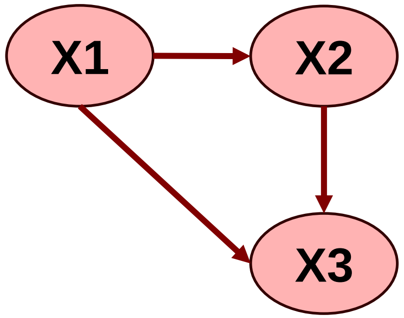
Causal Structure Learning under Distribution Shifts with Federated Learning

Zijiang Yang

Background and Related Work

- Causal Structure:
  - Underlying causal relationships between variables
  - Smoking -> Lung Cancer -> Cough
- DAG:
  - A graph composed of nodes connected by directed edges, but no cycles
- Adjacency Matrix

$$A_{ij} = \begin{cases} 1, & \text{if } X_j \to X_i \\ 0, & \text{otherwise} \end{cases}$$



$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Background and Related Work

- Linear Structural Equation Model (SEM): X = WX+Z   Z is noise
- Gradient-based DAG learning

$$\min_{W \in \mathbb{R}^{d \times d}} \quad F(W) \longrightarrow \frac{1}{2n}\|\mathbf{X} - \mathbf{X}W\|_F^2$$

$$\text{subject to} \quad h(W) = 0, \longrightarrow \operatorname{tr}\left(e^{W \circ W}\right) - d$$

- DAG-GNN

DAG-GNN: DAG Structure Learning with Graph Neural Networks

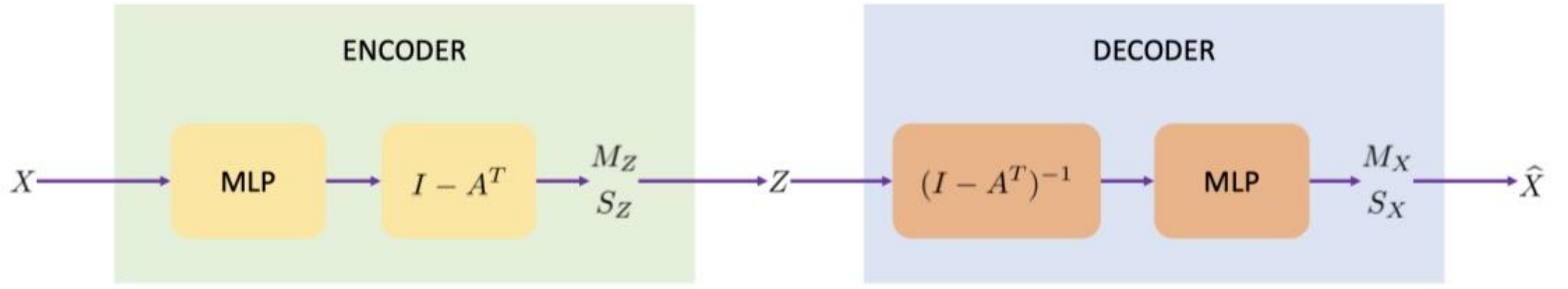


Figure 1. Architecture (for continuous variables). In the case of discrete variables, the decoder output is changed from M_X, S_X to P_X.

Background and Related Work

- FedDAG 2112.03555

Algorithm 1 FedDAG

Input: $\mathcal{D}$, $\mathcal{C}$, Parameter-list $= \{\alpha_{init},\ \rho_{init},\ h_{tol},\ it_{max},\ \rho_{max},\ \beta,\ \gamma,\ r\ \}$

Output: $\mathbb{E}g_{\tau}(\boldsymbol{U}_t), \boldsymbol{\Phi}_t$

#Parameter Initializing

$t \leftarrow 1,\ \alpha_t \leftarrow \alpha_{init},\ \rho_t \leftarrow \rho_{init}$

while $t \leq it_{max}$ and $h(\boldsymbol{U}_t) \geq h_{tol}$ and $\rho \leq \rho_{max}$ **do**

    #Sub-problem Solving

    $U_{t+1}, \Phi_{t+1} \leftarrow \mathrm{SPS}(\mathcal{D}, \mathcal{C}, \alpha_t, \rho_t, it_{in}, it_{fl}, r)$

    #Coefficients Updating

    $\alpha_{t+1} \leftarrow \alpha_t + \rho_t \mathbb{E}[h(\boldsymbol{U}_{t+1})], \quad t \leftarrow t+1$

    if $\mathbb{E}[h(\boldsymbol{U}_{t+1})] > \gamma \mathbb{E}[h(\boldsymbol{U}_t)]$ **then**

        $\rho_{t+1} = \beta \rho_t$

    else

        $\rho_{t+1} = \rho_t$

    end if

end while

Background and Related Work

- FedCDH 2402.13241

Algorithm 1 FedCDH: Federated Causal Discovery from Heterogeneous Data

Input: data matrix $\mathcal{D}_k \in \mathbb{R}^{n_k \times d}$ at each client, $k, \mho \in \{1, \ldots, K\}$
Output: a causal graph $\mathcal{G}$

Client executes:

1: *(Summary Statistics Calculation)* For each client k, use the local data $\mathcal{D}_k$ to get the sample size n_k and calculate the covariance tensor $\mathcal{C}_{\mathcal{T}_k}$, and send them to the server.

Server executes:

2: *(Summary Statistics Construction)* Construct the summary statistics by summing up the local sample sizes and the local covariance tensors: $n = \sum_{k=1}^{K} n_k$, $\mathcal{C}_{\mathcal{T}} = \sum_{k=1}^{K} \mathcal{C}_{\mathcal{T}_k}$.

3: *(Augmented Graph Initialization)* Build a completely undirected graph $\mathcal{G}_0$ on the extended variable set $\boldsymbol{V} \cup \{\mho\}$, where $\boldsymbol{V}$ denotes the observed variables and $\mho$ is surrogate variable.

4: *(Federated Conditional Independence Test)* Conduct the federated conditional independence test based on the summary statistics, for skeleton discovery on augmented graph and direction determination with one changing causal module. In the end, get an intermediate graph $\mathcal{G}_1$.

5: *(Federated Independent Change Principle)* Conduct the federated independent change principle based on the summary statistics, for direction determination with two changing causal modules. Ultimately, output the causal graph $\mathcal{G}$.

Problem Statement

1. **Current Popular Causal Assumptions**: Datasets are IID and share one causal structure, with the same weights (same W). Some research may assume non-IID setting, but still assume the same shared weights.
2. **Applying Fed-Learning to structure learning in heterogeneous setting:** Datasets are non-IID, sharing the same causal structure (same adjacency matrix A) but different weights W.

Goal:

In heterogeneous setting, learn a global encoder and decoder to capture the shared adj_A robustly

Experiment Setup

- DAG-GNN model:
  - MLPEncoder:
    - 2 MLP layers
    - Get sample z and q(z | X, adj_A)
  - MLPDecoder
    - 2 MLP layers
    - Get reconstructed x, new adj_A and p(x|z, adj_A)
  - Loss function:
    - ELBO+constraint

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z},\mathbf{A})\right] - \mathrm{KL}(q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})) + \lambda \cdot h(\mathbf{A})$$

Experiment Setup

- Generate a shared Directed Acyclic Graph A_true, with 10 nodes and 10 edges.
- **Heterogeneous** datasets
  - For k=1,…, 10:
    - generate W_k: Wij ~ Uniform([-3,-0.5] U [0.5, 3]) according to A_true
    - X and Z ~ Gaussian
    - D_k consists of 1000 data according to X = W_k ·X + Z

- Baseline
  - Combine 10 datasets together and directly use DAG-GNN to learn the adj_A
- FedAvg
  - 10 clients, 1 dataset for each client
  - Aggregate after 10 local update iterations
  - Learn the global encoder-decoder, apply it to all datasets

Experiment Setup

- Metrics
  - False Discovery Rate(FDR)

$$\mathrm{FDR} = \frac{\text{False Positives (FP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

  - False Positive Rate(FPR)

$$\mathrm{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

  - True Positive Rate(TPR)

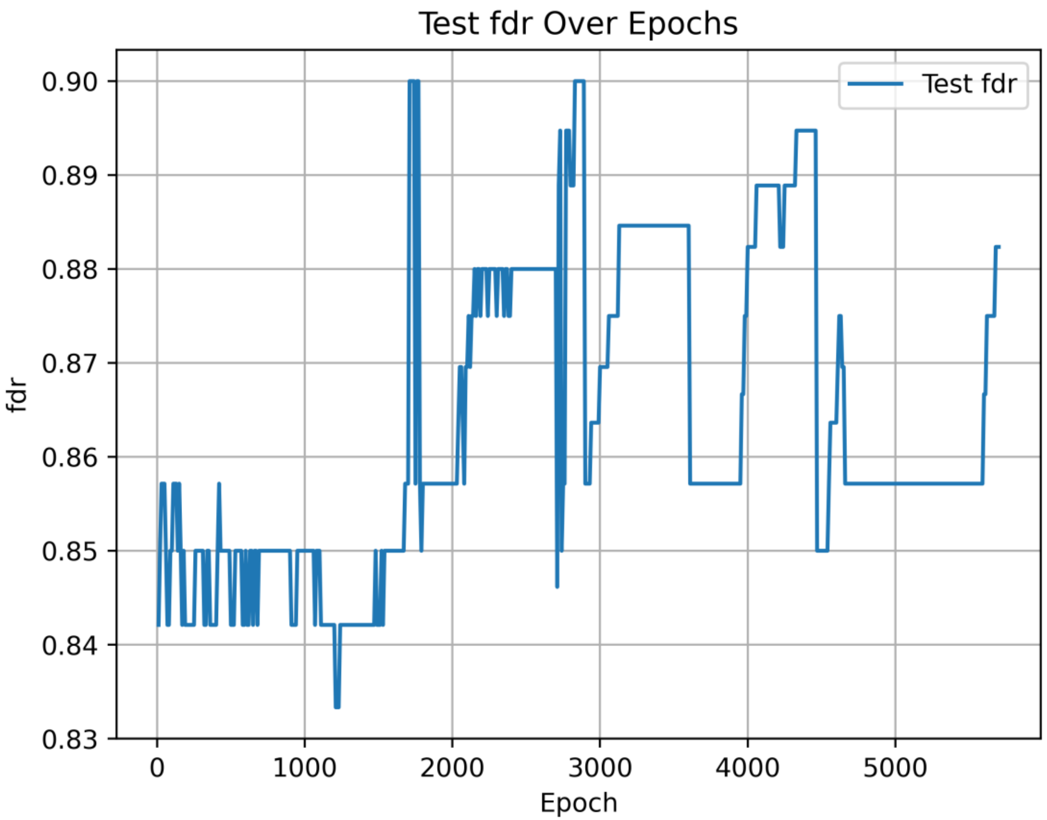
$$\mathrm{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

  - Structural Hamming Distance(SHD)

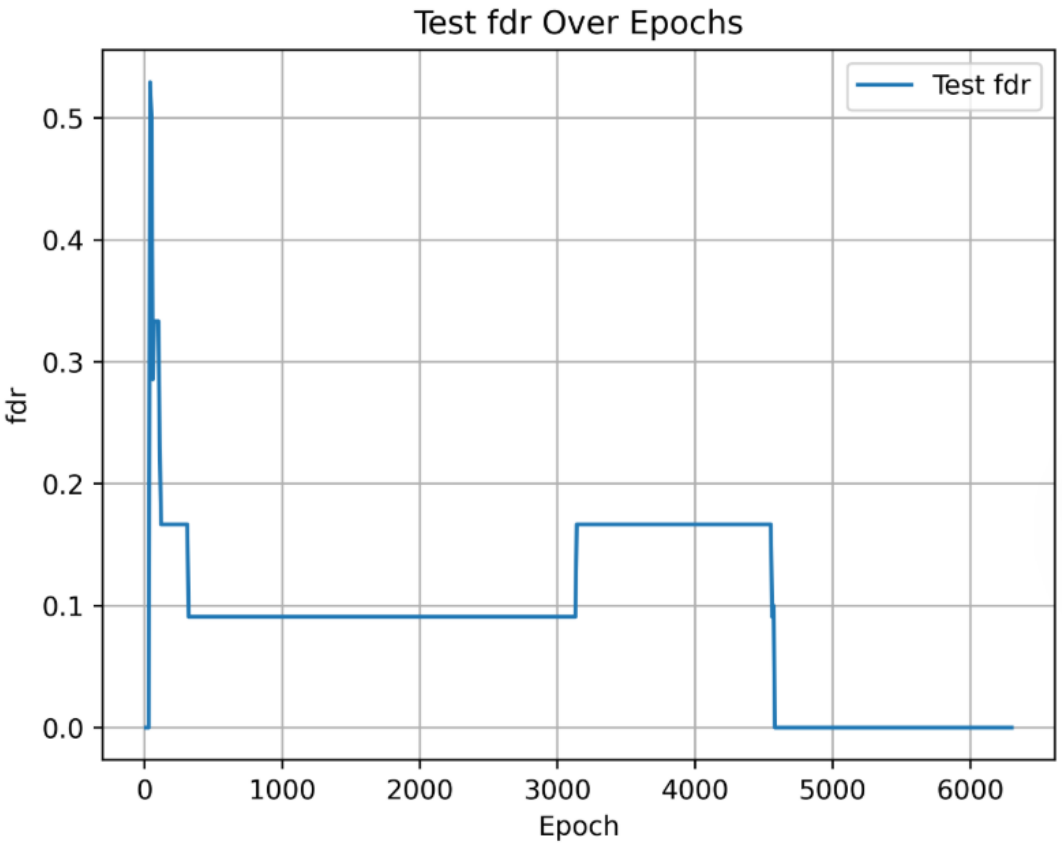
$$\mathrm{SHD} = \#(\text{Extra edges}) + \#(\text{Missing edges}) + \#(\text{Reversed edges})$$

Experiment Result

- Baseline

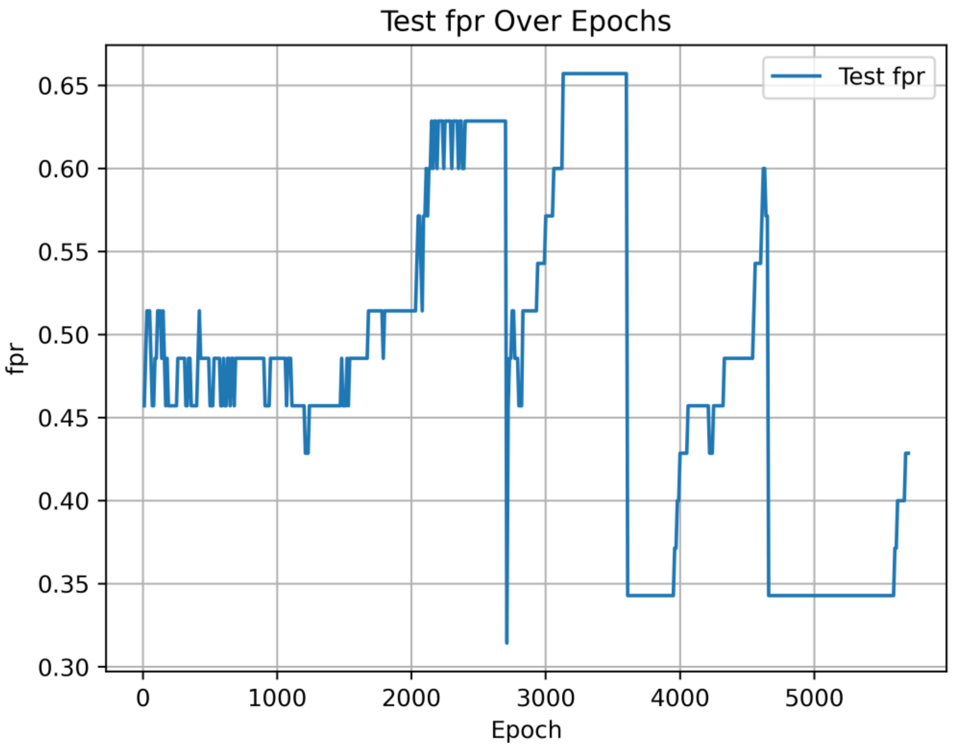


- FedAVG

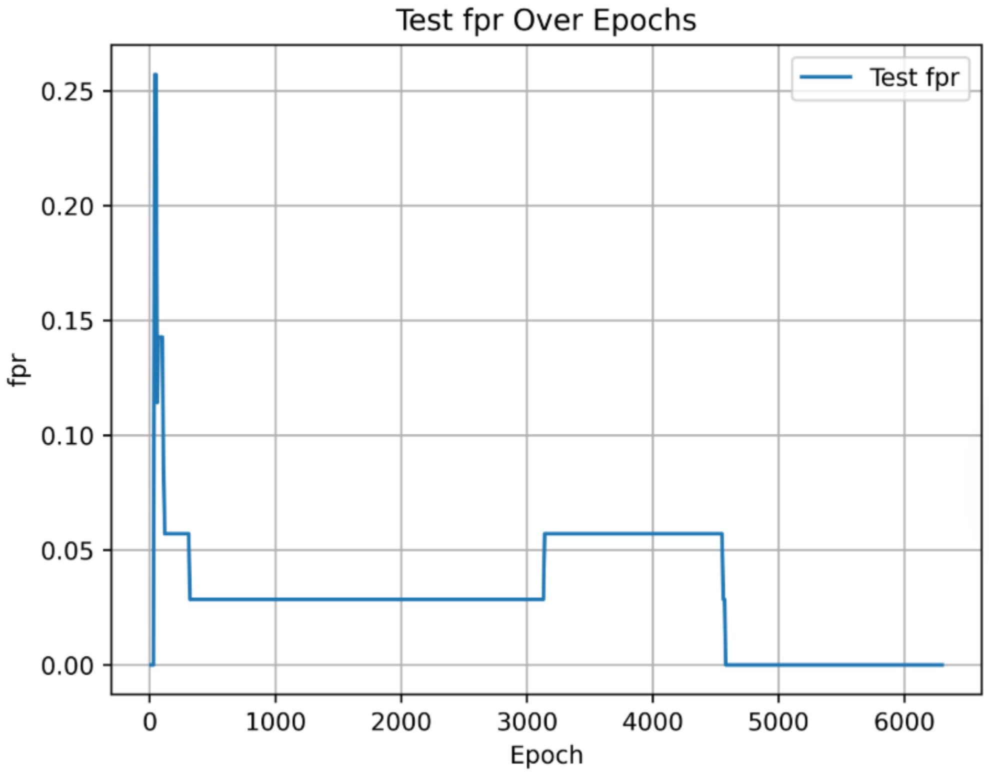


Experiment Result

- Baseline

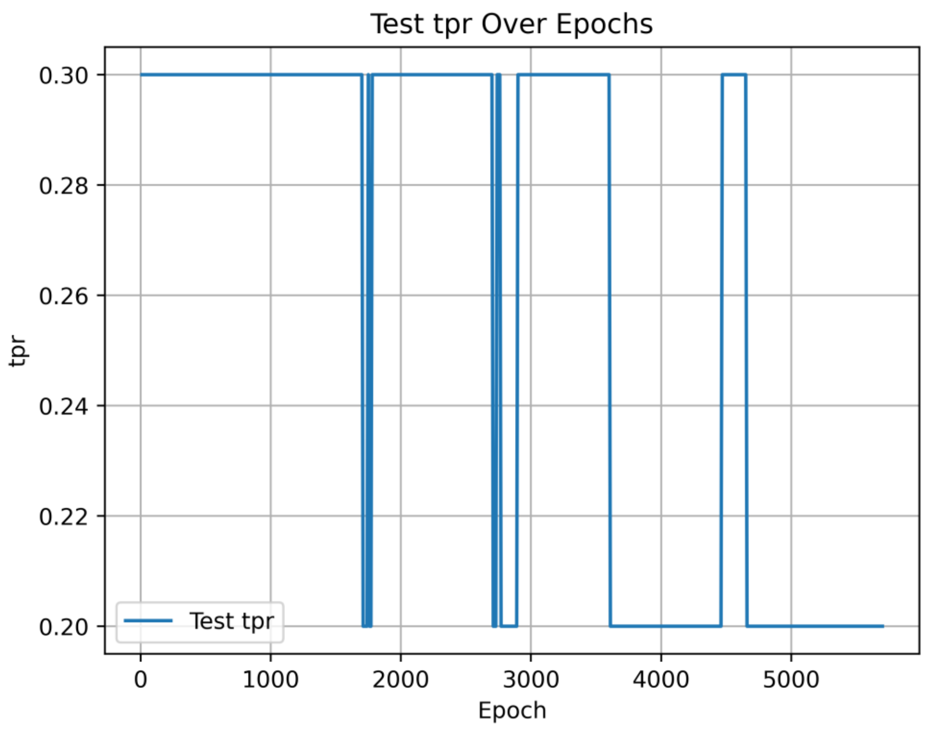


- FedAVG

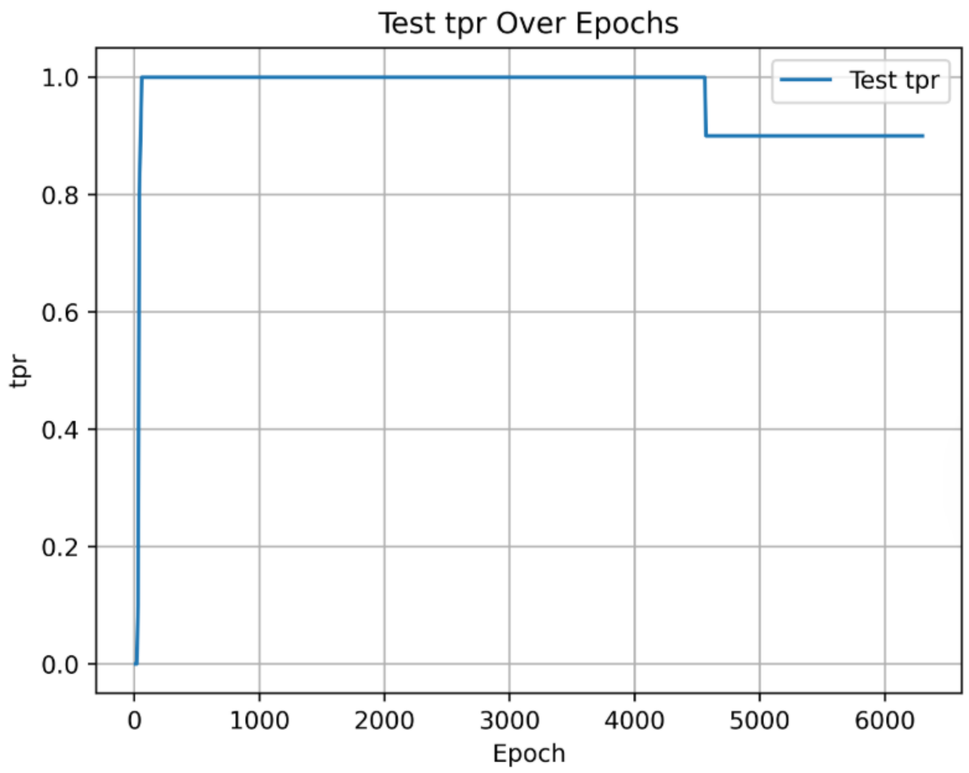


Experiment Result

- Baseline

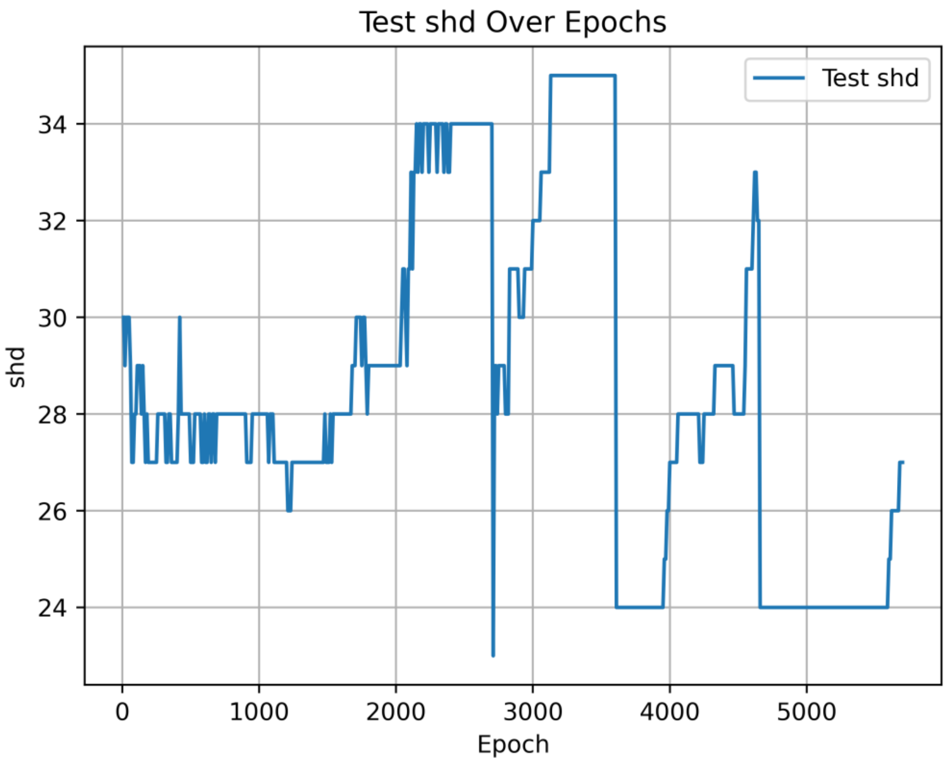


- FedAVG

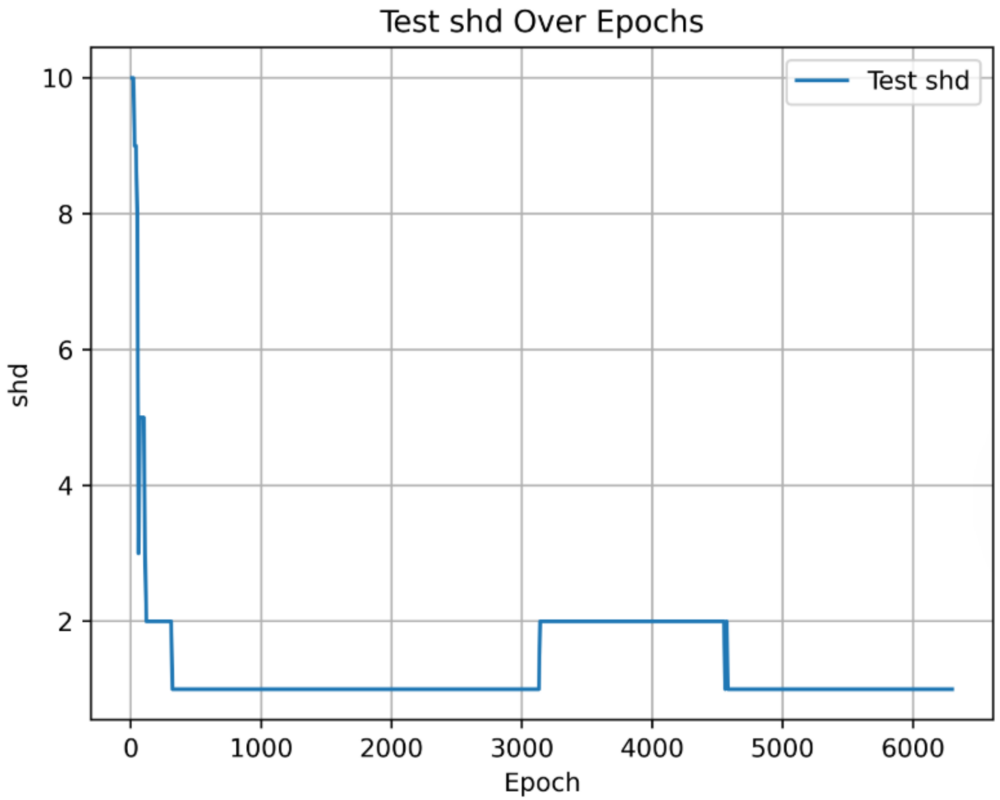


Experiment Result

- Baseline



- FedAVG



Further Research

- Apply the framework to unbalance datasets, use
- Use FedRobust instead to make the algorithm more robust to heterogeneous data settings
- Personalize encoder and decoder for each clients, but keep the models sharing some parameters

Q&A

- What kind of personalization method do you think is reasonable in this problem?

Reference

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. *DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization*. 2023. arXiv: 2209.08037 [cs.LG]. URL: https://arxiv.org/abs/2209.08037.

Xianjie Guo et al. "FedCSL: A scalable and accurate approach to federated causal structure learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 11. 2024, pp. 12235–12243.

Ignavier Ng and Kun Zhang. *Towards Federated Bayesian Network Structure Learning with Continuous Optimization*. 2022. arXiv: 2110.09356 [cs.LG]. URL: https://arxiv.org/abs/2110.09356.

Qiaoling Ye, Arash A. Amini, and Qing Zhou. "Federated Learning of Generalized Linear Causal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.10 (2024), pp. 6623–6636. DOI: 10.1109/TPAMI.2024.3381860.

Yue Yu et al. *DAG-GNN: DAG Structure Learning with Graph Neural Networks*. 2019. arXiv: 1904.10098 [cs.LG]. URL: https://arxiv.org/abs/1904.10098.

Xun Zheng et al. *DAGs with NO TEARS: Continuous Optimization for Structure Learning*. 2018. arXiv: 1803.01422 [stat.ML]. URL: https://arxiv.org/abs/1803.01422.

Thanks!