Causal inference for text data via convergent cross-mapping

MIAI — IA pour la Science

December 12th, 2025

Causal inference for text data

The goal of research

Propose the Bayesian causal inference in terms of generative modeling for text data.

The focus

Represent the text data as multivariate time series and use Convergent Cross-Mapping for Causal Inference.

1. Reformulate the control problem as the $|\text{do}(X)$ intervention for dynamical systems.
2. Prepare the model selection pipeline with domain localization and dimensionality reduction.
3. Generalise the CCM to Canonical Correlation Analysis and Attention.
4. Conduct a computational experiment on d-variate time series, including text and video.

Convergent Cross Mapping

- We observe a pair of dynamical systems X and Y, whose true behavior is described by the manifolds W_x and W_y in their corresponding phase spaces. We can only infer information about these manifolds from their projections M_x and M_y onto the observed values of the time series $\{x_t\}_{t=1}^{T}$ and $\{y_t\}_{t=1}^{T}$ corresponding to these systems.
- **Cross Mapping**. From the phase trajectory of one system, M_x, we can predict the values of the second system: $y_t \simeq \hat{y}_t | M_x$. Similarly, we can construct predictions $\hat{x}_t | M_y$. The accuracy of these predictions indicates whether a causal relationship exists between the systems.
- **Convergence**. If such a relationship exists, the prediction quality should improve as the considered time interval T increases.

Convergent Cross Mapping Algorithm (1)

1. There given two time series $\{x_1, x_2, ..., x_T\}$ and $\{y_1, y_2, ..., y_T\}$ of length T.
2. For the series $\{x_t\}_{t=1}^{T}$, construct history vectors of dimension E with time lag τ:

$$\mathrm{x}_t = \left(x_t, \quad x_{t-\tau}, \quad x_{t-2\tau}, \quad \ldots, \quad x_{t-(E-1)\tau}\right).$$

3. In the *phase* space $\mathbb{R}^E$, these history vectors form the *phase trajectory* of the system:

$$M_x = \{\mathrm{x}_t \mid t = 1 + (E-1)\tau, ..., T\}.$$

4. One has to build a prediction for the value y_t. To do this, find the $E+1$ vectors from M_x that are closest to x_t (in terms of, for example, the standard metric d in $\mathbb{R}^n$). Sort their time indices $t_1, ..., t_{E+1}$ from the nearest to the farthest points

$$d_i = d(\mathrm{x}_t, \mathrm{x}_{t_i}), \quad d_1 < d_2 < ... < d_{E+1}.$$

Convergent Cross Mapping Algorithm (2)

5. The estimate of the value y_t is then constructed as a weighted sum of the series values at times $t_1, ..., t_{E+1}$:

$$\hat{y}_t | M_x = \sum_{i=1}^{E+1} \omega_i y(t_i),$$

$$\omega_i = \frac{u_i}{\sum_{j=1}^{E+1} u_j}, \qquad u_i = \exp\Big(- \frac{d_i}{d_1} \Big), \qquad i = 1, ..., E+1.$$

6. To assess the existence of dependence between the time series $\{x_t\}_{t=1}^{T}$ and $\{y_t\}_{t=1}^{T}$, compute the Pearson correlation coefficient:

$$C_{yx} = \Big[\rho\Big(y, \hat{y} | M_x \Big) \Big]^2 .$$

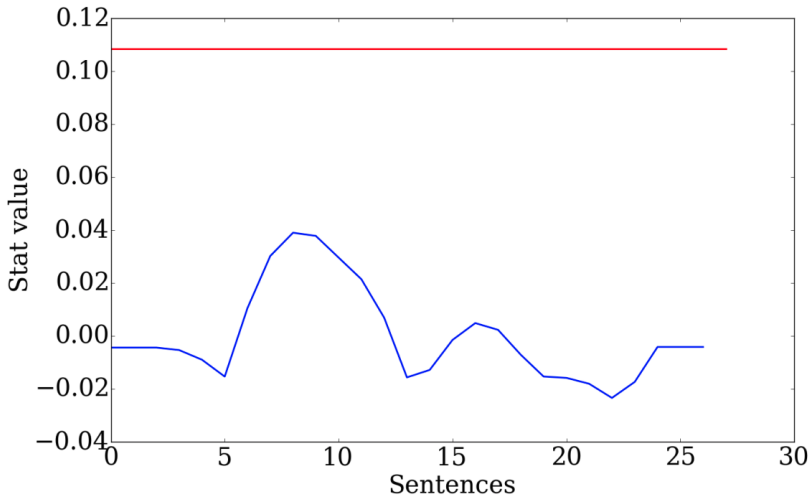
State space reconstruction

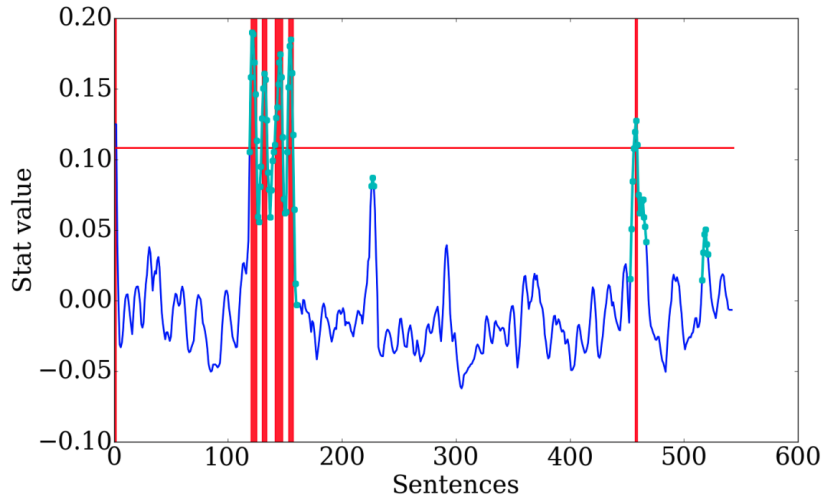
- Takens' theorem use delay vectors to reconstruct the internal structure of a dynamical system.
- When the condition $m \geq 2d + 1$ is satisfied, where d is the embedding dimension, it becomes possible to reconstruct the system's state space.
- In particular, the conclusions of Takens' theorem are used in CCM. The CCM algorithm is analogous to a statistical test and evaluates the causal relationship between two time series.

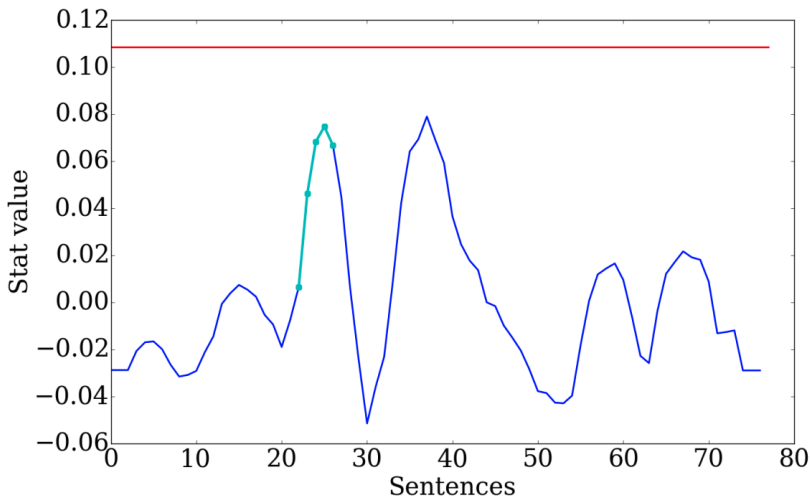
Questions to discuss

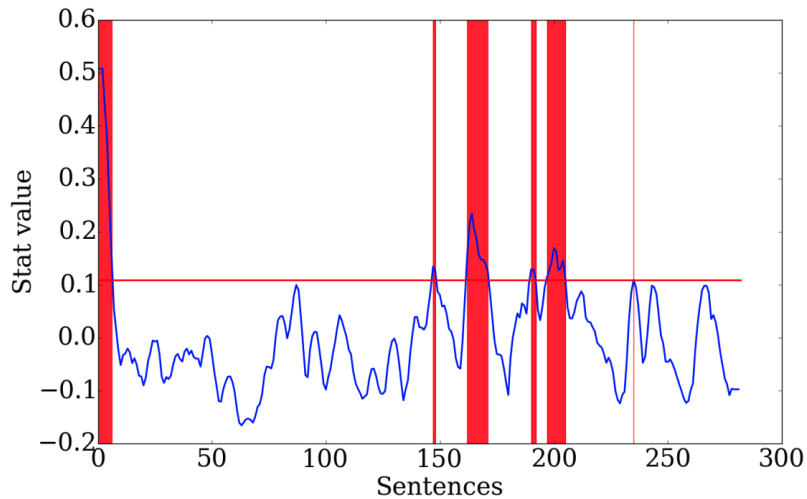
1. List machine learning problems with text data to illustrate cases of Causal Inference.
2. List cases where the text data are represented as time series.
3. List datasets for the experiment
    1) time series to illustrate,
    2) text data to test.

Appendix: text data as time series, illustration (1)



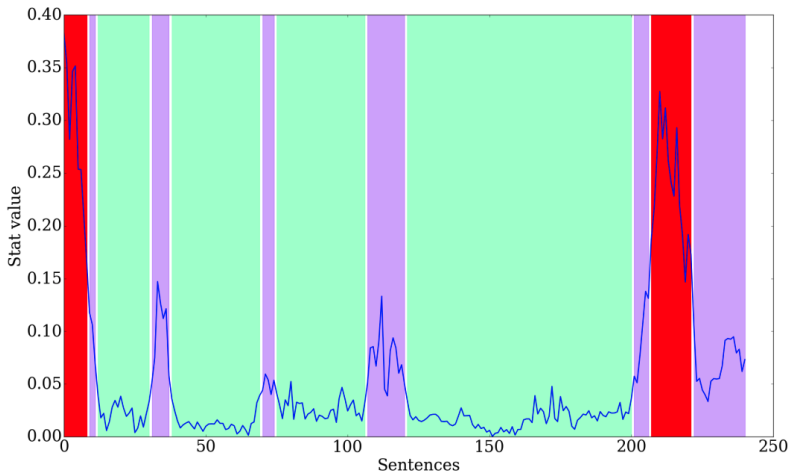


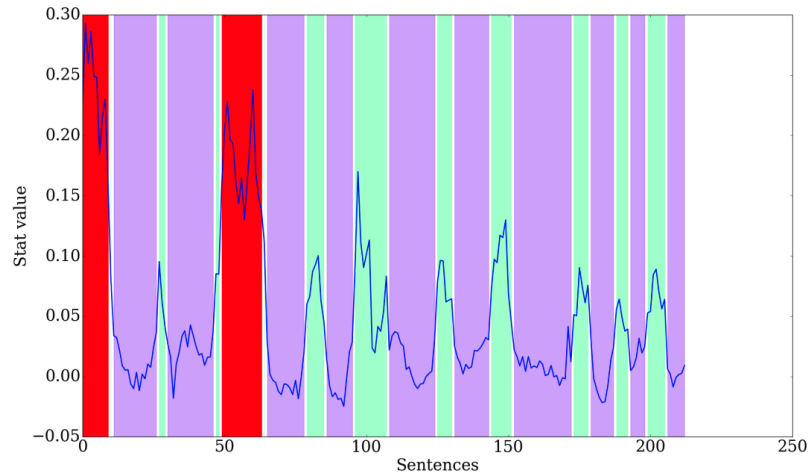




Plagiarism detection examples

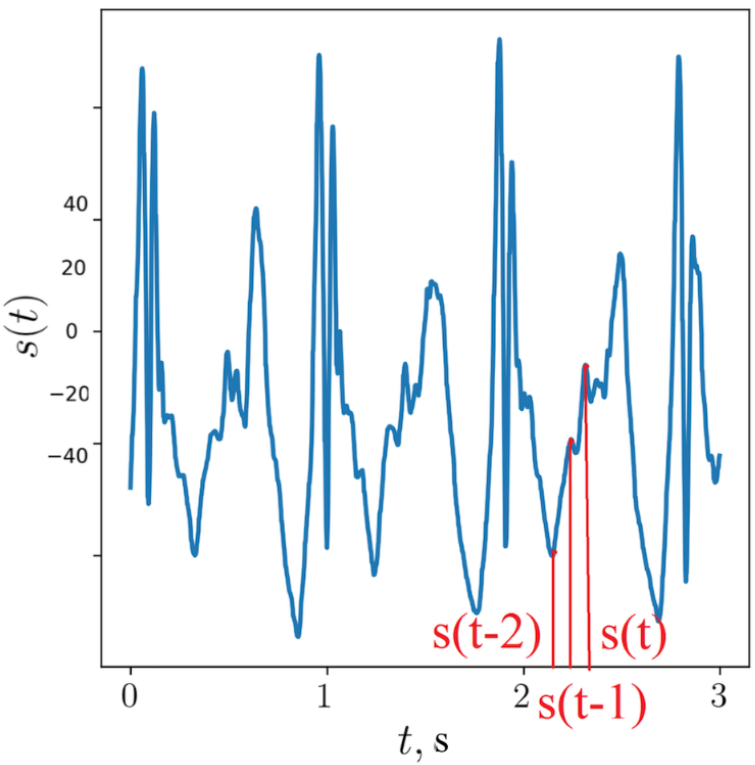
Appendix: text data as time series, illustration (2)

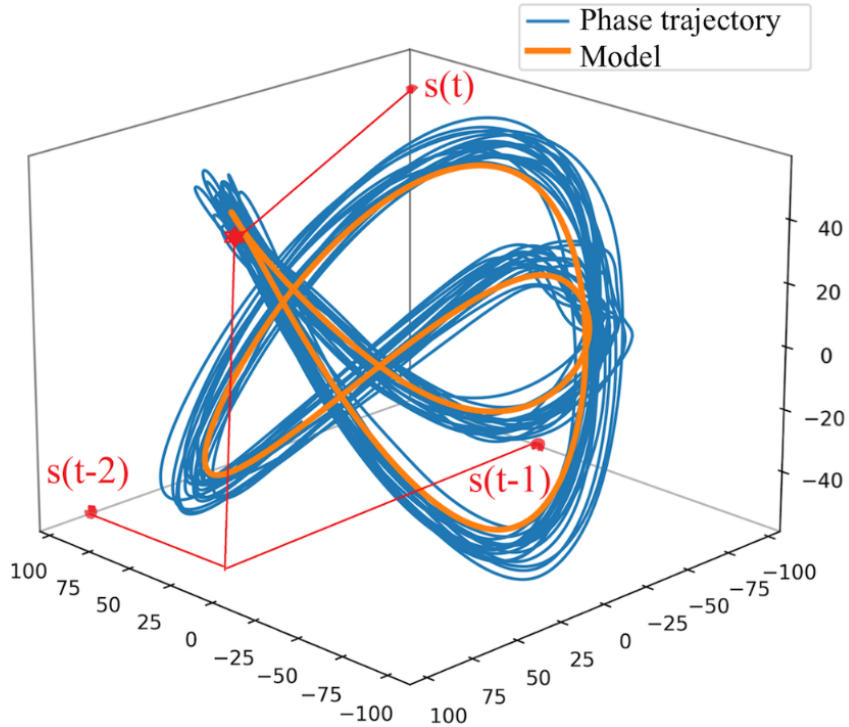




Text segmentation markdown

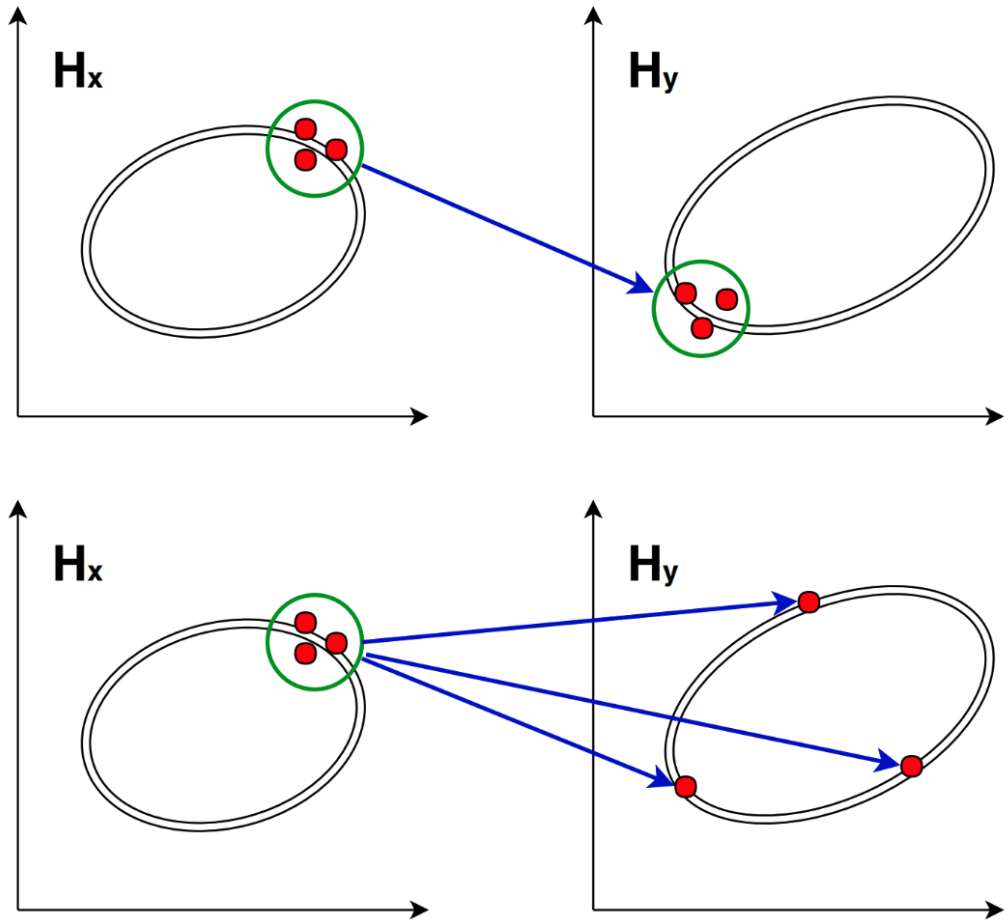
Appendix: phase trajectory of time series





Delay embedding delivers dimensionality reduction and keeps the reconstruction

Appendix: convergent cross-mapping



The time series y depends on the time series x, but not vice versus

The time series y depends on the time series x, if in the neighbourhood $(x, x') \in H_x$ there exists a *Lipschitz continuous map* $\varphi H_x \to H_y$ such that $\rho\big(H_y(\varphi(x, x')\big) \geqslant L\rho H_x(x, x')$.