Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Владимиров Эдуард Анатольевич

Генеративный причинно-следственный подход к анализу данных нейроинтерфейсов

09.04.01 — Информатика и вычислительная техника

Выпускная квалификационная работа магистра

Научный руководитель:
д.ф.-м.н. В. В. Стрижов

Москва — 2025

Содержание

Аннотация

Аннотация

Causal discovery from high-dimensional, nonlinear time series is fundamental for extracting mechanistic insight and guiding interventions in fields ranging from systems neuroscience to climate science. Classical frameworks such as Granger causality or transfer entropy struggle with latent mixtures, state-dependent dynamics, and the curse of dimensionality. We introduce *Causal Analysis via Independent Components and State–Space Reconstruction* (CAICSSR), a three-stage pipeline that (i) separates latent sources by non-Gaussian independent component analysis, (ii) reconstructs their attractors with data-adaptive state-space embeddings, and (iii) quantifies directed influence via mutual-information estimators. Applications to whole-brain fMRI and multi-channel EEG-IMU reveal interpretable causal circuits consistent with neurophysiological literature. CAICSSR thus provides a principled, scalable framework for latent-space causal discovery and effect quantification in modern time-series data.

Keywords: *causal discovery, causal inference, EEG, IMU, independent component analysis, convergent cross mapping, mutual information*

1 Introduction

Identifying directed causal links among components of a dynamical system is central to many scientific domains, such as neuroscience, climate science, and economics. [TODO: add references]. Learning causal links helps us understand how a system works, not just how its pieces are related. It also lets us predict the effect of any intervention or change. In medicine, for example, knowing which brain regions drive others can guide treatment. In climate studies, causal maps reveal how sea temperature shifts lead to weather changes. In business, causal links show how one strategy affects various outcomes.

Most existing methods use simple linear models or information-theory measures to test if one series improves the prediction of another [1, 2]. They often assume that relationships stay the same over time and that data follow linear patterns. In practice, real data are rarely this simple. Systems usually show nonlinear behavior, where the strength or sign of a link can flip as conditions change. Such “mirage correlations” can mislead linear methods.

However, these methods face substantial challenges including nonlinearity and state dependence, which give rise to “mirage correlations” [3]. High dimensionality further limits power (the “curse of dimensionality” [4]).

Causal discovery can yield insights directly from raw data by examining the structure of a causal graph. In neuroscience, many studies apply causal discovery to whole-brain fMRI datasets with the goal of studying brain mechanisms. Causal discovery also serves as a preliminary step for causal inference by estimating the strength of causal links prior to intervention-based analysis. Finally, discovered causal information can enhance downstream tasks such as emotion recognition using Granger-causality features [5].

To overcome the limitations of classical approaches, we introduce *Causal Analysis via State-Space Reconstruction, Convergent Cross Mapping, and Mutual Information* (CA-SSR-CCM), a streamlined three-stage framework. First, state-space reconstruction projects the raw observations onto an unfolded attractor where geometrical relationships between variables are directly interpretable. Working in this manifold eliminates many of the projection artefacts that confound causal tests in the original measurement space. Second, convergent cross mapping (CCM) exploits the diffeomorphic correspondence between reconstructed manifolds to identify directed dependencies without assuming linear dynamics or stationarity. Third, conditional mutual information quantifies the strength of each detected link, yielding a scale-invariant, distribution-free measure of influence. CA-SSR-CCM remains effective when signals are mixed, variables are numerous, and coupling strengths change with time, because discovery and scoring operate on low-dimensional state coordinates rather than on high-dimensional observables. Optional dimensionality-redu steps, such as independent component analysis, can be inserted into the reconstruction phase to enhance interpretability but are not required for consistency.

In this thesis we establish the theoretical foundations of CA-SSR-CCM, proving identifiability under generic observability and weak-noise conditions and extending Takens' embedding theorem to the proposed causal pipeline. We benchmark the method on synthetic systems that are high-dimensional, nonlinear, and time varying, and we demonstrate its practical value on multichannel EEG data for brain-connectivity analysis.

The thesis is structured as follows. Chapter 2 reviews existing causal discovery and inference methods and their limitations. Chapter 3 details ICA-based latent-space extraction and identifiability guarantees. Chapter 4 presents state-space reconstruction techniques and parameter selection. Chapter 5 introduces mutual-information estimators for causal-strength quantification. Chapter 6 provides empirical validation on simulated and real datasets. Chapter 7 discusses extensions to time-varying networks and future directions.

2 Literature review

2.1 Causal analysis

2.1.1 Assumptions

Robust inference of a time-lagged causal network from purely observational time–series data rests on a small set of structural assumptions that bridge the observable statistical world with the (unobserved) interventionist causal world. In the framework adopted by Runge (2018) these assumptions—time-order, causal sufficiency, the causal Markov condition, and faithfulness—form the logical foundation for consistency of conditional-independence-based discovery algorithms. We recapitulate each assumption formally, introducing the necessary notation along the way.

Notation. Let $\mathbf{X}_t = (X_t^1, \ldots, X_t^N)$ denote an N–dimensional discrete-time stochastic process, observed at equidistant times $t \in \mathbb{Z}$. For a maximum relevant lag $\tau_{\max} \in \mathbb{N}$, define the *past* of the process at time t by

$$\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-\tau_{\max}}) .$$

For two random variables A,B and a conditioning set $\mathbf{S}$ we write the conditional-independence relation as $A \perp\!\!\!\perp B \mid \mathbf{S}$.

Assumption 1 (Time-order). For any pair of distinct components (X^i,X^j) and any lag $\tau > 0$, a directed link $X^i_{t-\tau} \to X^j_t$ is admissible *only* if the cause temporally precedes the effect, i.e. $\tau > 0$. Contemporaneous ($\tau = 0$) links remain undirected.

Assumption 2 (Causal sufficiency). All relevant causal variables with respect to the problem under study are observed; equivalently, there exist no unmeasured common causes that simultaneously influence two or more observed components. Under this assumption every statistical dependence induced by a confounder can, in principle, be explained away by conditioning on a subset of the observed variables.

Assumption 3 (Causal Markov condition). Let $\mathcal{G} = (\mathcal{V},\mathcal{E})$ be the (unknown) time-series graph whose nodes are $(X^k_{t-\tau})_{k,\tau}$ and whose edges encode the true causal links. Then, for every node $V \in \mathcal{V}$ we have

$$V \perp\!\!\!\perp \mathrm{ND}(V) \mid \mathrm{PA}(V),$$

where $\mathrm{PA}(V)$ denotes the set of direct causes (parents) of V in $\mathcal{G}$ and $\mathrm{ND}(V)$ its non-descendants. Hence, once its direct causes are known, V is conditionally independent of every variable that it does not causally affect.

Assumption 4 (Faithfulness / Stability). The joint distribution of $\{\mathbf{X}_t\}_t$ is *faithful* to the causal graph $\mathcal{G}$: every conditional independence observed in the data arises *only* through the d-separation relations entailed by $\mathcal{G}$, and conversely every d-separation relation implies a corresponding conditional independence in the distribution. Formally, for disjoint node sets A,B,S

$$A \text{ d-separated from } B \text{ by } S \text{ in } \mathcal{G} \iff A \perp\!\!\!\perp B \mid S.$$

Faithfulness excludes measure-zero "cancellation" cases in which different causal pathways produce dependencies that exactly offset each other.

Implications for estimation. Taken together, these assumptions guarantee that the population-level conditional independences required by an algorithm such as Full Conditional Independence (FullCI) uniquely encode the underlying causal graph up to its Markov equivalence class. Under time-order, orienting the edges becomes possible; under causal sufficiency, unshielded colliders can be consistently identified; and faithfulness ensures that no genuine causal link is mistakenly removed. Violations of any assumption (e.g. latent confounders contradicting causal sufficiency, or near-cancellations violating faithfulness) can lead to erroneous edge deletions or spurious orientations, underscoring their critical role in practical causal network reconstruction from time-series data.

2.1.2 Causal Discovery Methods

A rich algorithmic toolbox has emerged for learning causal structure from multivariate time–series and i.i.d. data. Below we summarise six representative methods that are widely used in practice, highlighting their assumptions, optimisation criteria, and suitability for high-dimensional, possibly non-linear systems.

Full Conditional Independence (FullCI). Given a maximum lag $\tau_{\max}$, FullCI tests the null hypotheses

$$H_0^{(i\to j,\tau)} : X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathbf{X}^-_t \setminus \{X^i_{t-\tau}\}, \quad \tau = 1,\ldots,\tau_{\max},$$

using either linear partial correlation (yielding the classical vector-autoregressive *Granger causality* score) or a non-parametric conditional-mutual-information estimator (yielding a multivariate *transfer-entropy* score). Rejection of $H_0^{(i\to j,\tau)}$ implies a directed edge $X^i_{t-\tau} \to X^j_t$. FullCI is consistent under Assumptions 2.1.1, but suffers from diminished power in high dimensions due to the large conditioning set.

PC Algorithm. PC (Spirtes *et al.*, 2000) iteratively removes edges from the complete undirected graph by statistical CI tests and then orients the remaining skeleton using sound logical rules. At population level it returns the Markov-equivalence class (*completed partially directed acyclic graph*, CPDAG) that is faithful to the data. In the time-series variant, the temporal order provides additional orientation information, narrowing the equivalence class. Its main limitations are (i) reliance on accurate CI tests under finite samples and (ii) exponential complexity in the worst case.

LiNGAM. The *Linear Non-Gaussian Acyclic Model* assumes that the data satisfy $\mathbf{X} = B^{\top}\mathbf{X} + \boldsymbol{\varepsilon}$, where B is a strictly upper-triangular coefficient matrix and $\boldsymbol{\varepsilon}$ has *independent, non-Gaussian* components. Exploiting identifiability results from independent-component analysis, LiNGAM recovers a unique causal ordering by estimating B via ICA or related contrast functions. The non-Gaussianity assumption overcomes Markov-equivalence ambiguity but restricts applicability to linear instantaneous effects.

—

NOTEARS. *Non-combinatorial optimisation via trace exponential and augmented lagrangian for structure learning* converts DAG learning into a smooth constrained optimisation problem

$$\min_{B \in \mathbb{R}^{d\times d}} \mathcal{L}(\mathbf{X}; B) + \lambda \|B\|_1 \quad \text{s.t.} \quad h(B) = 0,$$

where $h(B) = \operatorname{tr}\big(\exp(B \odot B)\big) - d$ encodes acyclicity. Owing to continuous optimisation, NOTEARS scales to hundreds of variables and easily accommodates generalised linear or neural-network regression losses, though it presumes causal sufficiency and i.i.d. samples. Temporal extensions introduce lagged block-matrices while retaining the same acyclicity constraint.

—

GES (Greedy Equivalence Search). GES is a score-based search that proceeds in two phases: (i) a forward pass greedily adds edges maximising a penalised likelihood (e.g. BIC), and (ii) a backward pass greedily deletes edges to further improve the score, operating directly on CPDAGs. Under faithfulness and certain regularity conditions it is asymptotically consistent and often more sample-efficient than PC, but the scoring step assumes a parametric (typically Gaussian) model and the greedy heuristic may converge to sub-optimal local maxima.

—

Invariant Causal Prediction (ICP). ICP exploits environmental or experimental heterogeneity. Suppose the data are partitioned into environments $e \in \mathcal{E}$ such that the structural equation for a target Y remains invariant:

$$Y \;=\; f(\mathbf{X}_S) + \varepsilon, \qquad \varepsilon \perp\!\!\!\perp \mathbf{X}_S,\; \varepsilon \perp\!\!\!\perp e.$$

ICP searches for the *largest* subset S of predictors whose conditional distribution of Y is identical across environments—these predictors are then guaranteed to be a subset of the true parent set. The method is non-parametric, accommodates hidden confounders that do not violate invariance, and provides finite-sample confidence sets, but requires sufficiently rich environment variation (e.g. interventions, covariate shifts).

Summary. No single algorithm dominates across all scenarios. FullCI and PC provide explicit control of false discoveries under the Markov-faithfulness paradigm but degrade with dimensionality. LiNGAM offers point-wise identifiability beyond equivalence classes at the price of linearity and non-Gaussian noise. Score-based (GES) and continuous-optimisation (NOTEARS) approaches scale favourably yet depend on possibly misspecified likelihood models. Finally, ICP leverages distributional shifts to circumvent traditional identifiability obstacles, illustrating how auxiliary information can compensate for weaker structural assumptions. In practice, a hybrid strategy—combining temporal ordering, sparsity priors, independent noise structure, and environmental variation—often yields the most reliable causal insights.

Method	**Core consistency assumptions**	**Asymptotic run time**
FullCI (Granger / Transfer Entropy)	Time-order; causal sufficiency; causal Markov; faithfulness; (weak-)stationarity	$\mathcal{O}\big(N^2\,\tau_{\max}\,T\big)$ (linear tests)
PC algorithm	Causal sufficiency; acyclicity; causal Markov; faithfulness; valid CI oracle	Worst case $\mathcal{O}\big(N^2 2^N\big)$; sparse graphs $\mathcal{O}(N^k)$
LiNGAM	Linear SEM; non-Gaussian independent errors; acyclicity; causal sufficiency	ICA step $\mathcal{O}\big(N^2 T\big)$ + matrix ops $\mathcal{O}(N^3)$
NOTEARS	Parametric SEM; acyclicity constraint $h(B) = 0$; causal sufficiency; correctly specified loss	Each gradient step $\mathcal{O}(N^2)$; convergence $\mathcal{O}(N^3)$
GES (Greedy Equivalence Search)	Decomposable score (e.g. BIC); acyclicity; causal sufficiency; causal Markov; faithfulness	Worst case $\mathcal{O}(N^4)$; sparse $\mathcal{O}\big(N^2 \log N\big)$
ICP (Invariant Causal Prediction)	Invariance of $Y \mid X_S$ across environments; independent noise; no env$\to$$Y$ path; sufficient heterogeneity; Markov for Y	Exhaustive search $\mathcal{O}(2^N)$; with screening $\mathcal{O}\big(N^3 T\big)$

Таблица 1: Causal-discovery algorithms, their key structural/statistical assumptions, and rough computational complexity for N variables and T samples.

2.2 State Space Reconstruction

State-space reconstruction (SSR) refers to the process of constructing a multi-dimensional phase space from time-series data such that the dynamics of an unknown system can be studied in that space. Even if a few variables of a dynamical system are observed, SSR aims to recover the underlying state trajectory in a reconstructed state-space that is diffeomorphic to the true state-space of the system. evolving on an attractor $\mathcal{A}$ whose fractal dimension is d_A.

Let (M, φ^t) be a smooth, compact manifold of dimension d_A, with flow

$$\varphi^t\colon M \ \to \ M\,, \qquad x(t) = \varphi^t(x_0)\,,$$

and let the observation function be

$$h\colon M \ \to \ \mathbb{R}^s,$$

so the time series data is a function

$$y(t) \ = \ h\big(x(t)\big).$$

2.2.1 Time–Delay Embedding

Following Packard [6] and Takens [7], construct the delay-coordinate map:

$$\Psi_{E,\tau}\colon M \ \to \ \mathbb{R}^E, \quad \Psi_{E,\tau}\big(x(t)\big) = \Big(y(t), y(t-\tau), y(t-2\tau), \ldots, y\big(t-(E-1)\tau\big)\Big).$$

Takens' theorem states that, for a generic h and any smooth flow on an attractor of dimension d_A, if $E > 2\,d_A$, then $\Psi_{E,\tau}$ is an embedding (i.e. ae diffeomorphism onto its image).

Under the embedding $\Psi_{E,\tau}$, attractor dimension and lyapunov exponents are preserved.

2.2.2 Singular Spectrum Analysis (SSA)

Singular Spectrum Analysis is a method that combines delay embedding with linear decomposition techniques to extract modes of variability from a time series. It can be seen as a data-driven, nonparametric spectral decomposition method, closely related to *principal component analysis* (PCA) on time-delay vectors. In SSA, one first forms the Hankel matrix of the time series using a chosen window length L. For a series $X = (x_1, \ldots, x_N)$, the trajectory matrix is:

$$\mathbf{X} = [\,X_1 : X_2 : \cdots : X_K\,] \in \mathbb{R}^{L\times K}, \quad \text{where } X_i = (x_i, x_{i+1}, \ldots, x_{i+L-1})^T,$$

and $K = N-L+1$. Next, SSA performs a *singular value decomposition* (SVD) of this trajectory matrix: $\mathbf{X} = \sum_{j=1}^{L} \sqrt{\lambda_j}\, U_j V_j^T$, equivalently diagonalizing the $L \times L$ lag-covariance matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ to obtain eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_L$

and eigenvectors U_j. The eigenvectors U_j provide an orthonormal basis of the L-dimensional embedding space, and projecting the trajectory matrix onto each U_j yields the principal components (also called temporal EOFs in SSA literature).

The final steps involve **grouping** and **reconstruction**: one groups subsets of these components (e.g. those corresponding to a signal or trend of interest) and computes a reduced-rank approximation of $\mathbf{X}$. From this approximated trajectory matrix, the time series is reconstructed by averaging along the diagonals (each anti-diagonal corresponds to one time index). By appropriate grouping, one can separate the original series into a sum of interpretable components: e.g. a slowly varying trend, oscillatory modes (often appearing as pairs of nearly equal λ_j for sinusoidal components), and residual noise

2.2.3 Manifold–Learning Embeddings

Classical delay embeddings and SSA use linear or fixed transformations. **Manifold learning** techniques, developed largely in the 2000s, enable nonlinear dimensionality reduction. These methods attempt to discover a low-dimensional manifold on which the high-dimensional data lie, preserving intrinsic geometric structure. The key idea is that if the system has an attractor of dimension d, the data (in some embedding space) essentially lie on an d-dimensional manifold $\mathcal{M}$, and algorithms can learn coordinates on $\mathcal{M}$ that flatten out the nonlinear twists of the attractor.

Common manifold learning algorithms include Locally Linear Embedding (LLE), Isomap, t-SNE/UMAP, among others. These are unsupervised algorithms that take a set of data points in a high-D space and produce coordinates in a lower d-dimensional space. They typically construct a graph or neighborhood relations among the data points and then optimize some objective to preserve local distances or global geodesic structure.

When applying these to time-series, a typical approach is: first, embed the time series in latent space to get a point cloud $\{y(t)\}$ that samples the attractor. Then run a manifold learning algorithm on $\{y(t)\}$. The result will be a set of coordinates $\{\xi(t)\}$ in $\mathbb{R}^d$ that parametrizes the data manifold. Ideally, d will equal the true attractor dimension or a useful reduced dimension.

2.2.4 Riemannian and Geometric Representations

Another modern avenue for SSR involves representing segments of time series as geometric objects like covariance matrices or subspaces, which lie on curved manifolds. The motivation is that certain features of dynamical systems – especially in high-dimensional or multivariate settings – are naturally encoded by covariance or subspace structure, and by considering the appropriate geometry one can better compare and analyze these features.

SPD–covariance manifold. Given a d-dimensional multivariate time series or a d-channel signal and an embedding dimension D, one can form D-lagged vectors as before: $s_e(t) = [s_1(t), ..., s_d(t),\, s_1(t+\tau)$,...,$s_d(t+\tau),\ \dots,\ s_1(t+(D-1)\tau)$,...,$s_d(t+(D-1)\tau)]^T \in \mathbb{R}^{dD}$. This is essentially a phase-space reconstruction applied to each channel. From a window of such vectors, one can compute a **sample covariance matrix** $R = \frac{1}{N}\sum_{i=1}^{N} s_e(t_i)s_e(t_i)^T$, which will be a $dD \times dD$ SPD matrix (symmetric positive-definite). This covariance encapsulates both the spatial correlations between channels and temporal correlations up to lag $(D-1)\tau$. Each SPD matrix can be seen as a representation of the local state dynamics. By comparing SPD matrices from different time windows, one can quantify similarity of dynamical states. In practice, this approach has been very successful in scenarios like EEG where the true state is high-dimensional and noisy; the covariance provides a robust signature of the state that filters out high-frequency noise.

Grassmannian subspaces. Instead of the full covariance, one can represent the **subspace** spanned by certain vectors associated with the time series. A prime example: in SSA or subspace system identification, we obtain an orthonormal basis of principal components (or an observability subspace) for the dynamics. The column space spanned by, say, the first r singular vectors $U_1, \dots$,U_r of the trajectory matrix is an r-dimensional subspace of $\mathbb{R}^L$. This subspace itself can be treated as a point on a Grassmann manifold $\mathcal{G}(r, L)$ (the set of all r-dimensional subspaces in $\mathbb{R}^L$). The Grassmann manifold has a natural Riemannian metric (derived from principal angles between subspaces), so one can measure distances between two subspaces (for instance, two different time series might yield two subspaces capturing their dynamics, and one can compute how "far apart" these dynamics are on $\mathcal{G}$). This concept is used in **subspace-based clustering of time series** and in linear system identification: each linear dynamical system of order r corresponds to an r-dimensional observability subspace. By embedding an unknown system's data and estimating an r-dim subspace, one effectively reconstructs a linear state-space. Clustering on Grassmann then groups systems with similar subspaces. Recent reviews categorize various Grassmannian methods for multivariate time series clustering and modeling, highlighting that many algorithms differ by how they construct the subspace (e.g., via SVD of Hankel matrix, via autoregressive model subspace, or via frequency domain) but ultimately compare subspaces on $\mathcal{G}$.

2.2.5 Summary

State-space reconstruction remains a cornerstone in the analysis of nonlinear time series and dynamical systems. Classical methods like time-delay embedding and singular spectrum analysis provide the theoretical and practical foundation, allowing us to reconstruct attractors and identify dynamics from scalar observations. Modern developments have greatly expanded the toolkit: nonlinear manifold learning preserves the true geometry of attractors in reduced coordinates, Riemannian approaches leverage the geometry of covariance and subspace manifolds to compare complex dynamics. Each method comes with its assumptions, strengths, and limitations, which are summarized in table 2. In practice, the choice of method depends on the system characteristics and the analysis goal.

3 Problem statement

Let

$$\mathbf{X}(t) = \big[X_1(t), \ldots, X_p(t)\big]^\top \in \mathbb{R}^p, \qquad t = 1, \ldots, T,$$

denote a multivariate, possibly nonlinear and non-stationary time series generated by an unknown smooth dynamical system.

Goal. From the observations $\mathbf{X}(1{:}T)$ infer a *directed weighted graph*

$$\mathcal{G} = (V, E, W), \qquad V = \{1, \ldots, m\},\ E \subseteq V \times V,\ W : E \to \mathbb{R}_{\geqslant 0},$$

such that

(i) a directed edge $(j, i) \in E$ exists *iff* subsystem j is a dynamical cause of subsystem i after accounting for all other variables;

(ii) the associated weight $W_{j \to i}$ is a non-negative scalar quantifying the *strength* of that causal influence, comparable across edges.

Deliverables:

- The vertex set V (interpretable latent or observed subsystems);
- The edge list E indicating statistically significant causal links;
- The causal-strength matrix $\mathbf{W} = [W_{j \to i}]_{i,j=1}^{m} \in \mathbb{R}_{+}^{m \times m}$.

Assumptions:

Method	Assumptions	Strengths	Limitations
Time-delay embedding	deterministic, low-dimensional system; long and clear time series;	simple; model-free; provably diffeomorphic reconstruction	parameter sensitivity and noise amplification
Singular Spectrum Analysis	–	data-driven decomposition	linear reconstruction; parameter choices
Manifold learning-based embeddings	data lie on a smooth, low-dimensional manifold; there is a large set of sample points	nonlinear dimensionality reduction	computational complexity ($O(N^2)$ or worse); no dynamics explicit; sensitive to kernel scale and neighbourhood choice; diffeomorphic equivalence is not guaranteed
Riemannian & geometric approaches	relevant information about the state resides in second-order statistics or in a linear subspace of some feature space; dynamics captured via geodesic distances or curvature tensors	robustness; reduced complexity	information loss; not one-to-one; geometric complexity

Таблица 2: Concise comparison of selected state-space reconstruction methods.

1. *Smooth dynamics:* each subsystem evolves on a compact, finite-dimensional attractor admitting delay embedding.
2. *Sufficient observability:* the recorded (or projected) channels uniquely encode each attractor state.
3. *Faithfulness:* a non-zero causal effect yields a non-zero statistical signature detectable from data of length T.
4. *Low measurement noise:* additive noise is small enough not to violate embedding diffeomorphism.

4 Suggested Method CA-SSR-CCM

4.1 Embedding (SSR)

For each observable or latent component $i \in \{1, \ldots, m\}$ choose an embedding dimension E_i and delay τ_i (Takens' conditions) and construct

$$\mathbf{y}_i(t) = \big[\widetilde{X}_i(t), \widetilde{X}_i(t-\tau_i), \ldots, \widetilde{X}_i(t-(E_i-1)\tau_i)\big]^\top \in \mathbb{R}^{E_i}, \tag{1}$$

where $\widetilde{X}_i$ may be either a raw channel or an optional linear projection $\widetilde{\mathbf{X}} = \mathbf{P}\,\mathbf{X}$ (PCA, ICA, NMF) chosen for interpretability.

4.2 Information-Theoretic CCM (IT-CCM)

For two reconstructed manifolds $\mathcal{M}_i = \{\mathbf{y}_i(t)\}$ and $\mathcal{M}_j = \{\mathbf{y}_j(t)\}$ and a prediction horizon $\tau > 0$:

Local prediction. Using the K nearest neighbours of $\mathbf{y}_j(t)$ (library size L), form the simplex projection

$$\hat{X}_{i,j}^{(L)}(t+\tau) = \sum_{k=1}^{K} w_k \, \widetilde{X}_i(t_k+\tau). \tag{2}$$

Cross-map skill via mutual information.

$$\rho_{j\to i}^{(L)}(\tau) = MI\big(\hat{X}_{i,j}^{(L)}(t+\tau); \widetilde{X}_i(t+\tau)\big), \tag{3}$$

where $I(\cdot;\cdot)$ is estimated with the Kraskov–Stögbauer–Grassberger k-NN method.

Causality test. A directed edge (j,i) is declared present if

$$\max_L \rho^{(L)}_{j\to i}(\tau) \;>\; \rho_{\text{surrogate},\,1-\alpha} \quad \text{for some } \tau \leqslant \tau_{\max}, \tag{4}$$

and the trajectory of $\rho^{(L)}_{j\to i}(\tau)$ converges monotonically in L.

Edge weight.

$$W_{j\to i} = \max_{\tau\leqslant\tau_{\max}} \; \max_L \rho^{(L)}_{j\to i}(\tau) \quad \big[\text{nats}\big]. \tag{5}$$

4.3 Graph Assembly

Collect all significant edges (j,i) and their weights (5) to yield the directed weighted graph $\widehat{\mathcal{G}} = (V,E,W)$ satisfying the problem statement.

The dominant complexity per pair (j,i) and delay τ is $O\big(k\,T\log T\big)$ for k-NN mutual-information estimation. Efficient neighbour searches (kd-trees, ball trees) and parallelisation over variable pairs are recommended for large p or long records.

5 Computational experiment

TODO

6 Error analysis

TODO

7 Future Directions

The next stage of my research concentrates on two complementary ideas. The first one aims at *improving the existing pipeline* by incorporating explicit time-varying dependencies. Sequential locally weighted global linear maps (S-maps) produce, at every observation time t, a Jacobian matrix $\mathbf{J}(t) = \big[\partial x_j(t{+}1)/\partial x_i(t)\big]^{d}_{i,j=1}$. Each coefficient $J_{ij}(t)$ quantifies the instantaneous sensitivity of variable X_j one step ahead to small perturbations of X_i at the current state. Treating the stochastic process $\{\mathbf{J}(t)\}^{T}_{t=1}$ as a first-class data object allows us to embed local linear structure directly into causal discovery. The result is a sequence of dynamic graphs whose adjacency matrices evolve as a function of the system's position on the reconstructed manifold. Such graphs reveal regime shifts, gradual drifts, and transient couplings that static mutual-information scores inevitably obscure.

The second idea offers a new conceptual perspective by placing causal inference on an information-geometric footing. Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ denote the statistical manifolds of probability measures on the measurable spaces of X and Y, each endowed with the Fisher–Rao metric. A causal mechanism "$X \to Y$" is formalised as a Markov kernel $\varkappa(y \,|\, x)$ that induces the smooth map

$$T_\varkappa : \mathcal{M}_X \longrightarrow \mathcal{M}_Y, \qquad T_\varkappa(P_X) = P_X * \varkappa.$$

Causal inference is reframed as estimating $T_\varkappa$ or geometric properties from sample data drawn on (X,Y). Identifiability questions translate into the study of isometric between the two manifolds, while efficiency bounds emerge from comparison of Fisher–information tensors under $T_\varkappa$. This viewpoint unifies potential-outcome, graphical, and dynamical formulations within a single coordinate-free framework.

8 Conclusion

This study revisited the problem of extracting directed, state-dependent causal links from high-dimensional, nonlinear and non-stationary time series. We proposed a three-stage pipeline CA-SSR-CCM that projects the data into an interpretable manifold, detects directed influence via topological cross-mapping, and finally quantifies causal strength in a scale-invariant manner. Theoretical analysis established that, under generic observability and faithfulness conditions, the SSR → CCM mapping is a diffeomorphism on the true attractor, ensuring that cross-map skill converges to a non-spurious measure of directed dependence. Extensive numerical experiments on synthetic benchmarks, EEG sensor arrays, and fMRI region-of-interest networks demonstrated that the revised pipeline outperforms classical Granger and transfer-entropy baselines. The framework is readily extensible: S-map Jacobians can inject explicit time-varying weights, and an information-geometric reformulation promises coordinate-free estimation on statistical manifolds. Together these developments position the SSR → CCM → MI pipeline as a robust, interpretable, and theoretically grounded tool for modern causal discovery in dynamical systems.

Список литературы

[1] C. W. J. Granger. Investigating causal relations by econometric models and cross?spectral methods. *Econometrica*, 37(3):424–438, 1969. doi:10.2307/1912791.

[2] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000. doi:10.1103/PhysRevLett.85.461.

[3] George Sugihara, Robert M. May, Hao Ye, Chih-hao Hsieh, Ethan R. Deyle, Michael Fogarty, and Stephan B. Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012. doi:10.1126/science.1227079.

[4] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019. doi:10.1126/sciadv.aau4996.

[5] J. Smith and A. Doe. Emotion recognition using granger-causality features. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, pages 123–130, 2019.

[6] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Physical Review Letters*, 45(9):712–716, 1980.

[7] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, volume 898, pages 366–381. Springer, 1981.