NEURALK

Technical Interview: Research Intern Position

This technical interview is designed to be a collaborative and creative exercise. We are not looking for a single "correct" solution. Instead, we want to see how you approach a challenging problem, how you think about complex systems, and how you translate your ideas into code.

A Synthetic Data Generation Pipeline for Pre-training a Time Series Tabular Foundation Model

Context:
The success of foundation models in NLP and Vision is largely attributed to the vast scale and diversity of their pre-training data. Replicating this success for time series and tabular data is a major challenge due to the scarcity of large, diverse, and publicly available datasets. A promising direction is the creation of high-fidelity synthetic data that can mimic the complexity and variety of real-world processes. Causal models, such as Structural Causal Models (SCMs), offer a principled framework for generating such data. Recent models like TabPFN [Hollmann et al., 2023] and TabICL [Qu et al., 2025] demonstrate the power of pre-training, but they also highlight the dependency on high-quality synthetic data.

The new central challenge is not just pre-training on existing data, but *designing the data generation process itself*—a process that is scalable, controllable, and can produce data with the rich structural and temporal dynamics necessary to teach a foundation model generalizable representations.

Objective:
You will start with an existing open-source tool from a project called **TabICL**.

Path to Code Repository: https://github.com/soda-inria/tabicl/tree/main/src/tabicl/prior

Guidance on Where to Start:
- Familiarize yourself with the codebase.
- Brainstorm on ideas to add temporal dynamics. (Autoregression, Lagged Effects, Trend and Seasonality…)

Design and implement a novel pipeline for generating synthetic time series tabular data. This pipeline will be based on adapting the static, graph-based Structural Causal Model (SCM) generator from TabICL open-source repository to incorporate temporal dynamics. The generated data is intended for the pre-training of a hypothetical **TempTabFM**, a time series tabular foundation model. This exercise focuses on your ability to creatively solve a complex data generation problem, articulate your design principles, and deliver a robust, high-quality implementation.

Instructions

- **Duration:** 4-6 hours (this is a guideline; you can take more time if needed).
- **Submit:**
    - The complete source code project for the data generation pipeline.
    - A brief report (1-2 pages) explaining your approach, hypothesis, design choices, and a framework for evaluating the quality of the generated data.
    - A comprehensive README file for project setup and code execution.

Problem Statement

Context

Your research group is developing **TempTabFM**, a new foundation model for time series tabular data. The model's architecture is ready, but it requires a massive and diverse pre-training corpus that simply does not exist. Your task is to build the engine that creates this data.

Your starting point is the well-established concept of using a graph-based SCM to generate *static* tabular data. In this approach, a Directed Acyclic Graph (DAG) defines the causal relationships between variables, and data is sampled by propagating values through the graph. Your core task is to **extend and adapt the static SCM paradigm from TabICL to generate complex time series tabular data.**

Exploration Ideas

Your data generation pipeline could have some of the following features:

1. **Temporally Coherent:** The generated data must exhibit meaningful temporal dependencies (e.g., autoregression, trends, seasonality). Simply generating independent tabular samples at each time step is not enough.
2. **Structurally Diverse:** The pipeline must be able to generate a wide variety of datasets, controlled by parameters. This includes varying the number of variables, the underlying causal graph structure, the complexity of relationships, and the nature of the temporal dynamics.

Evaluation:

Since there is no "ground truth" for pre-training data, evaluating the quality of your generator is a creative and critical task.

- Design a rigorous evaluation framework to assess the quality and diversity of the data produced by your two strategies.
- Explain why your chosen evaluation metrics are appropriate for determining if the data is "good" for pre-training a foundation model.
- Propose a method to measure the "diversity" of the datasets your pipeline can generate.
- Justify your design choices and their potential impact on a downstream model.
- *Be creative and rigorous.* The quality of your evaluation design is a key part of this exercise.

Report (1-2 pages)

- **Literature:** Briefly position your approach within the context of synthetic data generation, particularly for time-series and causal modeling.
- **Approach:** Describe your thought process and the architectural decisions you made for the generation pipeline. Compare and contrast your two implemented temporal strategies.
- **Challenges:** Outline any conceptual or technical issues you encountered. The evaluation of synthetic data is a known hard problem; discuss your thoughts on this.
- **Potential Improvements:** Suggest ways to further enhance the generator. For example, how would you introduce non-stationarity, event-driven changes, or more complex variable types?

Submission Guidelines

- **Code:** Provide clean, well-commented, and modular code. The project should be self-contained and easy to run in a private github repo.
- **Report:** Provide a clear explanation of your methodology, design decisions, and evaluation framework. Show a solid understanding of the challenges involved in creating high-quality data for pre-training large models.
- **README:** Include clear instructions on how to use your data generator, explaining its parameters, and how to run any evaluation scripts.

Evaluation Criteria

We will evaluate your out-of-the-box idea for generating data that is meaningful to train a tabular model for time series and your capacity to evaluate your generated data without relying on a pre-training of a foundational model.

- **Innovation and Creativity:** Thoughtfulness in designing the temporal adaptation strategies and in devising a meaningful evaluation framework for the synthetic data.
- **Technical Execution:** Correctness and robustness of the data generation pipeline. Overall code quality: clean, well-organized, and reusable.

- **Analytical Thinking:** A logical approach and well-reasoned explanations in the report. The ability to reason about the complex relationship between data generation properties and the downstream utility for pre-training a foundation model.

References

- [Hollmann et al., 2023] Noah Hollmann, Samuel Müller, Katharina Eggensperger, Frank Hutter. (2023). TabPFN: A Transformer That Solves Small Tabular Data Problems in a Second. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- [Qu et al., 2025] Jingang Qu, David Holzmüller, Gaël Varoquaux, Marine Le Morvan. (2025). TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *International Conference on Machine Learning (ICML)*.