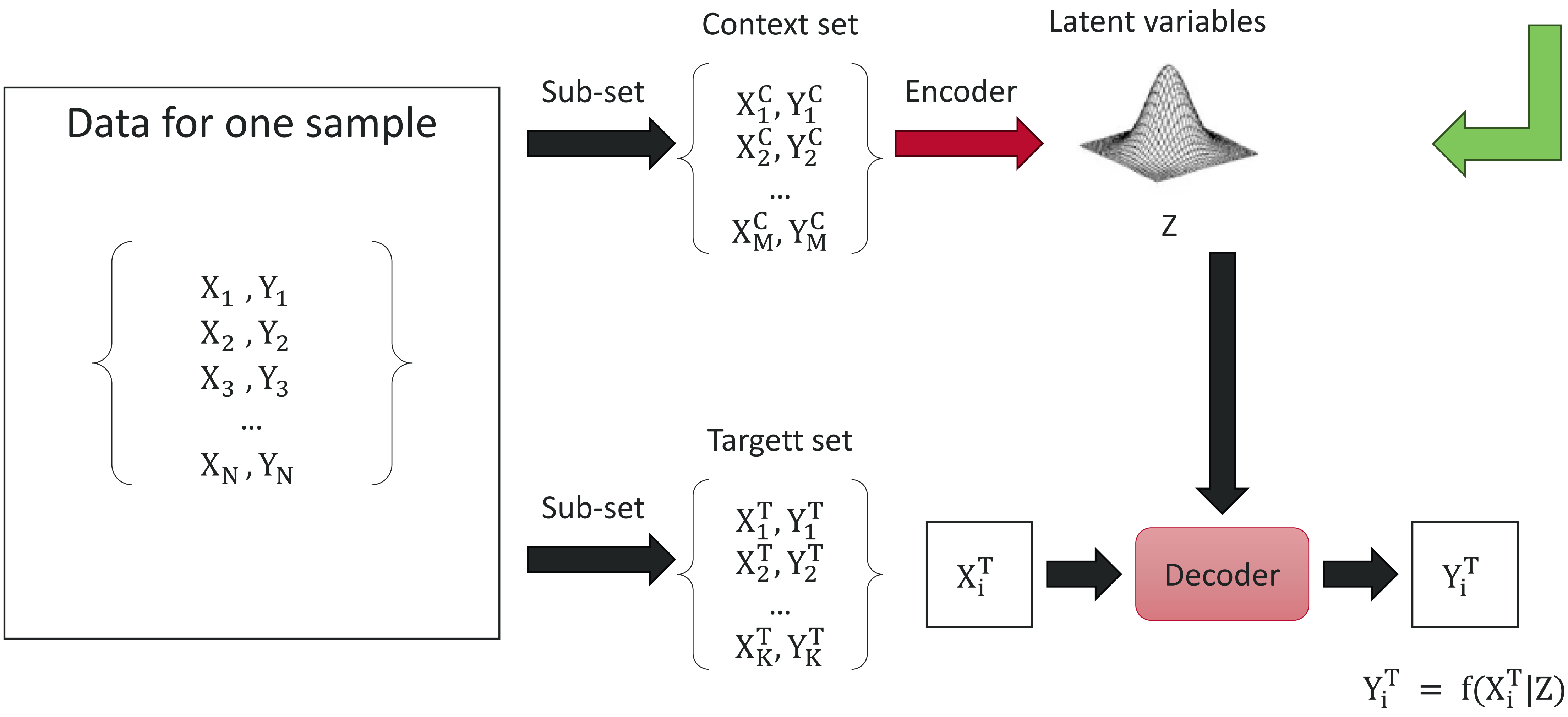
Neural network structure

- Neural Processes (Garnelo et al., 2018)

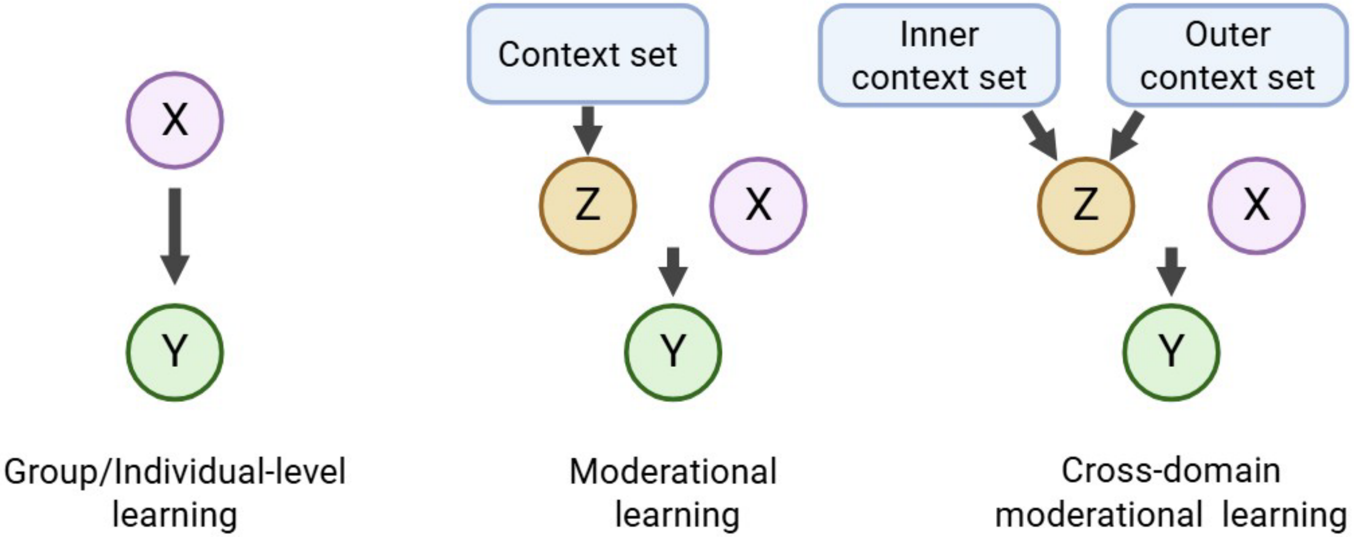


Neural network structure

- Neural Processes (Garnelo et al., 2018)



Neural network structure



psyarxiv/b865r

Overview

i.i.d. Trial Data: Point-Level Reconstruction

Model structure

- Encoder: $Z = q_\phi(data)$
- Decoder: $\hat{Y}_i = f_\theta(X_i, Z)$

Learning Objective with Regularization

$$\text{reconstruction loss} + \beta L_{KL} + \lambda L_{\text{sparse}}$$

Hyperparameter Sweep and Model Selection

For each $(\beta, \lambda_{\text{sparse}})$ combination, we examined: reconstruction loss, legalization loss, number of active dimensions, PCA structure …

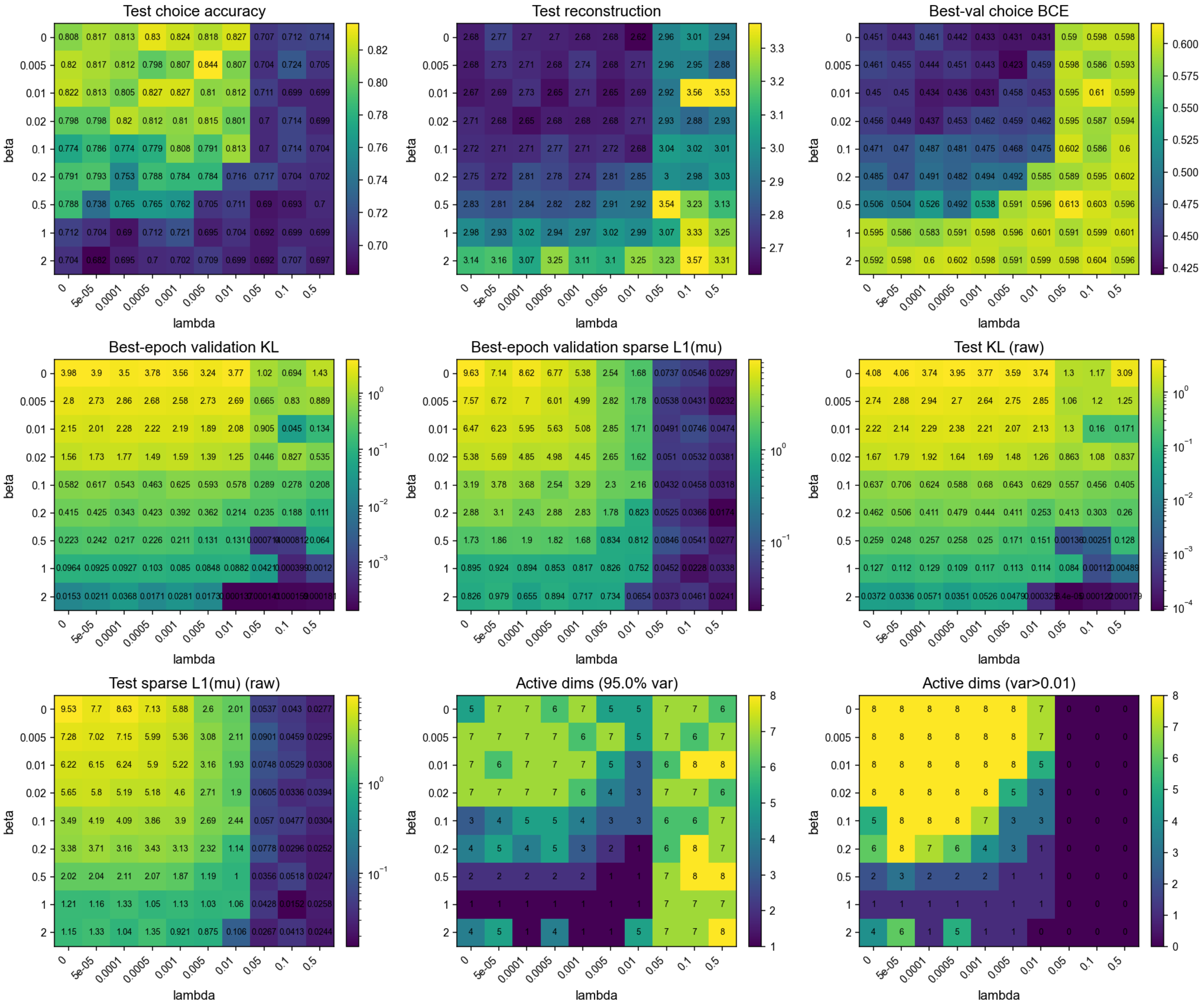
Key results

- Improved held-out predictive performance
- Recovered meaningful within-task and cross-task latent factors
- Revealed new stimulus–response mechanisms
- Translated the decoder's learned X,Z→Y mappings into interpretable equations using symbolic regression

$$\hat{p}^{RiskP} = money + 3.953 - 2.341z_1 + (1.395 - z_4)probability$$

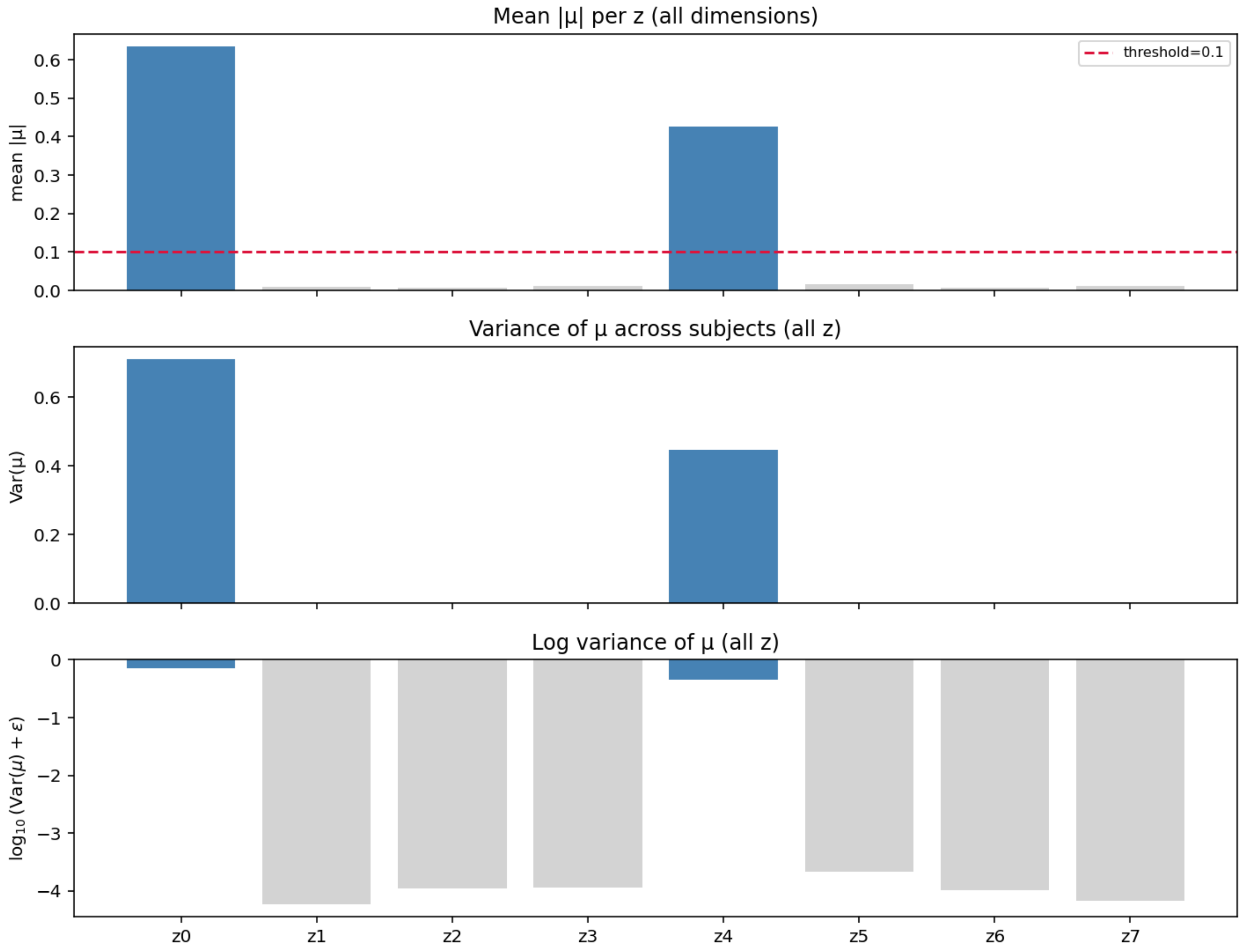
Latent factor discovery

- Sweeping hyper-parameters



Latent factor discovery

- Alive dimensions



Simulation study 1

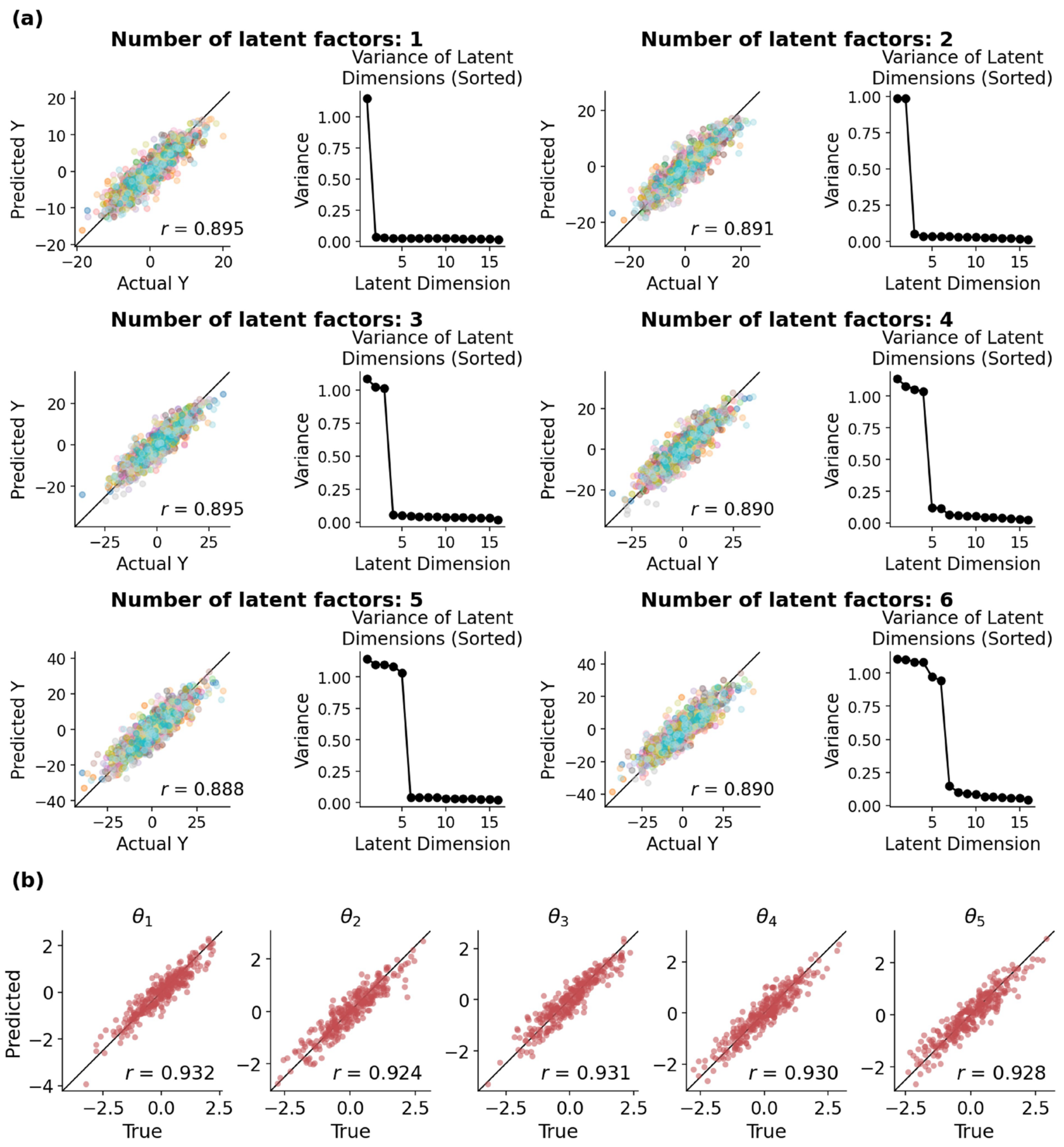
Data generation

- Number of ground-truth latent factors: k∈[1,6]
- 2000 simulated subjects.
- 150~200 trials each subject.

$$Y = \sum_{j=1}^{6} \theta_j X_j + \varepsilon, \varepsilon \sim \text{Normal}(0, \sigma^2)$$

$$X_j \sim \text{Unif}(-1,2), j = 1, \ldots, 6$$

$$\theta_j \sim \text{Normal}(4,1)\ for\ j \leq k, \theta_j = 1\ for\ j > k$$



Simulation study 2

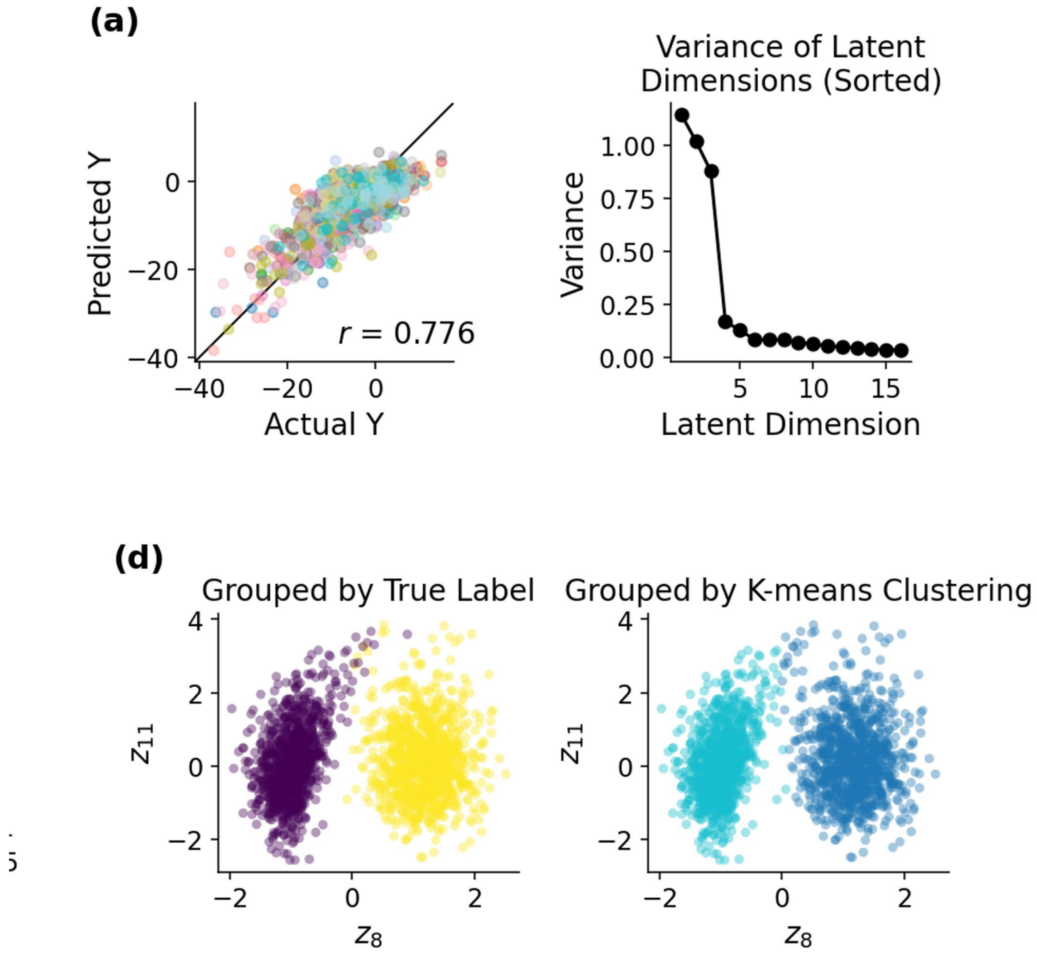
Data generation

$$X = (X_1, X_2, X_3, X_4)$$

$$\theta = (\theta_1, \theta_2, \theta_3)$$

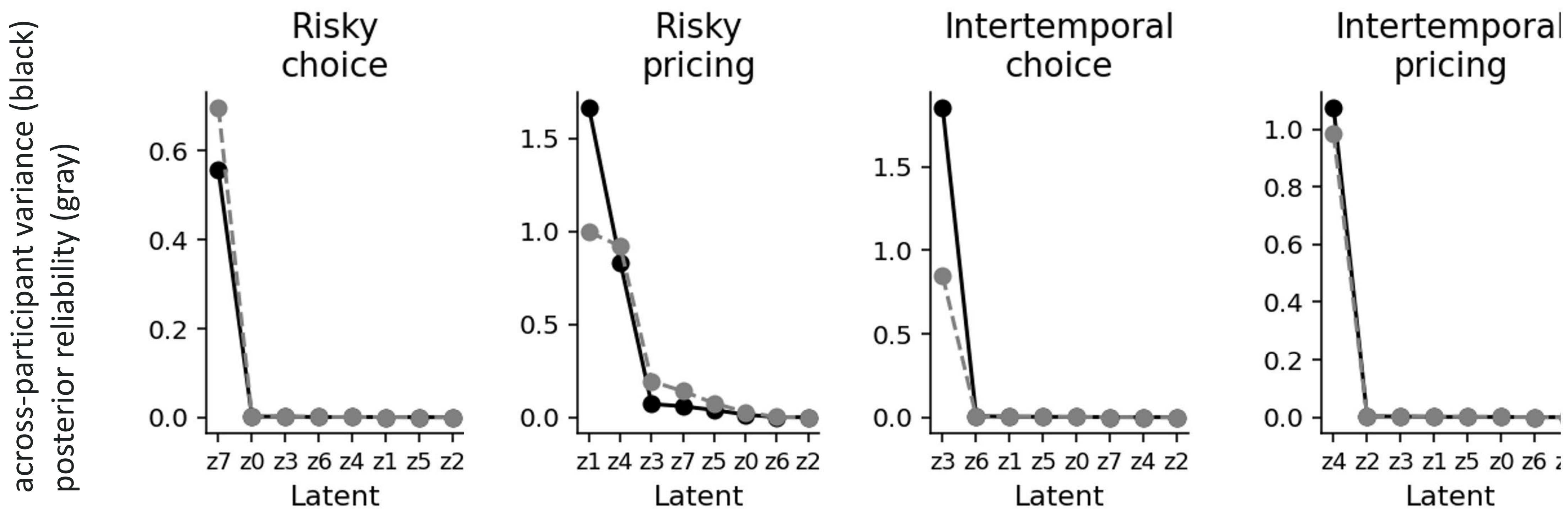
$\theta_1, \theta_2 \sim \text{Normal}(1.5, 0.25);$ $\theta_3 \in \{0, 1\}$

$$Y = \begin{cases} \theta_1(X_1 X_2) - \theta_2(X_3 + X_4)^2, if\ \theta_3 = 0 \\ \theta_1(X_1 X_3) - \theta_2(X_2 + X_4)^2, if\ \theta_3 = 1 \end{cases}$$



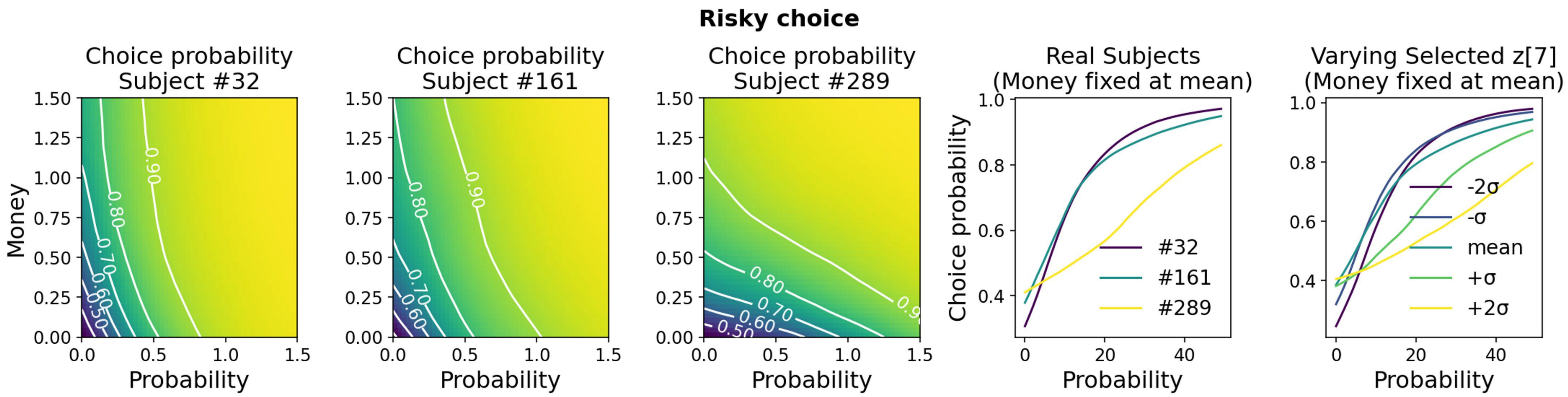
psyarxiv/b865r

Latent factor extraction



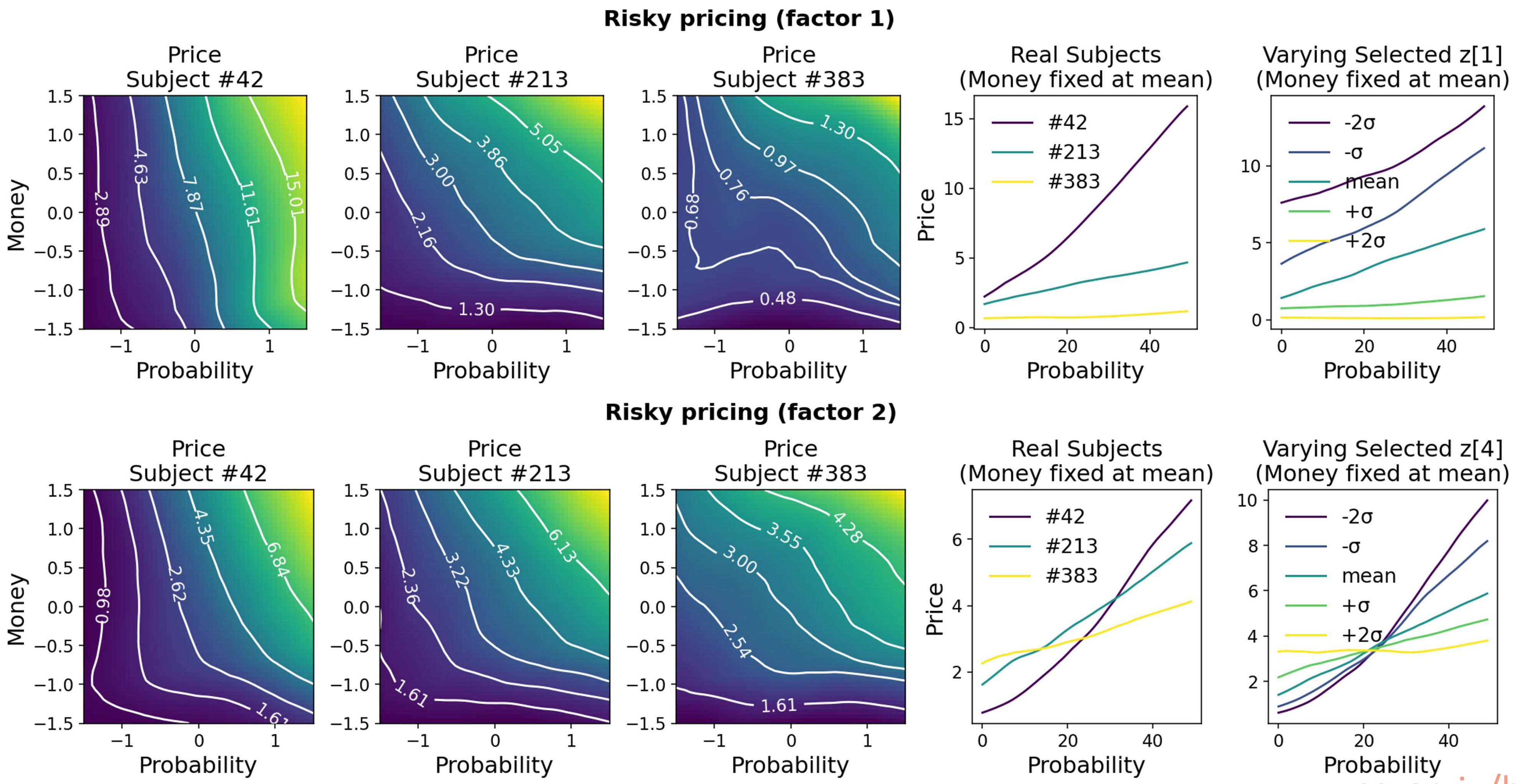
psyarxiv/b865r

Visualization of (X,Z)→Y mappings



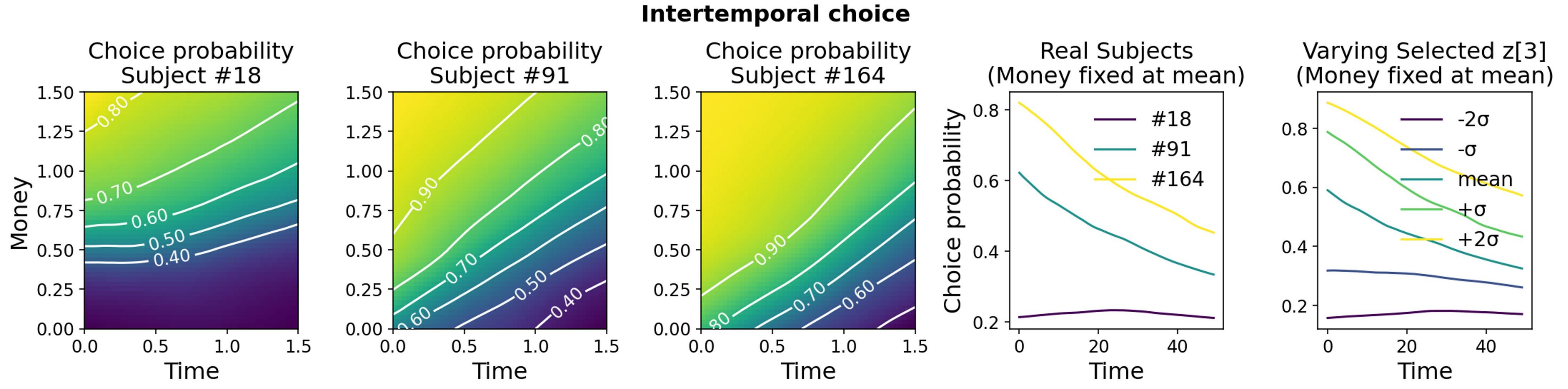
psyarxiv/b865r

Visualization of (X,Z)→Y mappings



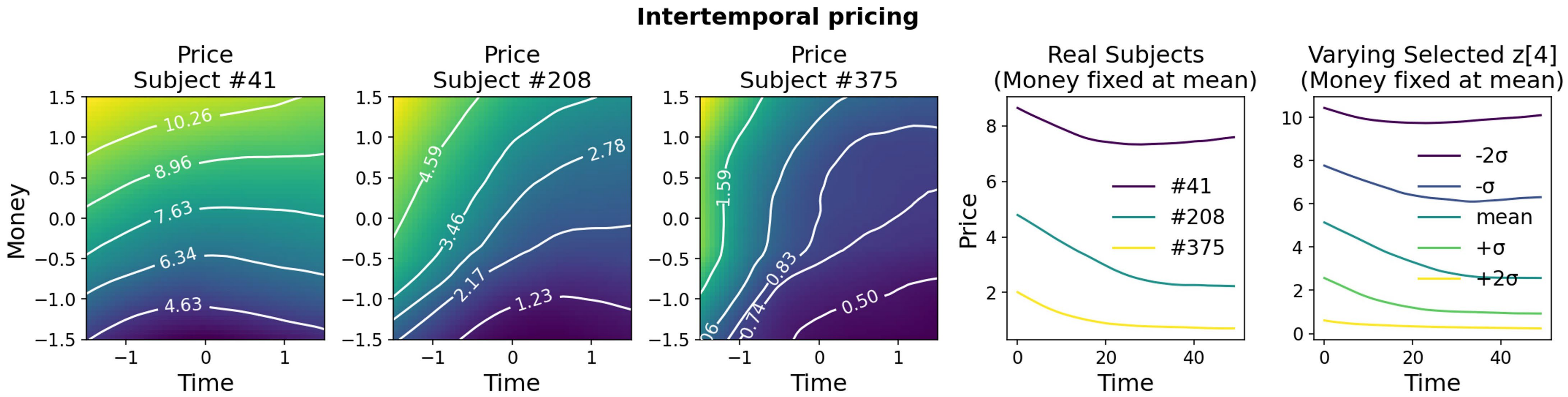
psyarxiv/b865r

Visualization of (X,Z)→Y mappings



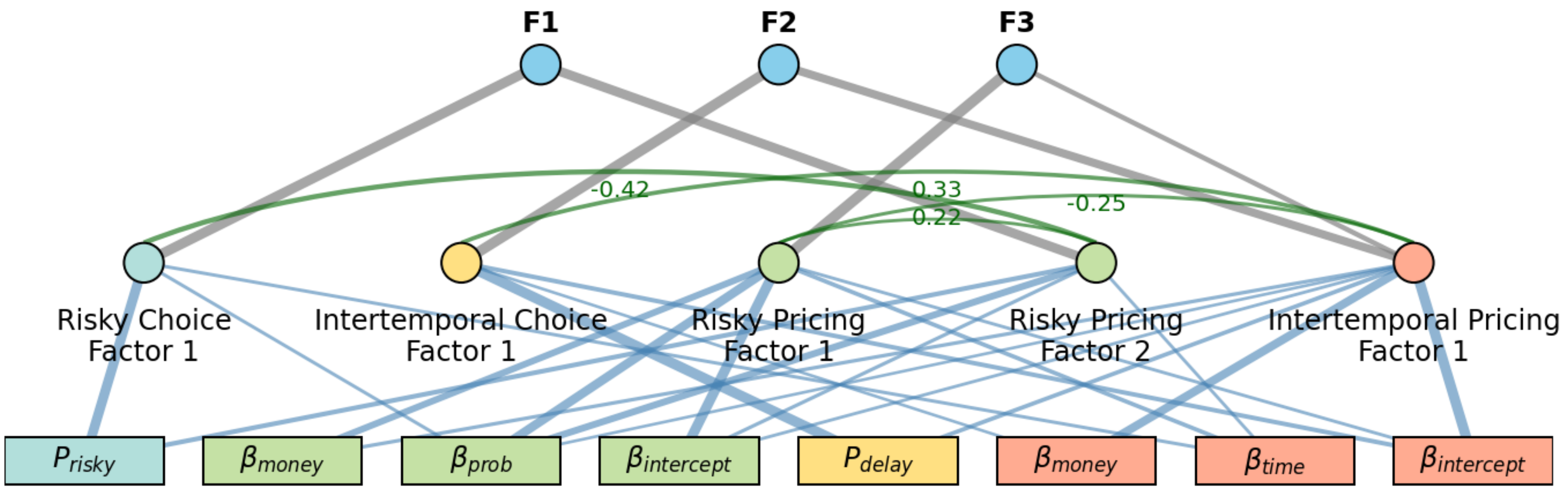
psyarxiv/b865r

Visualization of (X,Z)→Y mappings



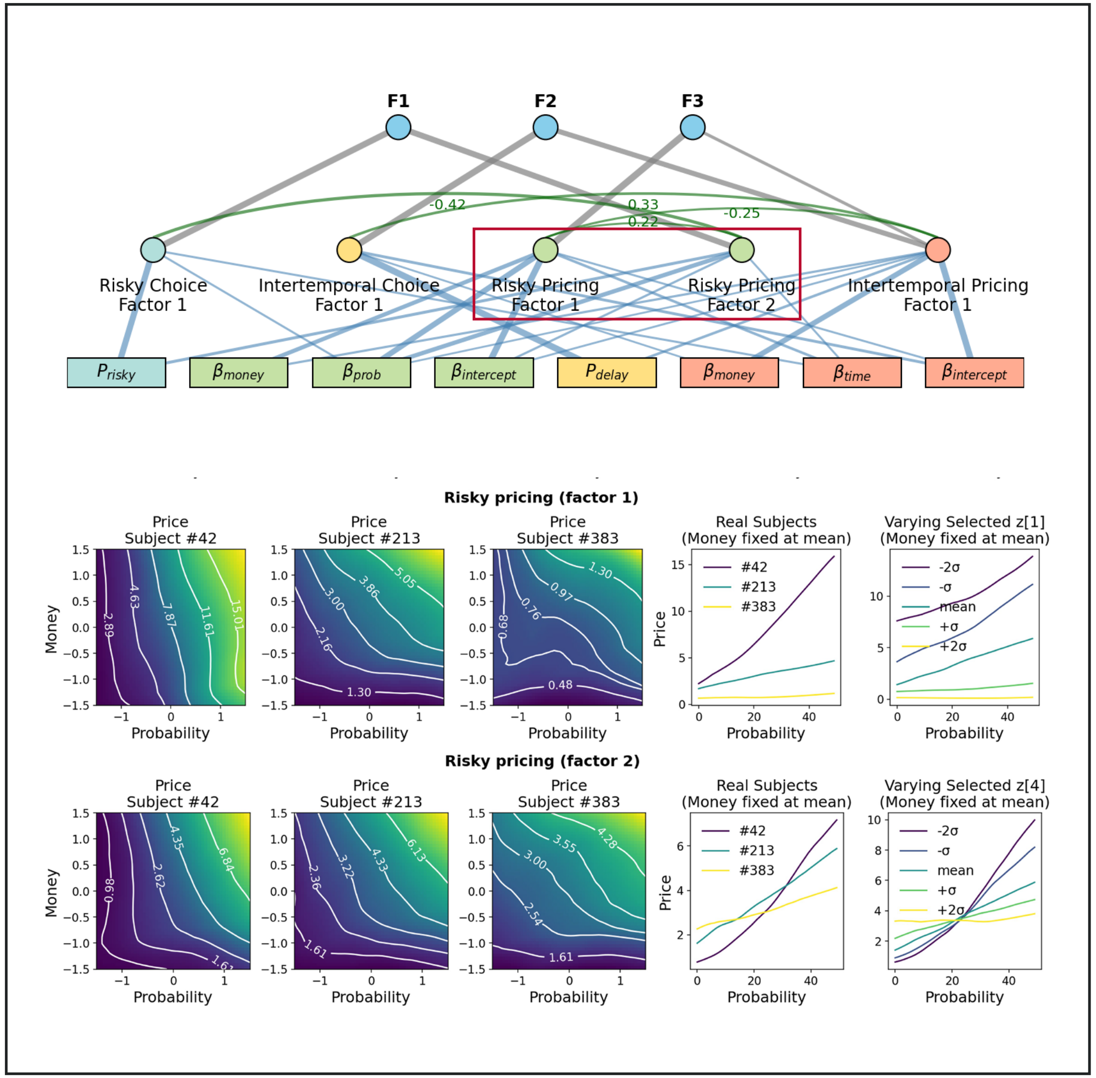
psyarxiv/b865r

Identifying cross-task factors



psyarxiv/b865r

Model discovery



$$\hat{y} = \underbrace{0.316(1.794 - z_1)}_{\text{overall gain term}} \left[6.768 + \underbrace{(0.680 + 0.269 z_2) M}_{\text{reward-magnitude weighting}} + \underbrace{(1.192 - z_2)(P + 0.329)}_{\text{probability weighting}} \right]$$

- z_1 scales the overall valuation strength
- z_4 shifts the trade-off between M and P
- Higher z_4 emphasizes M over P

Symbolic regression (PySR)

i.i.d. trial data: distributional reconstruction

Factor and model discovery for choice-RT data

Method

i.i.d. trial data: distributional reconstruction

Key Idea

- Extend moderational learning from point prediction to the joint modeling of discrete and continuous outcomes.

Instantiation

- **Discrete outcome:** choice
- **Continuous outcome:** response time

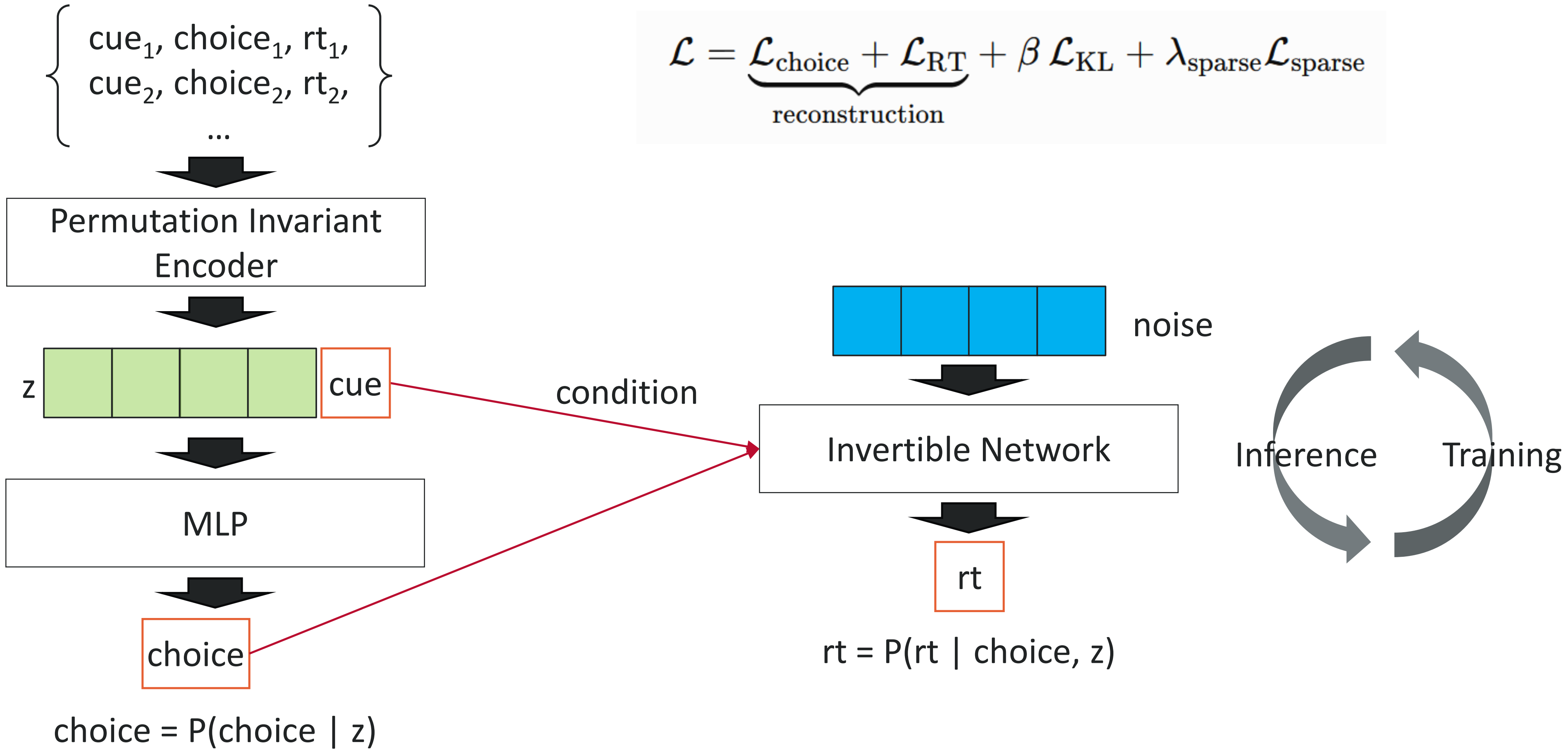
Method

- Model choice using a categorical distribution.
- Model the conditional RT distribution using a flexible **normalizing flow**.
- Apply the chain-rule decomposition:

$$p(choice,RT|X,Z)=p(choice|X,Z)p(RT|choice,X,Z)$$

Method

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{choice}} + \mathcal{L}_{\text{RT}}}_{\text{reconstruction}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}}$$



Simulation Study

Data-Generating Process

For each participant s, define the standard four-parameter DDM:

$$\boldsymbol{\theta}_s = \left(a_s, v_s, z_s, t_{0,s}\right)$$

For each parameter designated as a source of individual differences,

$$\theta_{s,j} \sim p_j\left(\theta_j\right)$$

whereas all remaining parameters are fixed across participants:

$$\theta_{s,j} = \theta_j^{\text{fixed}}$$

Trial-level responses are then generated as

$$\left(c_{si}, RT_{si}\right) \sim \text{DDM}\left(a_s, v_s, z_s, t_{0,s}\right)$$

Manipulating the Number of Free Parameters

- **4 free parameters:** a, v, z, t_0
- **3 free parameters:** a, v, z, a, v, t_0, a, z, t_0, or v, z, t_0
- **2 free parameters:** all two-parameter combinations
- **1 free parameter:** a, v, z, or t_0
- **0 free parameters:** all parameters fixed across participants

Key results

•**Accurate distributional reconstruction**
The model reproduced the joint choice–RT distributions generated by the DDM.

•**Recovery of latent dimensionality**
The number of active latent dimensions matched the true number of free DDM parameters.

•**Recovery of parameter information**
The learned latent dimensions strongly encoded the true participant-level DDM parameter values.

Method

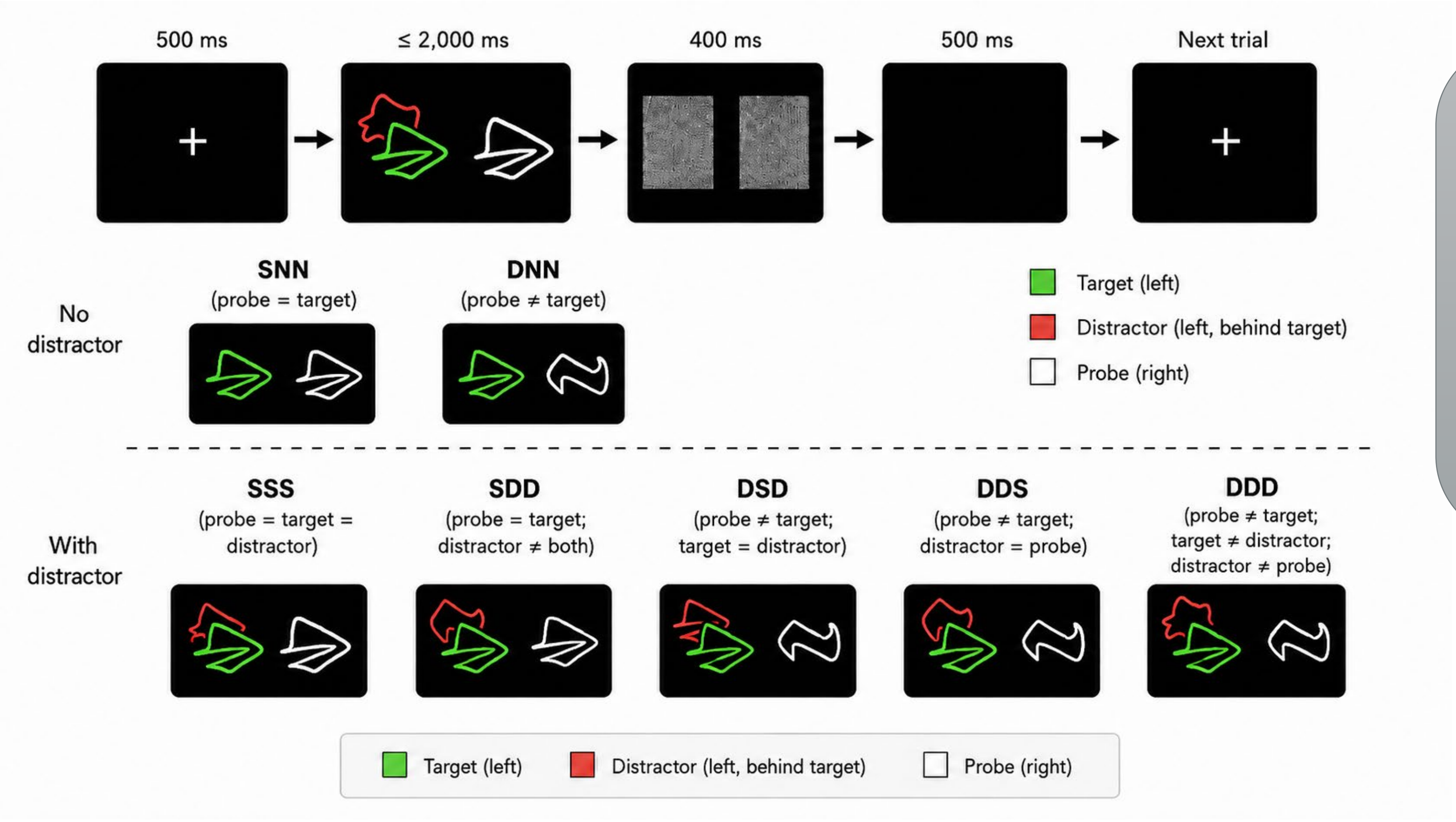
ARTICLE

https://doi.org/10.1038/s41467-019-10301-1 OPEN

Uncovering the structure of self-regulation through data-driven ontology discovery

Ian W. Eisenberg[1], Patrick G. Bissett[1], A. Zeynep Enkavi[1], Jamie Li[1], David P. MacKinnon[2], Lisa A. Marsch[3] & Russell A. Poldrack[1]

Shape-matching task

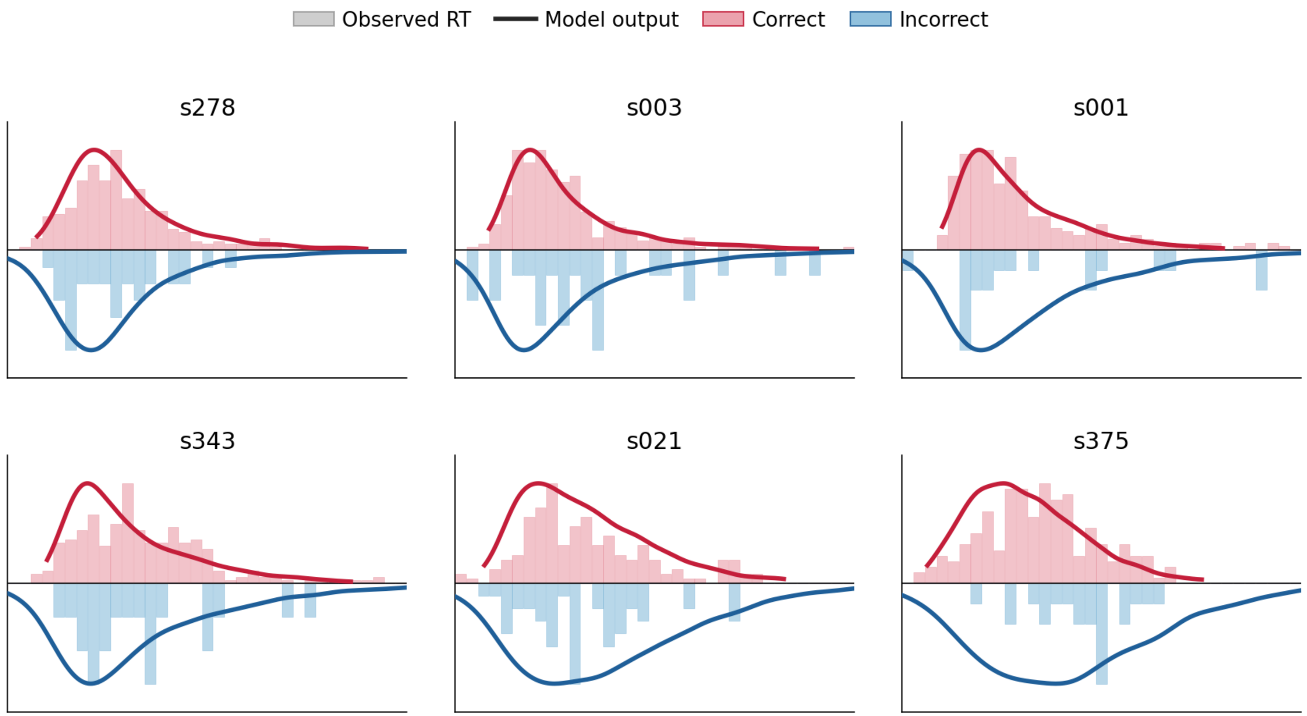


- Many cognitive tasks
- Fitted parameter of basic cognitive models
- Many questionnaires
- Demographic information

Results

1. Capturing Joint Choice–RT Distributions

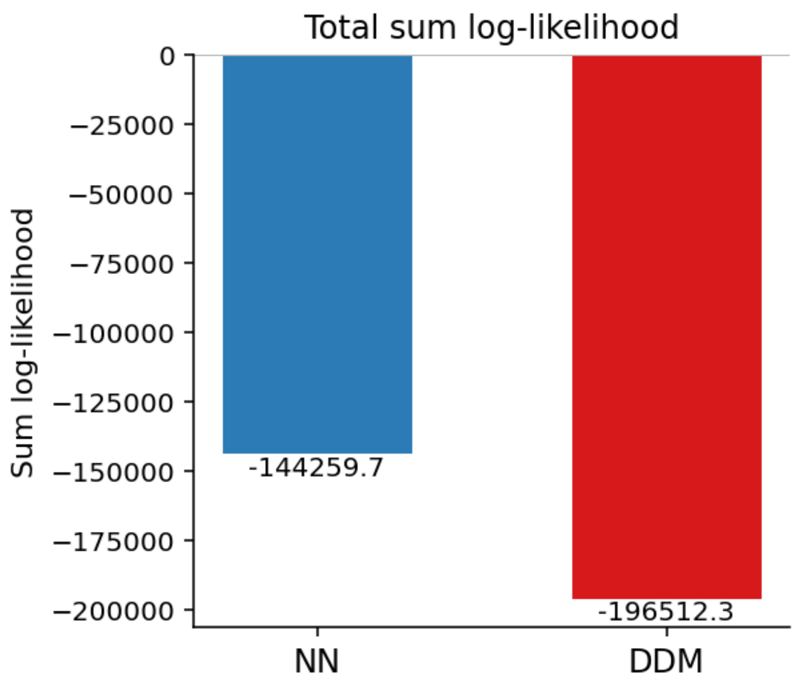
- Predicted and real choice-RT distribution

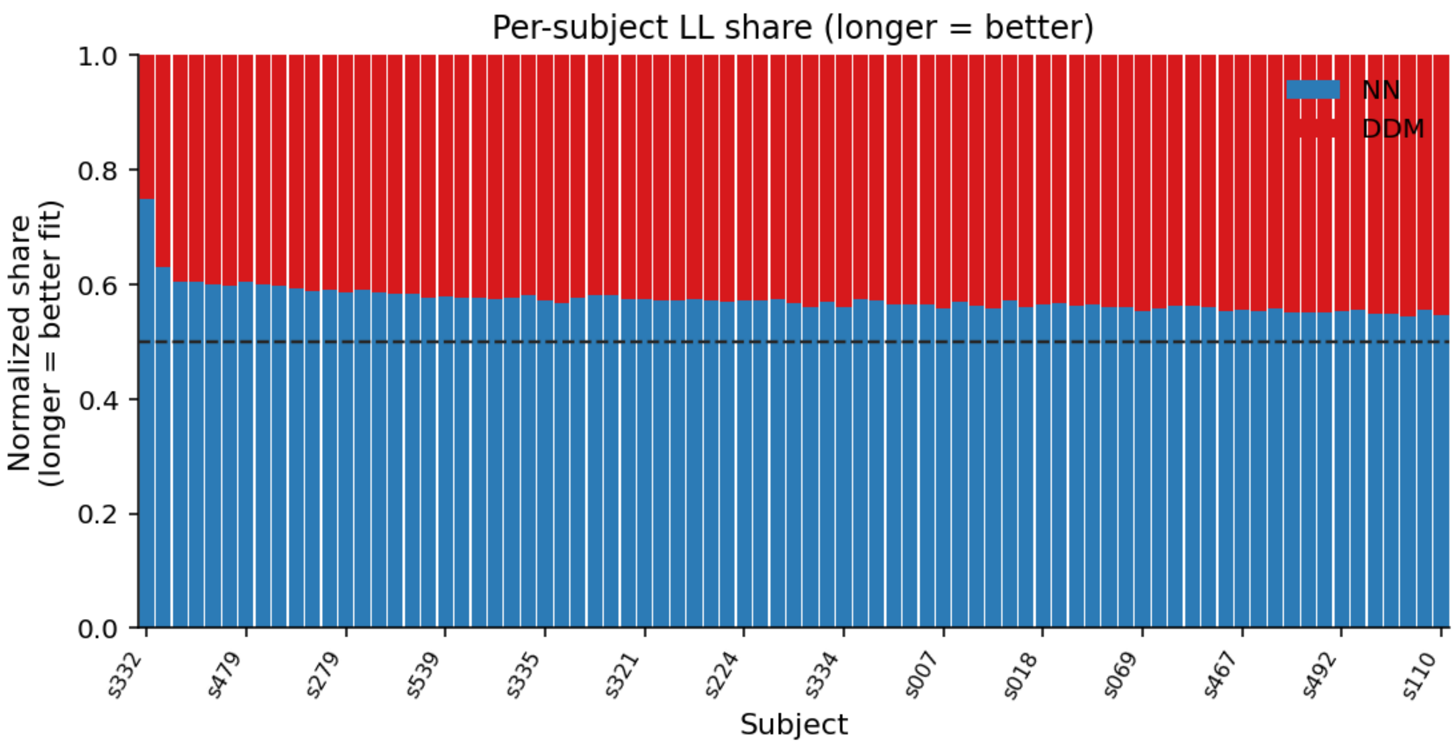


Results

1. Capturing Joint Choice–RT Distributions

- Simulation-based likelihood

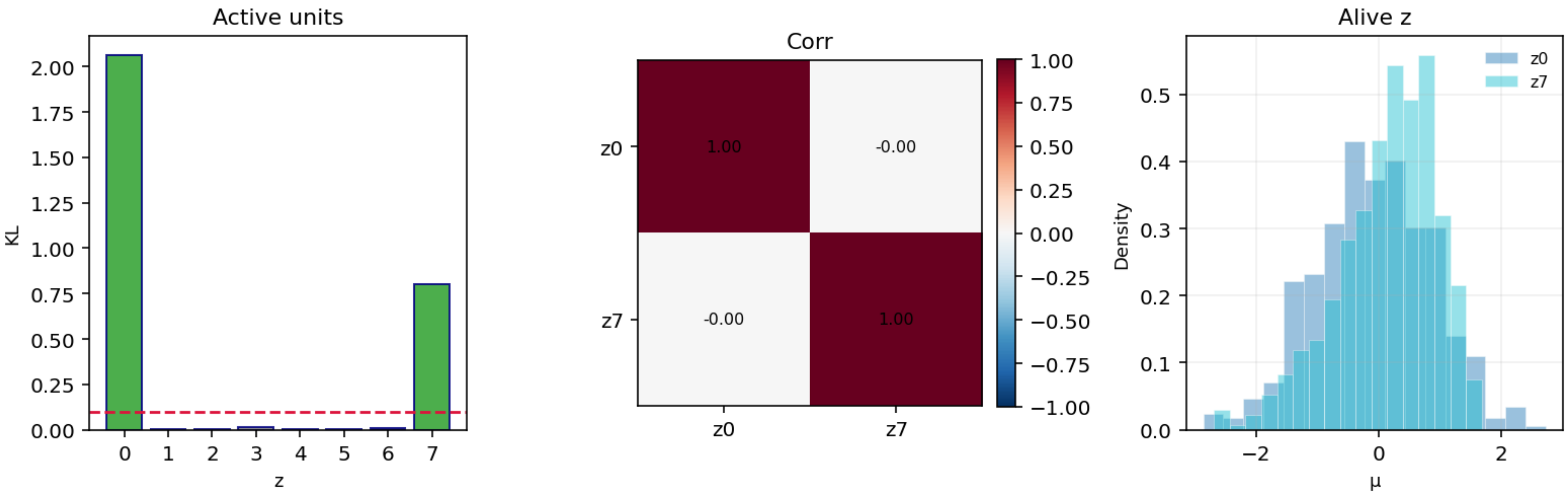




Results

2. Extracting Latent Dimensions

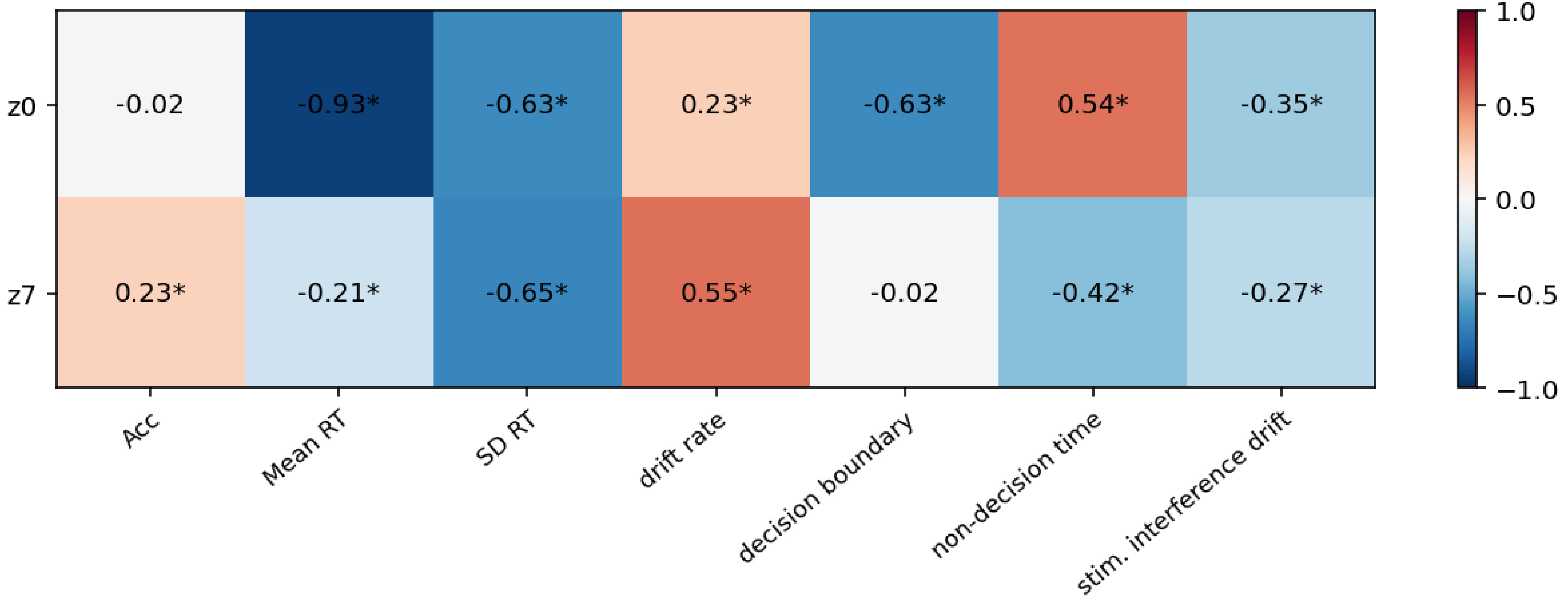
- Visualization of the latent dimensions (mean KL > 0.1)



Results

2. Extracting Latent Dimensions

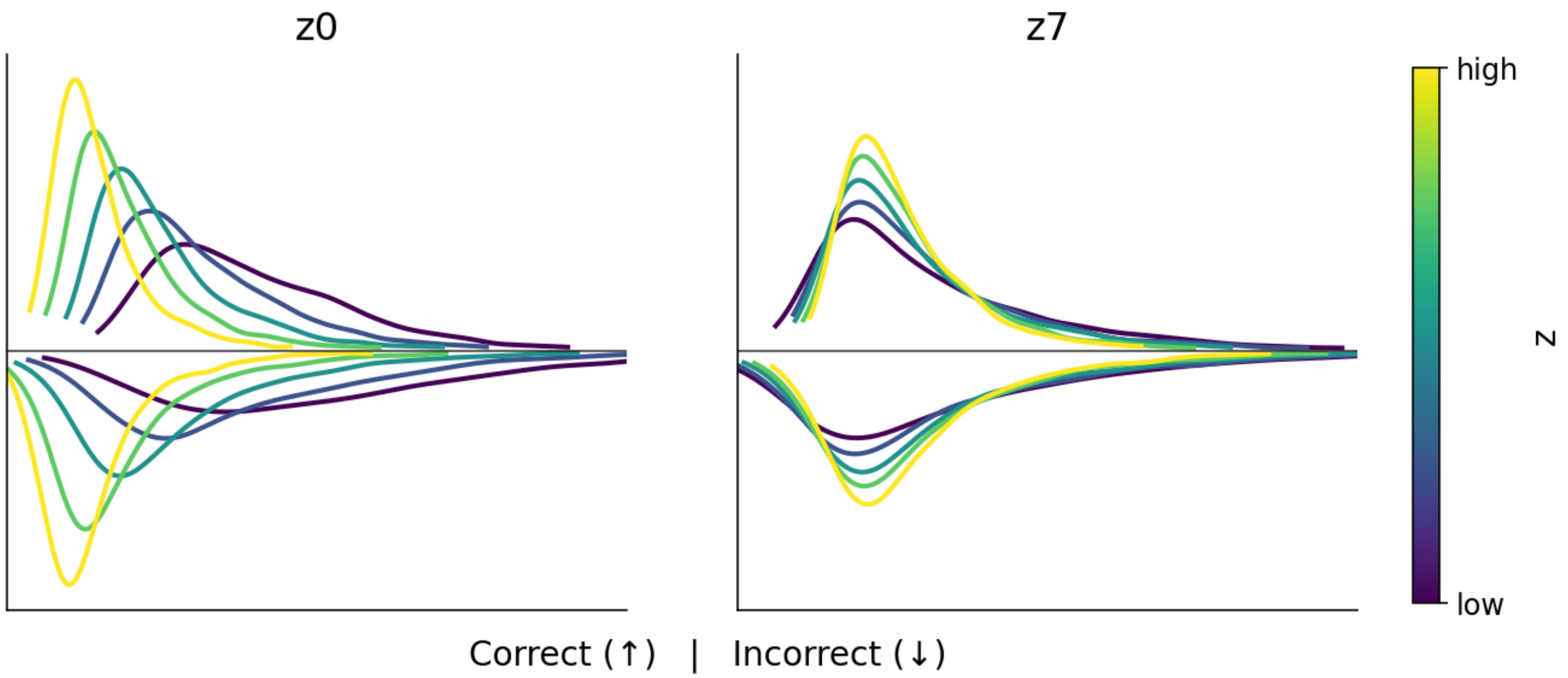
- Active latent dimensions correlating with behavioral summaries & HDDM parameters



Results

2. Extracting Latent Dimensions

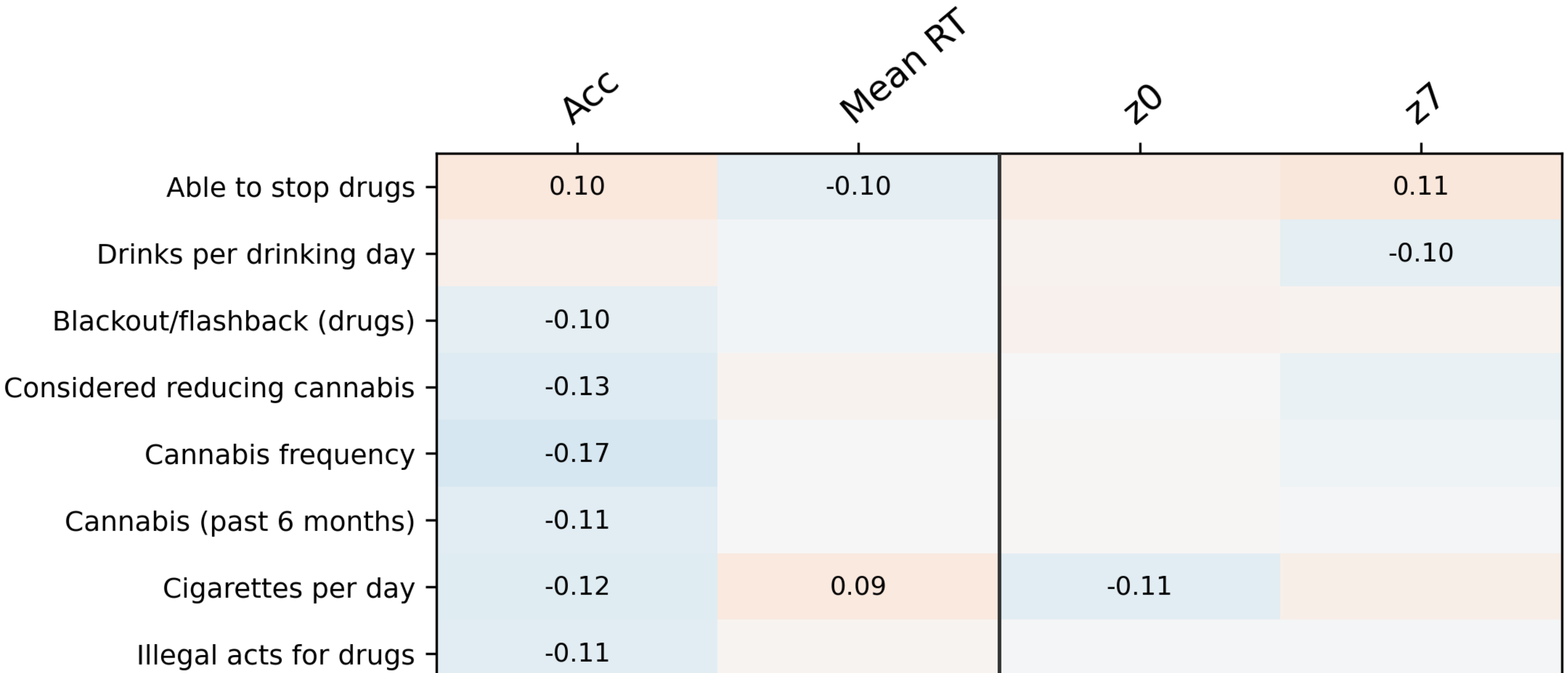
- Active latent dimensions correlating with behavioral summaries & HDDM parameters



Results

3. External Validity of the Latent Dimensions

- Associations with external measures



Results

3. External Validity of the Latent Dimensions

- **Accuracy: Behavioral dysfunction and adverse outcomes**

  e.g., cannabis-related problems, financial debt

- **Mean RT: Demographic and life-course variables**

  e.g., age, family and relationship history, selected impulsivity traits

- **z_0: Risk exposure and risk evaluation**

  e.g., smoking history, perceived risk, perceived benefits of risky behaviors

- **z_7: Self-regulation versus impulsivity**

  e.g., future orientation, self-control, urgency, sensation seeking

Model discovery

A General Stochastic Evidence-Accumulation Framework

GENERAL GENERATIVE FORMULA

X: trial information, Z: subjective representation → choice, response time

Evidence dynamics

$$dA\tau = \mathbf{b}(A\tau, \tau, X, Z)d\tau + \boldsymbol{\Sigma}(A\tau, \tau, X, Z)^{1/2}dB\tau$$

Initial state

$$A_0 = \mathbf{y_0}(X, Z)$$

First-passage stopping rule

$$\tau^* = \inf\{\tau : \max_k \mathbf{S_k}(A\tau, \tau, X, Z) \geq 0\}$$

Observed response

$$c = \mathrm{argmax}_k\, S_k(A\tau^*, \tau^*, X, Z) \qquad RT = \mathbf{T_{er}}(X, Z) + \tau^*$$

FIVE FUNCTIONS TO DISCOVER

- b — **drift / dynamics**: how evidence grows, leaks, competes, or interacts
- Σ — **noise structure**: trial-level stochasticity and accumulator correlations
- y_0 — **starting point**: initial bias before the accumulation process begins
- S_k — **boundary / choice rule**: when to stop and which response is selected
- T_{er} — **non-decision time**: encoding and motor delay added to decision time

Model discovery

Teacher-Guided Symbolic Regression Workflow

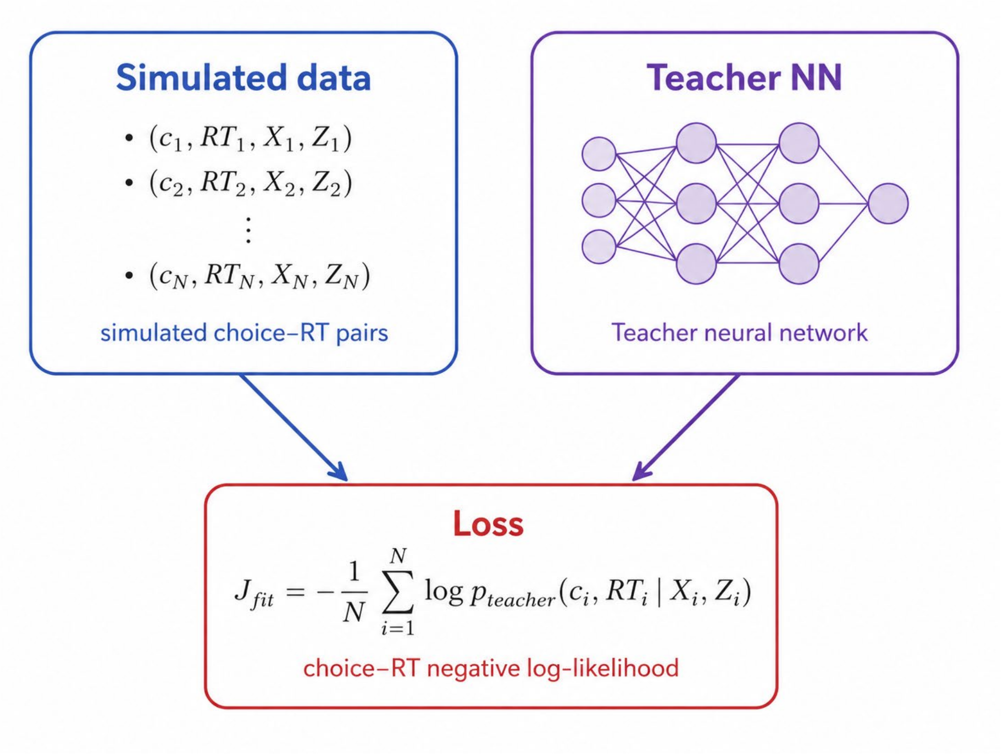
1. Search space
variables: X, Z, Aτ, τ
operators + constraints

2. Candidate model
symbolic set:
$\{b, \Sigma, y_0, S_k, T_{er}\}$

3. First-passage simulation
simulate trajectories
→ choice and RT

4. Teacher scoring
evaluate choice–RT
negative log-likelihood

5. Evolutionary update
select, mutate, crossover,
refit constants



Non-i.i.d. trial data

Factor and model discovery for learning data

History-dependent trial-by-trial behavioral data

Consider a sequential decision-making setting in which each participant s performs a sequence of trials indexed by $t = 1, \ldots, T_s$. At each trial, the observed data are

$$D_{s,t} = (C_{s,t}, Y_{s,t}, F_{s,t}),$$

The behavioral history available before trial t is defined as

$$H_{s,t} = \{(C_{s,\tau}, Y_{s,\tau}, F_{s,\tau})\}_{\tau=1}^{t-1}.$$

We assume that each participant is characterized by a latent vector of true generative factors,

$$\boldsymbol{\theta}_s = (\theta_{s,1}, \ldots, \theta_{s,K}),$$

where K denotes the number of underlying psychological or computational factors governing individual behavior. The true data-generating process is assumed to take the form

$$Y_{s,t} \sim P_{\boldsymbol{\theta}_s}(Y_{s,t} \mid H_{s,t}, C_{s,t}).$$

Learning objectives

- Train an encoder qΦ and decoder pψ:

$$\boldsymbol{z}_s \sim q_\phi\left(\boldsymbol{z}_s \mid D_{s,1:T_s}\right),$$

$$\hat{Y}_{s,t} \sim q_\psi\left(Y_{s,t} \mid H_{s,t}, C_{s,t}, \boldsymbol{z}_s\right).$$

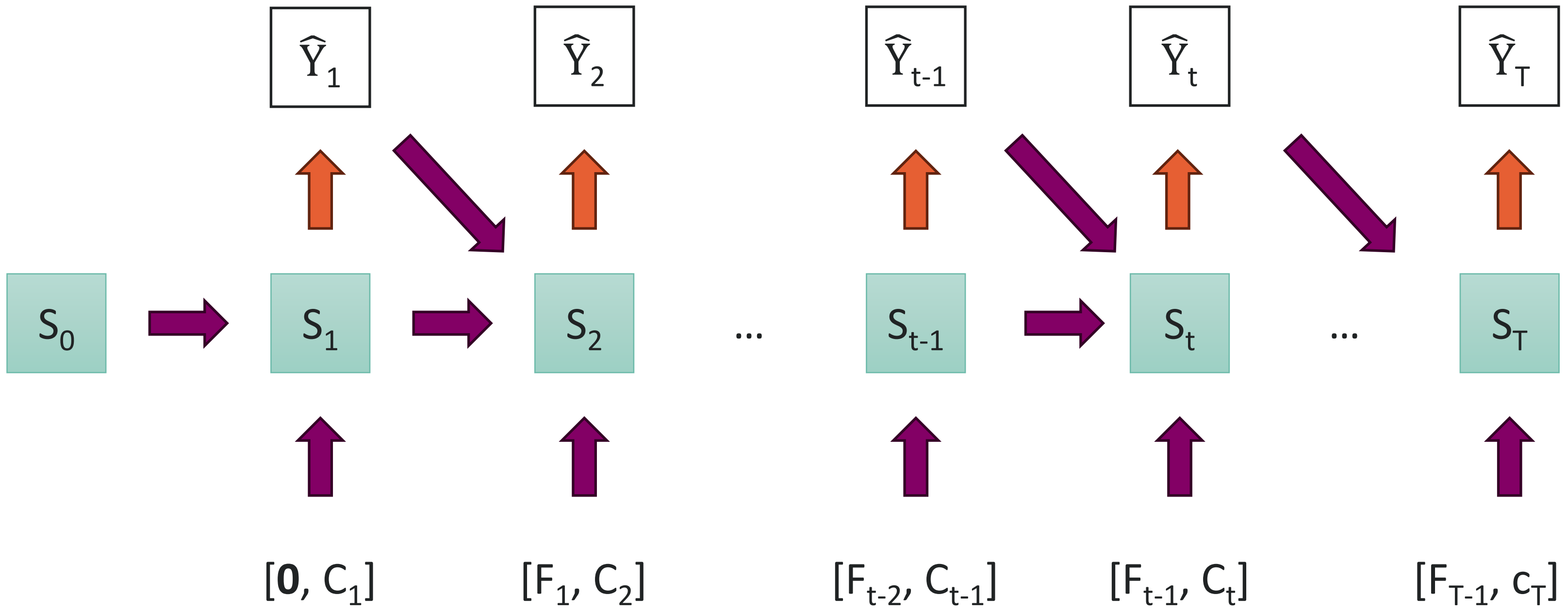
- Aim 1: maximize mutual information between Zs and θs

$$\max_{\phi}\ I(\boldsymbol{\theta}_s; \boldsymbol{z}_s),$$

- Aim 2: accurately reconstruct behavior

$$\min_{\phi,\psi} \sum_{s=1}^{S} \sum_{t=1}^{T_s} \mathcal{L}\left(Y_{s,t}, \hat{Y}_{s,t}\right),$$

Inductive bias: Markov process



Encoder network: $Z \sim q_\varphi(Z \mid [C_{1:T}, Y_{1:T}, F_{1:T}])$

Transition network: $S_t = NN([S_{t-1}, Y_{t-1}, G_{t-1}, C_t] ; Z)$

Emission network: $\widehat{Y}_t = NN(S_t ; Z)$

Preventing posterior collapse

- Fatal hazard

  Learned latent variables barely carry individual difference information, while the data still well reconstructed

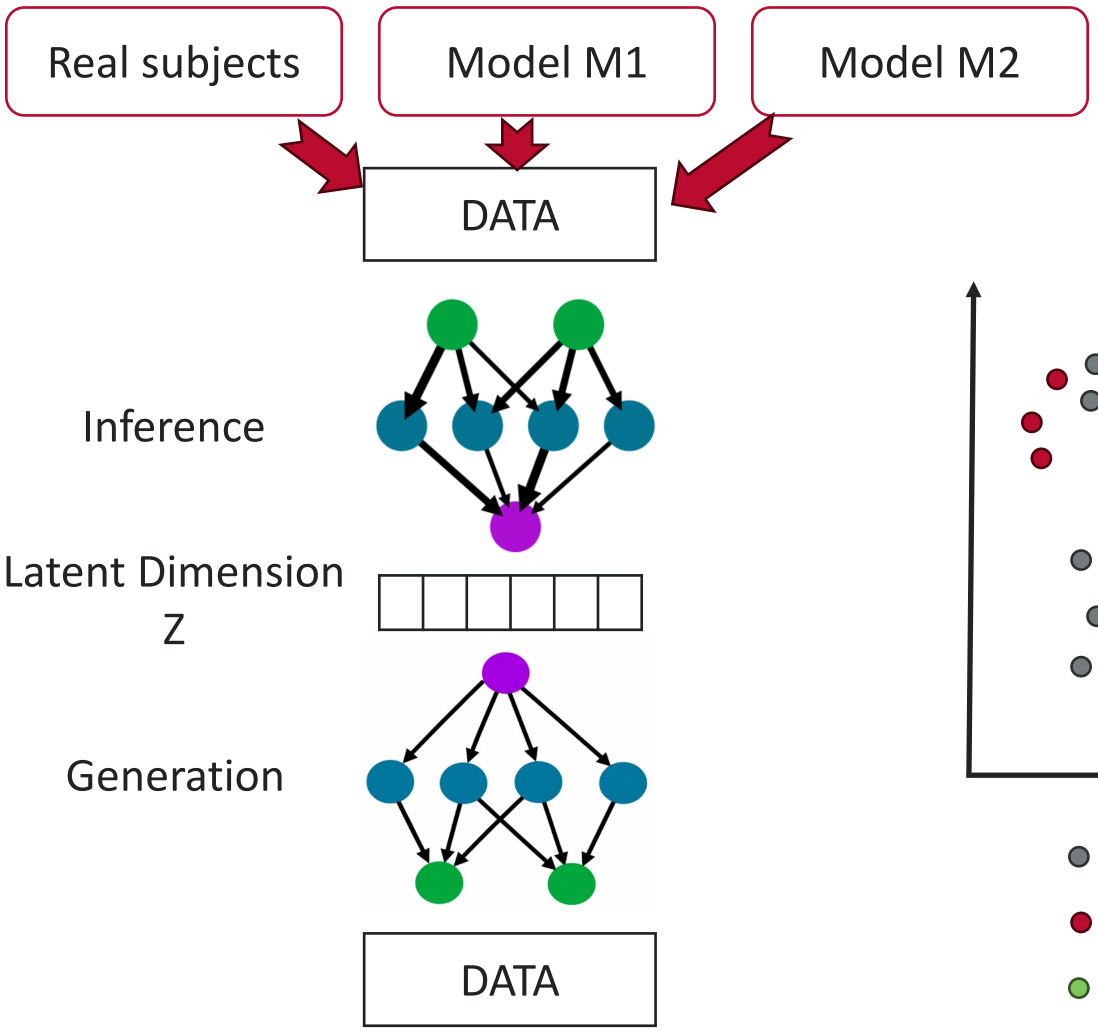
$$I(\boldsymbol{\theta}_s; \boldsymbol{z}_s) \rightarrow 0.$$

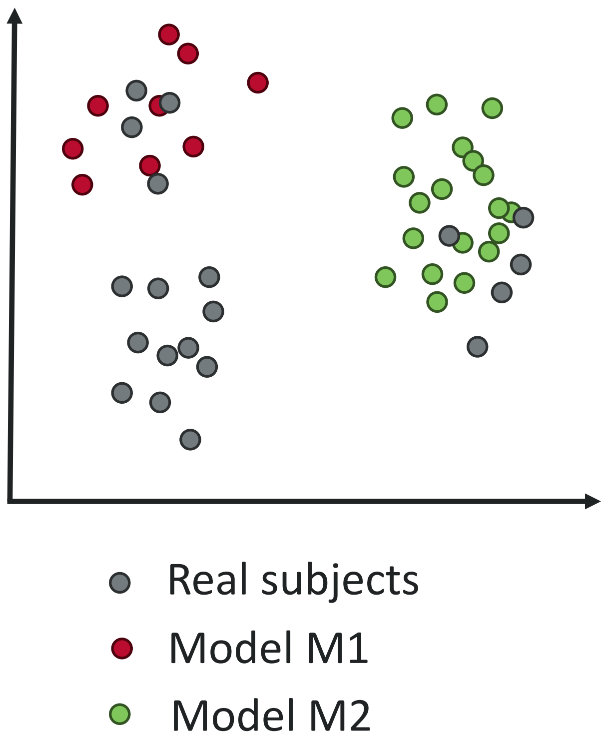
- Treatments
  - **Contrastive loss** — Pull latents from the same subject closer and push latents from different subjects apart.
  - **Switch loss** — Swap z_s between subjects and force predictions to change accordingly.
  - **Decoder capacity limit** — Prevent the decoder from reconstructing behavior using history alone.
  - **KL annealing / free bits** — Avoid early posterior collapse by gradually applying latent regularization.

Moderational learning × SBI

Ongoing projects

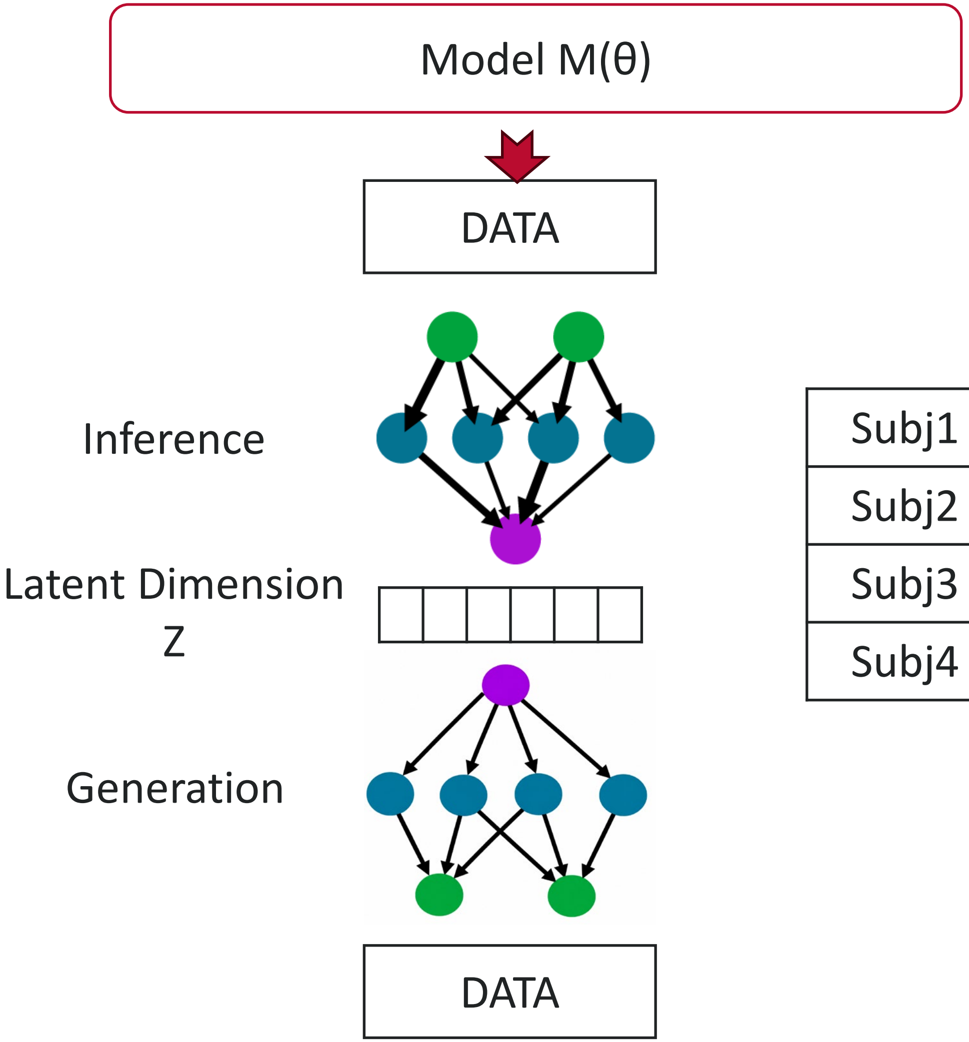
1. Latent space clustering





- ✓ Subject heterogeneity discovery
- ✓ Model separation check
- ✓ Model selection
- ✓ Model family selection

2. Model specification check



Subj1	θ1	Z1
Subj2	θ2	Z2
Subj3	θ3	Z3
Subj4	θ4	Z4

- **Θ-Z mutual information**

How much information about each parameter can be recovered from the learned representation

- **Pairwise Conditional Mutual Information**

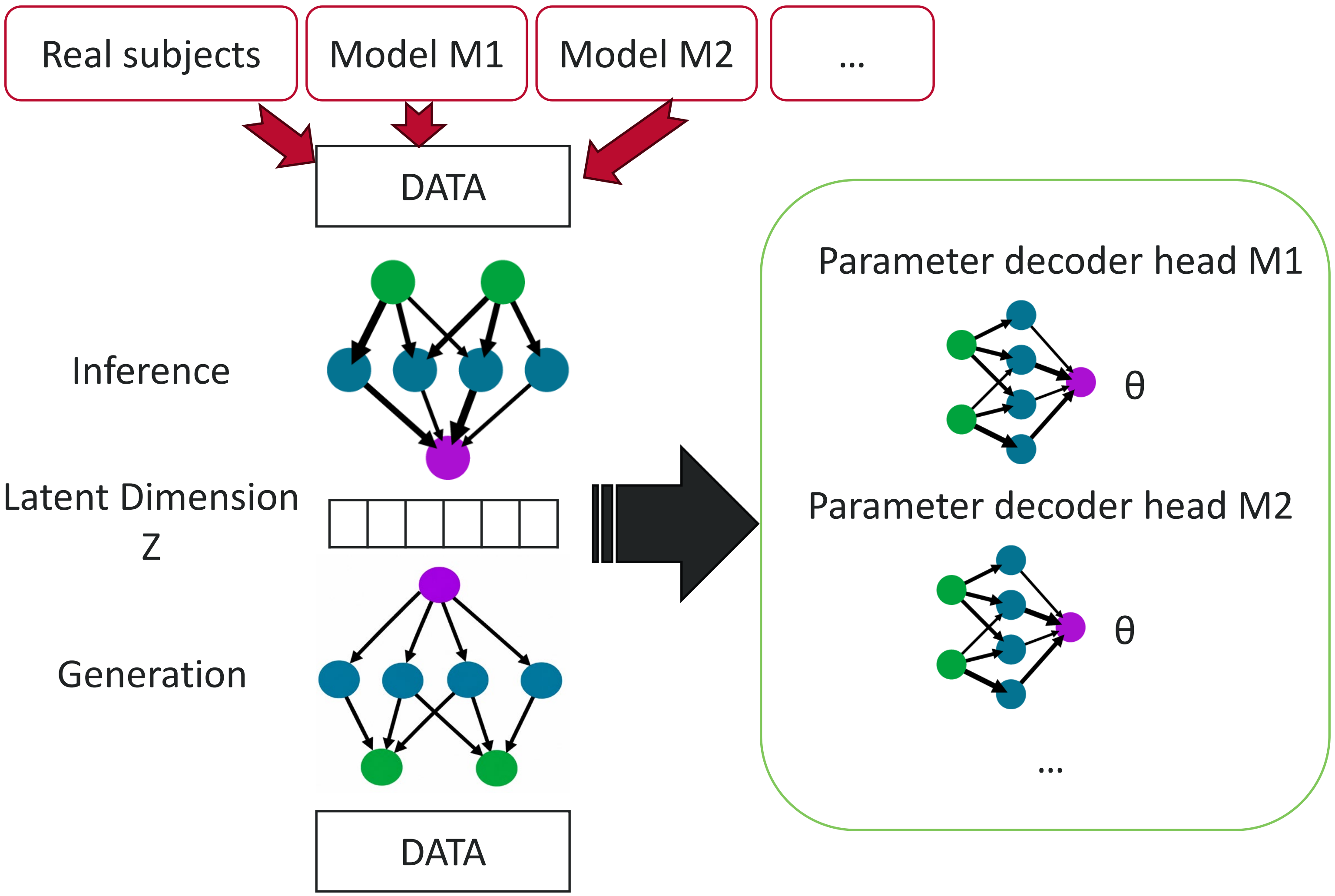
Pairwise parameter trade-offs

- **Parameter-Specific Conditional Mutual Information**

How much information each parameter provide

More …

3. Semi-supervised learning preventing posterior collapse



Loss =
Reconstruction loss
+ regularization
+ parameter recovery

Extended reading

Other machine-learning approached to do model discovery

LLMs-based program search

Article

Mathematical discoveries from program search with large language models

https://doi.org/10.1038/s41586-023-06924-6

Received: 12 August 2023

Accepted: 30 November 2023

Published online: 14 December 2023

Open access

Check for updates

Bernardino Romera-Paredes[1,4] ✉, Mohammadamin Barekatain[1,4], Alexander Novikov[1,4], Matej Balog[1,4], M. Pawan Kumar[1,4], Emilien Dupont[1,4], Francisco J. R. Ruiz[1,4], Jordan S. Ellenberg[2], Pengming Wang[1], Omar Fawzi[3], Pushmeet Kohli[1] ✉ & Alhussein Fawzi[1,4] ✉

Large language models (LLMs) have demonstrated tremendous capabilities in solving complex tasks, from quantitative reasoning to understanding natural language. However, LLMs sometimes suffer from confabulations (or hallucinations), which can

Discovering Symbolic Cognitive Models from Human and Animal Behavior

Pablo Samuel Castro[1] Nenad Tomasev[1] Ankit Anand[1] Navodita Sharma[1] Rishika Mohanta[2 3] Aparna Dev[2] Kuba Perlin[1] Siddhant Jain[1] Kyle Levin[1] Noémi Éltető[1 4] Will Dabney[1] Alexander Novikov[1] Glenn C Turner[2] Maria K Eckstein[1] Nathaniel D Daw[1 5] Kevin J Miller[* 1 6] Kimberly L Stachenfeld[* 1 7]

RNN-based model discovery

Article

Discovering cognitive strategies with tiny recurrent neural networks

https://doi.org/10.1038/s41586-025-09142-4

Received: 15 May 2023

Accepted: 12 May 2025

Published online: 2 July 2025

Li Ji-An[1], Marcus K. Benna[1] & Marcelo G. Mattar[2,3]

Understanding how animals and humans learn fr
decisions is a fundamental goal of neuroscience a
frameworks such as Bayesian inference[1] and reinf

Cognitive Model Discovery via Disentangled RNNs

Kevin J. Miller
Google DeepMind and University College London
London, UK
kevinjmiller@deepmind.com

Maria Eckstein
Google DeepMind
London, UK
mariaeckstein@deepmind.com

Matthew M. Botvinick
Google DeepMind
London, UK
botvinick@deepmind.com

Zeb Kurth-Nelson
Google DeepMind and University College London
London, UK
zebk@deepmind.com

Take-Home Messages

1. **Moderational Learning**

Jointly learns interpretable latent dimensions of individual differences and shared stimulus–response generating process moderated by the latent dimensions.

2. **Distributional Moderational Learning**

Extending the decoder with a normalizing flow enables individual differences to be learned from, and used to reconstruct, full distributions of continuous responses.

3. **From Prediction to Model Discovery**

Using data-driven choice–RT modeling for extracting individual-difference dimensions and discovering new evidence-accumulation models.

Thank you for your attention!

Q & A

PeterKvam.com