Causal inference for domain localization and dimensionality reduction

with examples on multimodal data

Abstract

Using Bayesian causal inference methods to reveal the localization dimensionality reduction seems fruiful. The outcome is high-accuracy forecasting models that work in domains of causality. Machine learning models that work with sequential data (time series, texts, videos) in a sequence-to-sequence format meet the fact that data samples from source and target spaces are not related to each other. An optimal structure of a forecasting model takes into account two aspects of data representation: 1) the domain where the causality holds, and 2) the embedding of optimal dimensionality.

We start with the application of Causal Inference for Convergent cross mapping to elaborate the statistical tests. Then we use these tests to construct sequential models: canonical correlation analysis and transformers for linear and nonlinear cases. This research project continues the topic “Causal Discovery in Observational Time Series”.

Research plan

1. Present the Bayesian causal inference in terms of generative modeling for time series; expand to CCM, CCA, and Attn.
2. Reformulate the control problem as the **|do(X)** intervention for dynamical systems.
3. Prepare the model selection pipeline with domain localisation and dimensionality reduction optimization problems.
4. Conduct a computational experiment on d-variate time series, including text and video.

Datasets

1. Quasiperiodic biomedical time series from IMU for starters
2. Textual documents as time series: Intrinsic plagiarism and obfuscation problems.
3. Causality detection in text narratives and video sequences.
4. Generative causality inference for fMRI images and video sequences. The relationship between fMRI images and video sequences viewed by humans remains complex and is often studied using large transformer models.

Tools

1. Write a PyPL library with basic operations for Bayesian causal inference for time series.
2. Write the computational experiment code for various State Space Models.

3. The alternatives: CG, PMCI, VarLiNGAM, TiMINo, NBCB

References:

1. Clément Yvernes, Emilie Devijver, and Eric Gaussier. Complete Characterization for Adjustment in Summary Causal Graphs of Time Series, PMLR, 2025.
2. Clément Yvernes, Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Identifiability by common backdoor in summary causal graphs of time series. ArXiv, 2025.
3. Lei Zan, Anouar Meynaoui, Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test. Entropy, 2024.
4. Daniil Dorin, Nikita Kiselev, Andrey Grabovoy, and Vadim Strijov. Forecasting fMRI images from video sequences, Health Inf Sci Syst, 2024
5. Kuznetsov M.P., Motrenko A.P., Kuznetsova M.V., and Strijov V.V. Methods for intrinsic plagiarism detection and author diarization // Working Notes of CLEF, 2016/

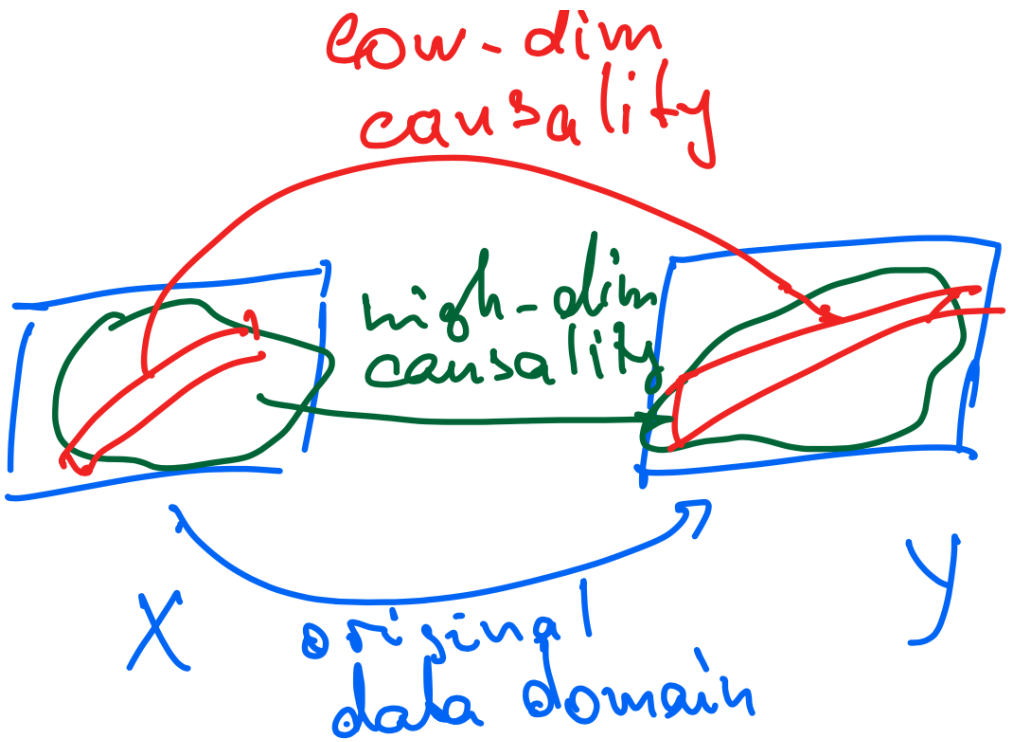

Figure 1: An optimal structure of a forecasting model should consider two aspects of data representation: 1) the domain where the causality holds (green), and 2) the embedding of optimal dimensionality (red).

Appendix: A note from 01/12/2025

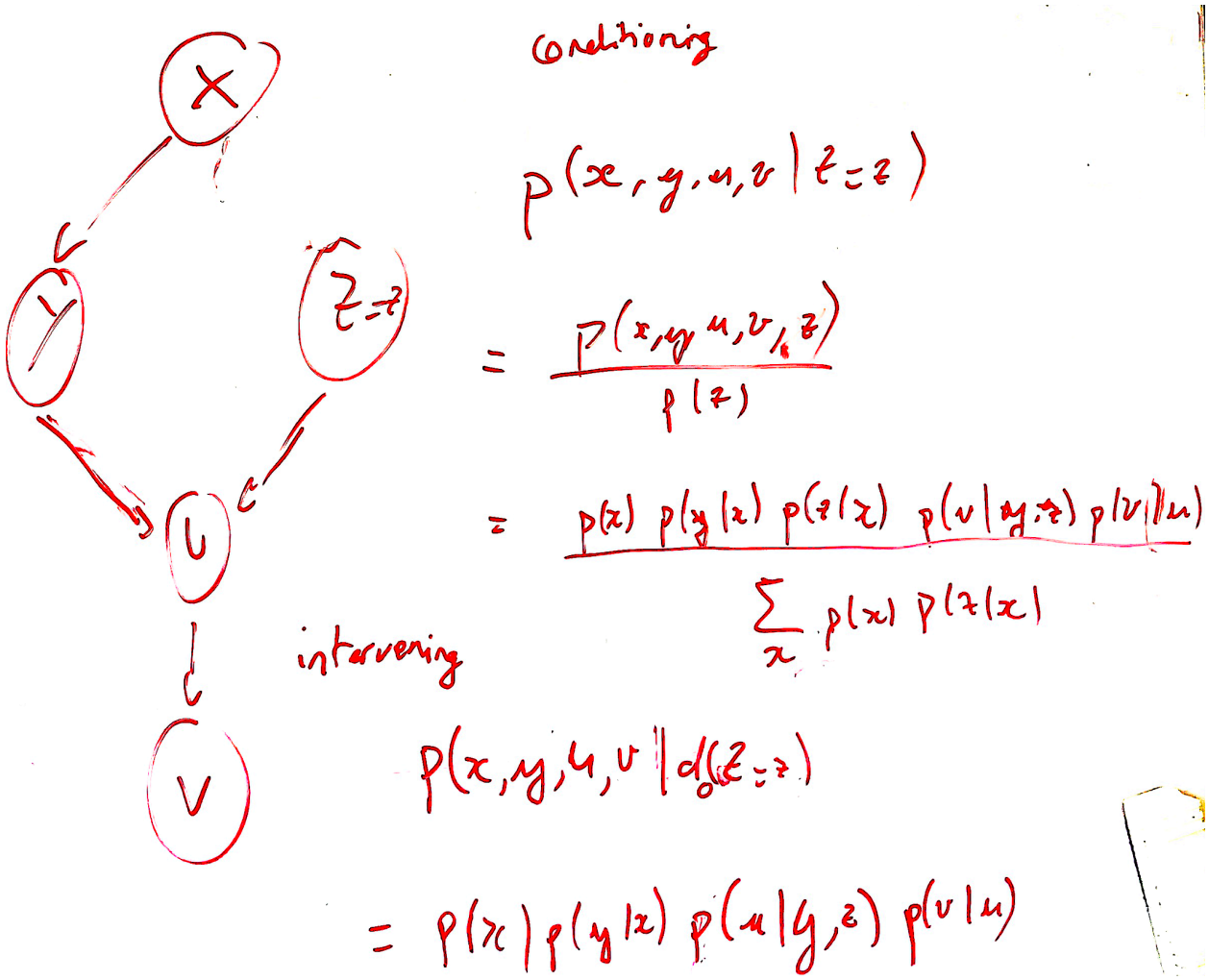

Figure 2: The causal effect of X on V is expressed using the do-operator: P(v,...|do(Z=z)). This represents the distribution of V when we force Z=z, breaking its natural causes. Here, the observation P(V,...|Z=z) represents correlation, and P(V,...|do(Z=z)) represents the causal effect.