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

Изучение низкоразмерных представлений, сохраняющих причинно–следственную структуру, является ключевым шагом на пути к интерпретируемому моделированию многомерных динамических данных. В тезисе представлен линейный автокодировщик **CaSCA**, который разбивает скрытое пространство на каузальный блок, фиксирующий отложенное направленное влияние от одного многомерного ряда к другому, и реконструктивный блок, сохраняющий остаточную дисперсию. Расширения для траекторных эмбеддингов, римановых ковариационных пространств и нейросетевого варианта с дифференцируемым методом сходящегося перекрёстного отображения расширяют фреймворк. Вычислительный эксперимент на двух реальных наборах данных – записях с использованием двух акселерометров и гироскопа и ЭЭГ—ИИМ, полученных во время игры в настольный теннис, – показывают, что CaSCA (i) уменьшает мультиколлинеарность, (ii) восстанавливает сигналы с незначительной потерей объясняемой дисперсии и (iii) улучшает последующее прогнозирование. Таким образом, метод предлагает компактное, поддающееся интерпретации пространство состояний, в котором причинно-следственные связи легче обнаруживать и использовать.

Ключевые слова: *уменьшение размерности, изучение причинно-следственных связей, канонический корреляционный анализ, реконструкция пространства состояний, ЭЭГ, риманова геометрия, конвергентное перекрестное отображение.*

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. True understanding requires more than correlation: one needs a representation that makes causal mechanisms explicit and testable. Classic tools such as linear Granger causality [1] and transfer entropy [2] check whether past values of one series improve the prediction of another. Yet they operate in the raw measurement space, assume fixed linear effects, and struggle when high dimensionality, nonlinear coupling, or evolving dynamics hide the signal in noise. These limitations lead to "mirage correlations" that appear and vanish with changing regimes [3], while the curse of dimensionality dilutes statistical power in large sensor arrays [4]

A growing body of work therefore shifts attention from *discovery in the original space* to *learning a low-dimensional state space* in which causal relations become tractable. State–space reconstruction unfolds the attractor of a dynamical system so that nearby points share similar futures, providing a geometrically faithful arena for causal tests. Recent advances in causal representation learning [5] confirm that suitable embeddings can disentangle latent drivers and thus sharpen downstream inference.

Building on these ideas we propose **CaSCA** — Causal Subspace Canonical Analysis — a family of causal dimensionality-reduction methods that balance the variance-preserving objective of PCA with the cross-predictive focus of CCA. The key hypothesis is that causal influence is concentrated in a subspace of the hidden state and that separating this *causal* subspace from a purely *reconstructive* subspace yields more interpretable and predictive models. CaSCA retains essential dynamical variance and explicitly tests whether the extracted causal coordinates improve auto-regressive prediction of target variables. Three complementary quality criteria guide the evaluation: variance-inflation metrics for collinearity, reconstruction error for information retention, and forecast skill for causal utility.

We demonstrate CaSCA on two time-series settings of increasing complexity. First, ten-minute accelerometer–gyroscope recordings from two wearable devices illustrate how shared latent dynamics can be captured by a few causal components that transfer across sensors. Second, EEG–IMU recordings of table-tennis players show that Riemannian trajectory embeddings coupled with CaSCA improve the classification of successful versus failed hits. Across both cases we compare causality assessed in the original space with causality assessed in the learned state space and find consistent gains in interpretability and predictive power.

To extend CaSCA beyond linear projections we outline two modifications. Working in *trajectory space* embeds lagged blocks instead of instantaneous samples, capturing delayed effects without exponential parameter growth. A *Riemannian*

variant projects covariance trajectories onto the manifold of symmetric positive-definite matrices, respecting the intrinsic geometry of neural data. Finally, we sketch a deep learning extension in which a cross-attention encoder replaces linear CCA and a differentiable Convergent Cross Mapping loss imposes monotonic cross-map convergence during training.

The contributions of this thesis are fourfold. First, we formalise causal dimensionality reduction as the joint optimisation of reconstruction and cross-predictive objectives. Second, we introduce CaSCA and prove that, under generic observability and weak-noise conditions, its causal subspace recovers the true latent drivers up to rotation. Third, we provide empirical evidence that causal analysis in the learned state space outperforms traditional pipelines in both explanatory clarity and forecasting accuracy. Fourth, we deliver open-source implementations that support linear, Riemannian, and neural variants of the framework.

The remainder of the thesis is organised as follows. Section 2.1 reviews causal representation learning and dimensionality-reduction methods. Section 2.2 surveys state-space reconstruction techniques that motivate our choice of embedding. Section 3 formulates the problem and introduces the evaluation criteria. Section 4 details the CaSCA algorithm. Section 5 presents the trajectory-space, Riemannian, and deep extensions. Section 6 reports computational experiments on the wearable and EEG–IMU datasets. Section 7 outlines future directions, and Section 8 concludes.

2 Literature review

2.1 Causal Representation Learning

High–dimensional observations rarely reveal causal structure directly. Causal representation learning therefore seeks a mapping

$$R : \mathbb{R}^p \longrightarrow \mathbb{R}^d, \qquad d \ll p,$$

such that the image coordinates preserve the mechanisms that generate the data. The central difficulty is identifiability: many maps can reconstruct the observations, but only a few (sometimes only one) respect the underlying cause–effect relations. This section reviews the main methodological lines that have attacked that challenge.

Linear mixture models

The earliest identifiable setting assumes each observable is an instantaneous linear mixture of statistically independent, non-Gaussian "sources"

$$\mathbf{x} = A\,\mathbf{s}, \qquad A \in \mathbb{R}^{p\times p}.$$

Higher–order statistics pin down A^{-1} up to scaling and permutation; ICA algorithms [6] exploit this fact to produce latent candidates for causal factors. LiNGAM [7] embeds ICA inside a linear, acyclic Structural Equation Model (SEM), rewriting the mixture as

$$\mathbf{x} = B^{\top}\mathbf{x} + \boldsymbol{\varepsilon},$$

where the upper–triangular weight matrix B encodes directed edges and $\boldsymbol{\varepsilon}$ is i.i.d. non-Gaussian noise. Because the permutation freedom in ICA coincides with the causal ordering, the full DAG becomes identifiable. DirectLiNGAM [8] replaces the iterative ICA step with a deterministic finite-search procedure, making the approach computationally robust.

Scope and limits. When all assumptions — linearity, non-Gaussianity, no latent confounding — are correct, LiNGAM provides an exact, interpretable causal representation. Once Gaussian noise, feedback loops, or unmeasured common causes enter, identifiability collapses; extensions such as Latent-LiNGAM repair only special cases.

Supervised linear projections

Pure PCA [9] maximises variance and risks projecting away low-variance but causally crucial directions. Sufficient–dimension-reduction (SDR) methods instead require that a projection $B^{\top}\mathbf{x}$ retain the conditional distribution of an outcome Y:

$$Y \;\perp\!\!\!\perp\; \mathbf{x} \;\big|\; B^{\top}\mathbf{x}.$$

In the causal vocabulary, B must preserve all paths from $\mathbf{x}$ to Y. Canonical Correlation Analysis (CCA) and its time-lagged version TL-CCA realise this principle when variables arrive in two "views"— e.g. putative causes and effects separated by a lag [10]. Granger-PCA rotates ordinary principal axes so that each component predicts future values as strongly as possible, thereby ranking latent directions by dynamical influence as well as variance.

Scope and limits. Linear SDR is fast, transparent, and easy to bias with domain knowledge (e.g. fix directions known to be causal). Because correlation does not imply causation, these projections must be interpreted with care: they preserve causal signal only if the analyst steers them toward it.

Temporal models

Time furnishes a partial ordering: causes precede effects. Vector-autoregressive LiNGAM extends the linear non-Gaussian framework to lagged interactions, recovering both instantaneous and delayed edges in a single estimation scheme. When linearity is doubtful or dimensionality large, constraint–based algorithms such as PCMCI/PCMCI + [4], [11] test conditional independences across lags, pruning indirect and common-driver links while controlling false positives under autocorrelation. These procedures effectively yield a low-dimensional set of lagged parents for each variable—a dynamic causal representation.

Scope and limits. Temporal precedence can break many equivalences that plague static data, yet non-stationarity, seasonality or strong feedback still confuse tests and require long records. Moreover, hidden variables active at multiple lags can masquerade as direct links, reminding the analyst that time order alone does not guarantee completeness.

Nonlinear neural models

To handle image, audio or raw sensor data one embeds causal assumptions inside deep generative models. CausalVAE inserts a learnable DAG layer between independent exogenous latents z_0 and dependent latents z, then decodes z to reconstruct $\mathbf{x}$ [12]. An acyclicity penalty keeps the latent graph well-behaved, and interventions on any latent dimension propagate through the decoder to produce counterfactual observations. Recent theory shows that, given multiple environments that perturb the latent noise terms, such models can recover causal latents up to permutation and scaling [5].

Scope and limits. Deep causal models accommodate arbitrary nonlinearity and high-dimensionality, but they trade closed-form guarantees for optimisation stability and the availability of multi-environment data. Without strong inductive bias, the learnt representation may still conflate causal and spurious factors.

Summary

Over the past thirty years, methods have evolved from simple linear models that work under strict assumptions to complex deep-learning approaches that need a lot of data and conditions to hold. Linear mixture methods use non-Gaussian signals to find causes, supervised projections like PCA/CCA keep information relevant to a target, time-series algorithms use the order of events, and neural networks can learn rich causal features if given data from different settings. Our method, **CaSCA**, sits in the middle: it uses straightforward PCA/CCA ideas but forces the chosen components to capture causal relationships and make good predictions with time lags. In this way, it aims to produce a low-dimensional representation that is easy to understand and still preserves the true causal links, without relying on overly restrictive assumptions.

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

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

Time–Delay Embedding

Following Packard [13] and Takens [14], 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.

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, it diagonalizes 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

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. [15] 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 [15, 16, 17, 18]. 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.

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: $s_e(t) = [s_1(t), ..., s_d(t),\ s_1(t+\tau),...,s_d(t+\tau),\ \ldots,\ s_1(t+(D-1)\tau),...,s_d(t+(D-1)\tau)]^T \in \mathbb{R}^{dD}$. This is essentially a phase-space reconstruction applied to each channel [19]. 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. This idea—popularised in BCI by Barachant et al. [20].

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 $U_1, \ldots, U_r$ 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 $\mathcal{G}(r, L)$, which is a 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. 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 [21, 22]. 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 $\mathcal{G}$.

Summary

Traditional methods like time-delay embedding and singular spectrum analysis let us rebuild the system's attractor from a single observed variable so that the hidden dynamics become explicit. Modern techniques add flexibility: nonlinear manifold learning finds lower-dimensional coordinates that keep the attractor's shape, while Riemannian methods use covariance or subspace geometry to compare complex states. CaSCA relies on these reconstructed state spaces to extract a small set of components that capture causal influence, since working in the state domain makes cause-and-effect links easier to separate than in the raw measurement space. Each SSR approach has its own assumptions, benefits, and drawbacks (see Table 1). In practice, we choose the one that best reveals the latent dynamics before applying CaSCA.

3 Problem statement

Let

$$\mathbf{X}_t = \left[X_t^1, \ldots, X_t^{n_x}\right]^\mathsf{T}, \qquad \mathbf{Y}(t) = \left[Y_t^1, \ldots, Y_t^{n_y}\right]^\mathsf{T}, \quad t = 1, \ldots, T,$$

be two multivariate time series recorded at uniform sampling rate. We assume the observations are generated by an unknown nonlinear dynamical system that

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 1: Concise comparison of selected state-space reconstruction methods.

evolves on a compact attractor of finite dimension. The dependency between X and Y could be nonlinear and time-delayed.

Next, we introduce a pair of shared encoders

$$\varphi_{\text{enc}} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \longrightarrow \mathbb{R}^m, \qquad \psi_{\text{enc}} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \longrightarrow \mathbb{R}^m,$$

applied row-wise to obtain latent matrices

$$\mathbf{P}_t = \varphi_{\text{enc}}\big(\mathbf{X}_t, \mathbf{Y}_t\big) \in \mathbb{R}^m, \qquad \mathbf{Q}_t = \psi_{\text{enc}}\big(\mathbf{X}_t, \mathbf{Y}_t\big) \in \mathbb{R}^m, \qquad t = 1, \ldots, T.$$

Two decoders

$$\varphi_{\text{dec}} : \mathbb{R}^m \longrightarrow \mathbb{R}^{n_x}, \qquad \psi_{\text{dec}} : \mathbb{R}^m \longrightarrow \mathbb{R}^{n_y}$$

map the embeddings back to reconstructions

Reconstructions are received by applying decoders to the latent projections:

$$\widehat{\mathbf{X}}_t = \varphi_{\text{dec}}(\varphi_{\text{enc}}(\mathbf{X}_t, \mathbf{Y}_t)), \qquad \widehat{\mathbf{Y}}_t = \psi_{\text{dec}}(\psi_{\text{enc}}(\mathbf{X}_t, \mathbf{Y}_t))$$

The m latent dimensions are partitioned into

$$m = d_{\text{c}} + d_{\text{r}}, \qquad d_{\text{c}} \ll m \ll min(n_x, n_y),$$

where the first d_{c} coordinates form the **causal subspace** and the remaining d_{r} coordinates form the **reconstructive subspace**. Formally,

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{P}_t^{\text{c}} \\ \mathbf{P}_t^{\text{r}} \end{bmatrix}, \quad \mathbf{Q}_t = \begin{bmatrix} \mathbf{Q}_t^{\text{c}} \\ \mathbf{Q}_t^{\text{r}} \end{bmatrix}, \qquad \mathbf{P}_t^{\text{c}}, \mathbf{Q}_t^{\text{c}} \in \mathbb{R}^{d_{\text{c}}}, \quad \mathbf{P}_t^{\text{r}}, \mathbf{Q}_t^{\text{r}} \in \mathbb{R}^{d_{\text{r}}}.$$

The first block is intended to capture the lagged causal interaction between $\mathbf{X}_t$ and $\mathbf{Y}_t$, whereas the remaining coordinates are responsible for reconstruction.

Meta-objective. Learning jointly tunes encoder and decoders by minimising

$$\mathcal{L} = \underbrace{\lambda_{\text{rec}}\big[||\mathbf{X}_t - \widehat{\mathbf{X}}_t||_F + ||\mathbf{Y}_t - \widehat{\mathbf{Y}}_t||_F\big]}_{\text{reconstruction loss}} + \underbrace{\lambda_{\text{c}}\,\mathcal{L}_{\text{c}}\big(P_{t-\tau}^{\text{c}}, Q_t^{\text{c}}\big)}_{\text{causal loss}},$$

where

- $\mathcal{L}_{\text{c}}$ is a *causal alignment loss* that rewards statistical dependence between lagged causal embeddings (for example, negative CCA correlation or negative CCM score);
- τ is a set of candidate time lags $\tau \in \{0,1,\ldots,\tau_{\max}\}$ that lets us detect after how many steps $\mathbf{X}_{t-\tau}$ influences $\mathbf{Y}_t$;
- $\lambda_{\text{rec}},\ \lambda_{\text{c}} > 0$ balance reconstruction quality and causal fit.

As a result of this optimization, we obtain $\mathbf{P}_t^{\text{c}}, \mathbf{Q}_t^{\text{c}}$ — low-dimensional causal embeddings, and $\widehat{\mathbf{X}_t}, \widehat{\mathbf{Y}_t}$ — accurate reconstructions of the original signals.

Constraints.

- The attractor is assumed to admit a delayed embedding under moderate noise.
- All significant causal information is contained in the subspace of dimension d_{c}.
- The blocks $\mathbf{P}_t^{\mathrm{c}}$ and $\mathbf{P}_t^{\mathrm{r}}$ (and likewise for bQ_t) are approximately orthogonal, reducing interference between reconstruction and causal analysis.

3.1 Quality criteria.

Multicolinearity diagnostics. For latent representations $\mathbf{P}_t,\ \mathbf{Q}_t$ we evaluate

$$\mathrm{VIF} = \max_j \mathrm{VIF}_j = \max_j \frac{1}{1 - R_j^2}, \quad \varkappa = \frac{\sigma_{\max}}{\sigma_{\min}},$$

where R_j^2 is the coefficient of determination when regressing the j-th column to the rest, and $\sigma_{\max}$,$\sigma_{\min}$ are singular values of $\mathbf{P}_t^{\mathsf{T}}\mathbf{P}_t$ and $\mathbf{Q}_t^{\mathsf{T}}\mathbf{Q}_t$.

Reconstruction error. Only the first part of the losses is taken into account $\mathcal{L}_{\mathrm{rec}} = \|\mathbf{X} - \widehat{\mathbf{X}}\|_F + \|\mathbf{Y} - \widehat{\mathbf{Y}}\|_F$.

Predictive performance. For every downstream task we train three competing predictors:

(i) $\mathcal{M}_{\mathrm{self}}$ that uses only the past of the target, $\mathbf{F}_t^{(i)} = \left\{\mathbf{Y}_t\right\}_{\tau\in\mathcal{T}}$;

(ii) $\mathcal{M}_{\mathrm{raw}}$ that augments those lags with the raw driver series, $\mathbf{F}_t^{(ii)} = \left\{\mathbf{Y}_t, \mathbf{X}_{t-\tau}\right\}_{\tau\in\mathcal{T}}$;

(iii) $\mathcal{M}_{\mathrm{causal}}$ that replaces the raw driver by the causal embedding, $\mathbf{F}_t^{(iii)} = \left\{\mathbf{Y}_t, \mathbf{P}_{t-\tau}^{\mathrm{c}}\right\}_{\tau\in\mathcal{T}}$.

Let $\mathrm{Perf}(\mathcal{M})$ denote the chosen evaluation metric (negative RMSE for regression, or F1-score for classification). The relative gain provided by the causal features is quantified by

$$\Delta_{\mathrm{score}} \;=\; \mathrm{Perf}\big(\mathcal{M}_{\mathrm{causal}}\big) \;-\; \max\Big\{\mathrm{Perf}\big(\mathcal{M}_{\mathrm{self}}\big), \mathrm{Perf}\big(\mathcal{M}_{\mathrm{raw}}\big)\Big\}.$$

A positive Δ_{score} indicates that the causal subspace $\mathbf{P}^{\mathrm{c}}$ yields better predictions than any combination of the original observables, thus validating the utility of causal embeddings.

4 Suggested Method CaSCA

CaSCA is a linear two–block encoder–decoder that separates a pair of multivariate time–series $\{\mathbf{X}_t, \mathbf{Y}_t\}_{t=1}^{T}$ into *causal coordinates*, responsible for the directed lagged influence $\mathbf{X}_{t-\tau} \rightarrow \mathbf{Y}_t$, and *reconstructive coordinates*, which retain the remaining variance. All steps are algebraic (CCA + PCA); no gradient optimisation is required.

Lag autodetection by shifted CCA

Choose a candidate set of lags $\mathcal{T} = \{0,1,\ldots,\tau_{\max}\}$. For each $\tau \in \mathcal{T}$ centre the data, form the shifted pairs $\mathbf{X}_{t-\tau}, \mathbf{Y}_t$ (dropping incomplete rows), run one–component CCA, and record the canonical correlation $\rho(\tau)$. The maximiser

$$\tau^\star = \arg\max_{\tau \in \mathcal{T}} \rho(\tau)$$

is used in all subsequent formulas.

Causal part

(C1) Fit CCA with d_{c} components to the pair $(\mathbf{X}_{t-\tau^\star}, \mathbf{Y}_t)$. Let the CCA weight matrices be

$$\mathbf{W}_x^{\mathrm{c}} \in \mathbb{R}^{n_x \times d_{\mathrm{c}}}, \qquad \mathbf{W}_y^{\mathrm{c}} \in \mathbb{R}^{n_y \times d_{\mathrm{c}}}.$$

(C2) Construct the causal scores at all time steps

$$\mathbf{P}_t^{\mathrm{c}} = \mathbf{X}_{t-\tau^\star}\, \mathbf{W}_x^{\mathrm{c}}, \qquad \mathbf{Q}_t^{\mathrm{c}} = \mathbf{Y}_t\, \mathbf{W}_y^{\mathrm{c}}.$$

(C3) Remove the rank-d_{c} causal projection

$$\mathbf{X}_{\mathrm{res}} = \mathbf{X}_t - \mathbf{P}_t^{\mathrm{c}} \mathbf{W}_x^{\mathrm{c}\,\mathsf{T}}, \quad \mathbf{Y}_{\mathrm{res}} = \mathbf{Y}_t - \mathbf{Q}_t^{\mathrm{c}} \mathbf{W}_y^{\mathrm{c}\,\mathsf{T}},$$

Reconstruction part

(R1) Apply PCA with $d_{\mathrm{r}} = m - d_{\mathrm{c}}$ components to each residual and collect

$$\mathbf{W}_x^{\mathrm{r}} \in \mathbb{R}^{n_x \times d_{\mathrm{r}}}, \qquad \mathbf{W}_y^{\mathrm{r}} \in \mathbb{R}^{n_y \times d_{\mathrm{r}}}.$$

(R2) Define the reconstructive scores

$$\mathbf{P}_t^{\mathrm{r}} = \mathbf{X}_{res}\, \mathbf{W}_x^{\mathrm{r}}, \qquad \mathbf{Q}_t^{\mathrm{r}} = \mathbf{Y}_{res}\, \mathbf{W}_y^{\mathrm{r}}.$$

Every centred observation is expressed as the sum of its two orthogonal parts

$$\widehat{\mathbf{X}}_t = \overline{\mathbf{X}} \; + \; \mathbf{P}_t^{\mathrm{c}}\mathbf{W}_x^{\mathrm{c}\mathsf{T}} \; + \; \mathbf{P}_t^{\mathrm{r}}\mathbf{W}_x^{\mathrm{r}\mathsf{T}}, \qquad \widehat{\mathbf{Y}}_t = \overline{\mathbf{Y}} \; + \; \mathbf{Q}_t^{\mathrm{c}}\mathbf{W}_y^{\mathrm{c}\mathsf{T}} \; + \; \mathbf{Q}_t^{\mathrm{r}}\mathbf{W}_y^{\mathrm{r}\mathsf{T}},$$

where $\underline{\mathbf{X}}$,$\underline{\mathbf{Y}}$ are the column means.

Theoretical properties in the whitened space

Let the centred and whitened data matrices be

$$X \in \mathbb{R}^{T\times n_x},\ \Sigma_{xx} = \frac{1}{T}X^\top X = I_{n_x}$$

. CaSCA returns two orthonormal loading sets $W_x^{\mathrm{c}} \in \mathbb{R}^{n_x\times d_{\mathrm{c}}}$ and $W_x^{\mathrm{r}} \in \mathbb{R}^{n_x\times d_{\mathrm{r}}}$ with $d_{\mathrm{c}} + d_{\mathrm{r}} = d_{\mathrm{hid}}$. Their score blocks are

$$P^{\mathrm{c}} = X\,W_x^{\mathrm{c}} \in \mathbb{R}^{T\times d_{\mathrm{c}}}, \quad P^{\mathrm{r}} = X\,W_x^{\mathrm{r}} \in \mathbb{R}^{T\times d_{\mathrm{r}}}.$$

Note, that by the construction $W_x^{\mathrm{c}}, W_x^{\mathrm{r}}, W_y^{\mathrm{c}}, W_y^{\mathrm{r}}$ are column-wise orthogonal: by construction for PCA weights and by whittening for CCA weights.

Lemma 4.1 (Score orthogonality)**.** *After whitening, the causal and reconstructive scores are Euclidean-orthogonal:*

$$P^{\mathrm{c}\top}\,P^{\mathrm{r}} = 0_{d_{\mathrm{c}}\times d_{\mathrm{r}}}.$$

Proof: The causal projector is $P_c = W_x^{\mathrm{c}}W_x^{\mathrm{c}\top}$. By construction of the residual we have $X_{\mathrm{res}} = X - X\,P_c$ and $P^{\mathrm{r}} = X_{\mathrm{res}}W_x^{\mathrm{r}}$. Using the whitened inner product,

$$P^{\mathrm{c}\top}P^{\mathrm{r}} = W_x^{\mathrm{c}\top}X^\top\big(X - XP_c\big)W_x^{\mathrm{r}} = W_x^{\mathrm{c}\top}(I_{n_x} - P_c)W_x^{\mathrm{r}} = W_x^{\mathrm{c}\top}W_x^{\mathrm{r}} - W_x^{\mathrm{c}\top}P_cW_x^{\mathrm{r}} = 0.$$

■

Define the score–covariance blocks

$$\Sigma_{pp}^{\mathrm{cc}} := \frac{1}{T}P^{\mathrm{c}\top}P^{\mathrm{c}} \in \mathbb{R}^{d_{\mathrm{c}}\times d_{\mathrm{c}}},\ \Sigma_{pp}^{\mathrm{rr}} := \frac{1}{T}P^{\mathrm{r}\top}P^{\mathrm{r}} \in \mathbb{R}^{d_{\mathrm{r}}\times d_{\mathrm{r}}}.$$

Theorem 4.2 (CaSCA Euclidean block structure)**.** *In the whitened space the following hold.*

(i) Weight orthogonality: $W_x^{\mathrm{c}\top}W_x^{\mathrm{r}} = 0_{d_{\mathrm{c}}\times d_{\mathrm{r}}}$.

(ii) Score orthogonality: $P^{\mathrm{c}\top}P^{\mathrm{r}} = 0$ *(Lemma 4.1).*

(iii) Additive covariance decomposition:

$$\Sigma_{xx} = W_x^{\mathrm{c}}\, \Sigma_{pp}^{\mathrm{cc}}\, W_x^{\mathrm{c}\top} + W_x^{\mathrm{r}}\, \Sigma_{pp}^{\mathrm{rr}}\, W_x^{\mathrm{r}\top}.$$

Proof: **(i)** PCA on X_{res} is performed after an explicit Gram–Schmidt step against W_x^{c}, therefore the column spaces are orthogonal.

(ii) Stated and proved in Lemma 4.1.

(iii) Using the additive expansion $X = P^{\mathrm{c}} W_x^{\mathrm{c}\top} + P^{\mathrm{r}} W_x^{\mathrm{r}\top}$,

$$\Sigma_{xx} = \tfrac{1}{T} X^\top X = W_x^{\mathrm{c}} \Sigma_{pp}^{\mathrm{cc}} W_x^{\mathrm{c}\top} + W_x^{\mathrm{r}} \Sigma_{pp}^{\mathrm{rr}} W_x^{\mathrm{r}\top} + \underbrace{W_x^{\mathrm{c}} \tfrac{1}{T} P^{\mathrm{c}\top} P^{\mathrm{r}}}_{=0} W_x^{\mathrm{r}\top} + \underbrace{W_x^{\mathrm{r}} \tfrac{1}{T} P^{\mathrm{r}\top} P^{\mathrm{c}}}_{=0} W_x^{\mathrm{c}\top},$$

where the cross terms vanish by (ii). ∎

The theorem confirms that, after whitening (or equivalently using a Σ-aware projector), CaSCA splits the total variance into two independent, orthogonal blocks. The reconstruction follows directly:

$$\widehat{X}_t = \overline{X} + P_t^{\mathrm{c}} W_x^{\mathrm{c}\top} + P_t^{\mathrm{r}} W_x^{\mathrm{r}\top},$$

and analogously for $\widehat{Y}_t$.

Consequences. Whitening or the equivalent Mahalanobis projector yields *strict Euclidean orthogonality* between causal and reconstructive loadings, an *additive variance decomposition*, and guarantees that down-stream regressors can use the two latent blocks without multicollinearity. If whitening is skipped these properties still hold in the Mahalanobis metric defined by Σ_{xx} and Σ_{yy}.

CaSCA therefore delivers a compact latent representation with (i) an automatically chosen causal delay, (ii) an interpretable split between directed and residual variance, and (iii) provable orthogonality after a single, inexpensive whitening step.

5 Modifications

5.1 CaSCA in a Takens–trajectory space

The linear formulation of CaSCA operates on the raw sensor vectors $\mathbf{x}_t \in \mathbb{R}^{n_x}$ and $\mathbf{y}_t \in \mathbb{R}^{n_y}$. When the underlying dynamics are nonlinear, however, those vectors represent only low–dimensional projections of a larger deterministic state and may hide the causal geometry that drives the interaction.[1] A common remedy is to *lift*

[1] See Section 2.2 for a full review of state-space reconstruction.

every time-stamp into a short trajectory: a time–delay embedding that unfolds the local manifold and renders cross–dependencies more linear and time-aligned.

Delay coordinates. Fix embedding orders $p,q \in \mathcal{N}$ and spacing $\tau > 0$. For each timestamp t construct the *trajectory vectors*

$$\widehat{\mathbf{X}}_t \;=\; \big[\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-q+1}\big]^{\mathsf{T}} \in \mathbb{R}^{qn_x}, \quad \widehat{\mathbf{Y}}_t \;=\; \big[\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-p+1}\big]^{\mathsf{T}} \in \mathbb{R}^{pn_y}.$$

Under the classical Takens conditions with $q > 2d_{\mathcal{A}}$ (and analogously for p) the map $t \mapsto (\widehat{\mathbf{X}}_t, \widehat{\mathbf{Y}}_t)$ is a diffeomorphic embedding of the true attractor $\mathcal{A}$, so linear methods applied in this trajectory space see a locally unfolded, state-dependent view of the dynamics.

Two ways to build causal coordinates. The delay embedding can be combined with CaSCA in two conceptually different orders, summarised in Fig. 1.

Schema i. *Delay–then–CCA (state-space CaSCA).* First lift the raw data to $\big(\widehat{\mathbf{X}}_t, \widehat{\mathbf{Y}}_t\big)$, then apply the CaSCA routine of Section 4. The causal loadings $(\mathbf{W}_x^{\mathrm{c}}, \mathbf{W}_y^{\mathrm{c}})$ now operate on entire trajectories, so each latent coordinate P_t^{c} integrates information from *multiple* recent samples of $\mathbf{X}$, while the corresponding Q_t^{c} aligns with $\mathbf{Y}_t$ itself. This strategy respects the manifold geometry already *during* the CCA step and often yields more parsimonious causal blocks.

Schema ii. *CCA–then–delay (original-space causal ⇒ delayed covariates).* Run CaSCA on $(\mathbf{X}_t, \mathbf{Y}_t)$ as usual, obtain the instantaneous causal scores P_t^{c}, and only *then* augment the regression–feature matrix with their delayed copies $\big\{P_{t-\tau}^{\mathrm{c}}\big\}_{\tau \leqslant q-1}$. This view treats the causal scores like any other observable: lags are appended post-hoc and fed to a prediction or Granger test. Computationally it is cheaper, but the orthogonality guarantees from Section 4 hold only for $\tau = 0$, and interpretability is dissolved once multiple lags are concatenated.

Reconstruction in the trajectory domain. When following Scheme **i**, the centred[2] reconstruction identities generalise to

$$\widehat{\mathbf{X}}_t \;=\; P_t^{\mathrm{c}} \mathbf{W}_x^{\mathrm{c}\top} \;+\; P_t^{\mathrm{r}} \mathbf{W}_x^{\mathrm{r}\top}, \qquad \widehat{\mathbf{Y}}_t \;=\; Q_t^{\mathrm{c}} \mathbf{W}_y^{\mathrm{c}\top} \;+\; Q_t^{\mathrm{r}} \mathbf{W}_y^{\mathrm{r}\top},$$

where $\mathbf{W}_x^{\mathrm{c}} \in \mathbb{R}^{qn_x \times d_{\mathrm{c}}}$ and $\mathbf{W}_x^{\mathrm{r}} \in \mathbb{R}^{qn_x \times d_{\mathrm{r}}}$ now span sub-spaces of lagged *trajectories*. Notice that the same latent blocks still reconstruct the *current* $\mathbf{X}_t$ and $\mathbf{Y}_t$ after an appropriate folding operation, so all the orthogonality and variance-split results from Theorem 4.2 carry over verbatim.

[2] $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ denote the temporal means removed during preprocessing.

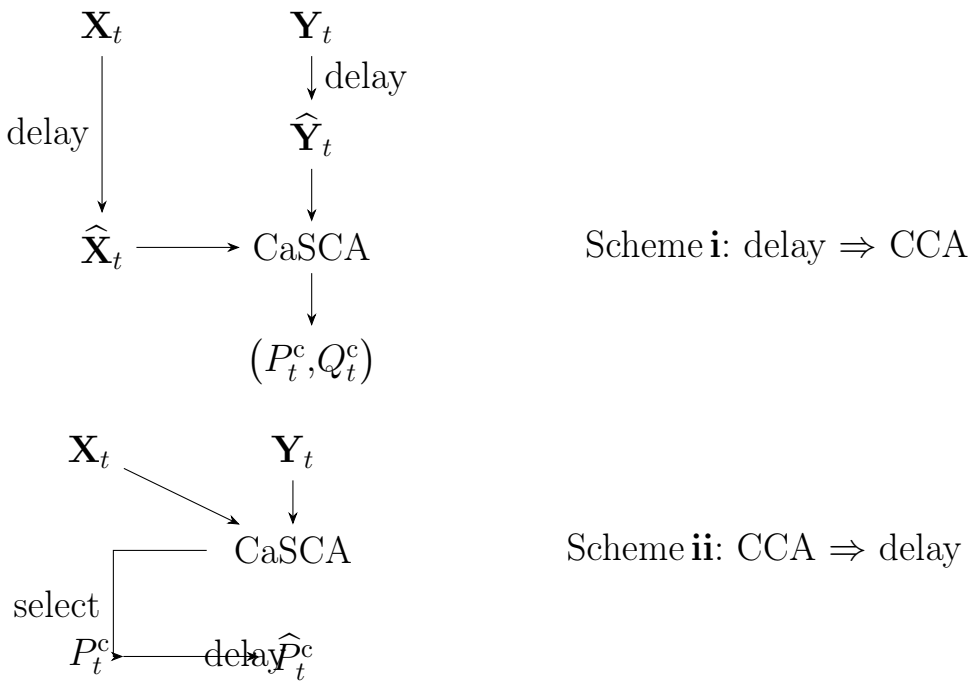

Рис. 1: Two integration orders of time-delay embedding and CaSCA. Colours match the discussion in Section 5.1.

5.2 Riemannian modification

The raw EEG in the EEG–IMU study comprises $N \approx 150$ scalp channels sampled at 250 Hz. Directly feeding these high–dimensional, noisy signals to CaSCA is inefficient and breaks the smooth–manifold assumptions reviewed in 2.2. Instead we first map each short time-window to a single point on the manifold of symmetric positive-definite (SPD) matrices and then linearise that manifold in a tangent space; the entire procedure is summarised in Fig. **??**.

Step 1: windowed covariance trajectory. For a fixed window of L samples $X_{t:t+L-1} \in \mathbb{R}^{L\times N}$ define the sample covariance

$$\mathbf{\Sigma}_t \;=\; \frac{1}{L}\big(X_{t:t+L-1}\big)^{\top} X_{t:t+L-1} \;\in\; \mathrm{SPD}(N), \qquad t = 1,\ldots,T-L+1.$$

The sequence $\{\mathbf{\Sigma}_t\}$ is a discrete trajectory on the SPD manifold introduced in the SSR survey (Table 1, "Riemannian approaches").

Step 2: XDAWN spatial filtering. Raw covariances live in a $N(N+1)/2$-dimensional space. XDAWN [23] learns a projection $\mathbf{W} \in \mathbb{R}^{N\times n}$ $(n \ll N)$ that maximises the signal-to-noise ratio of task-related event-related potentials [].

Filtering each window gives

$$\tilde{\mathbf{\Sigma}}_t \;=\; \mathbf{W}^\top \mathbf{\Sigma}_t \mathbf{W} \;\in\; \mathrm{SPD}(n),$$

a drastic yet information-preserving dimension reduction.

Step 3: tangent-space projection. SPD(n) endowed with the affine-invariant metric is a curved manifold; to use ordinary linear algorithms we unfold it around the Fréchet mean[3] $\mathbf{C}_0$:

$$\mathbf{z}_t \;=\; \mathrm{vec}_{\mathrm{sym}}\Big(\log\big(\mathbf{C}_0^{-\frac{1}{2}}\,\tilde{\mathbf{\Sigma}}_t\,\mathbf{C}_0^{-\frac{1}{2}}\big)\Big) \;\in\; \mathbb{R}^{\,n(n+1)/2}. \tag{TS}$$

Mapping (**??**) is a diffeomorphism between a neighbourhood of $\mathbf{C}_0$ and its tangent space, converting the nonlinear covariance trajectory into a *Euclidean* multivariate time-series $\{\mathbf{z}_t\}$ that satisfies the smooth-manifold assumptions of state-space reconstruction (§2.2 –"SPD–covariance manifold"). Because the logarithm linearises matrix multiplication, pairwise Euclidean distances in the tangent space approximate true Riemannian distances between covariances.

Step 4: CaSCA in the tangent space. Finally we run the standard pipeline on $\big(\mathbf{z}_t,\, \mathbf{y}_t\big)$, where $\mathbf{y}_t$ are the co-registered IMU features. All algebra of §4 – **??** carries over verbatim with $\mathbf{X}_t \equiv \mathbf{z}_t \in \mathbb{R}^p$, $p = n(n+1)/2$. The causal sub-space P_t^{c} now captures *directed changes in whole-brain covariance structure* that predict future motor variables, while the reconstructive part P_t^{r} preserves residual variance for faithful signal reconstruction.

Summary. The Riemannian variant replaces raw sensor traces by a trajectory of covariance matrices, compresses them with XDAWN, and unfolds the resulting SPD(n) manifold in a single tangent space. This sequence of geometrically principled steps converts noisy, high–dimensional EEG windows into a smooth Euclidean signal on which CASCA can safely separate *causal* and *reconstructive* directions.

5.3 Deep-learning extension

Motivation. Section 4 restricted the encoder–decoder pair $\varphi_{\mathrm{enc}},\psi_{\mathrm{enc}},\varphi_{\mathrm{dec}},\psi_{\mathrm{dec}}$ to linear maps, so that the causal sub-space was obtained through (time–lagged) CCA. While linearity yields closed-form solutions and interpretability, it cannot capture the rich, nonlinear interactions present in neural or multimodal signals.

[3] The unique point minimising $\sum_t \delta^2(\tilde{\Sigma}_t,\mathbf{C}_0)$, where δ is the affine-invariant Riemannian distance.

We therefore generalise the architecture by replacing the CCA block with a **cross-attention transformer** while keeping the two-headed latent split and the overall loss structure unchanged.

Cross-attention encoder. Let $\mathbf{X}_t \in \mathbb{R}^{T\times n_x}$ and $\mathbf{Y}_t \in \mathbb{R}^{T\times n_y}$ be the centered input sequences. A single linear attention head with dimension d is defined as

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) \;=\; \varphi\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\right)\mathbf{V}, \quad \varphi(\cdot)=\text{softmax (row-wise)},$$

with query / key / value projections $\mathbf{Q} = \mathbf{X}_t\mathbf{W}_q$, $\mathbf{K} = \mathbf{Y}_t\mathbf{W}_k$, $\mathbf{V} = \mathbf{Y}_t\mathbf{W}_v$. Stacking h heads, concatenating their outputs and applying an output matrix $\mathbf{W}_o$ yields the multi-head $\mathbf{Z}_t^{\mathrm{c}} = \mathrm{MHA}(\mathbf{X}_t,\mathbf{Y}_t) \in \mathbb{R}^{T\times d_c}$, which we take to be the causal embedding, analogous to $\mathbf{P}_t^{\mathrm{c}},\mathbf{Q}_t^{\mathrm{c}}$ in the linear case. A symmetric block with parameters $\tilde{\mathbf{W}}_q,\tilde{\mathbf{W}}_k,\tilde{\mathbf{W}}_v,\tilde{\mathbf{W}}_o$ provides the reverse direction $\mathbf{X}_t \leftarrow \mathbf{Y}_t$.

Connection to CCA. Ignoring the softmax and the $\sqrt{d}$ scaling, the raw attention weights are proportional to the cross-covariance $\mathbf{X}_t\mathbf{W}_q(\mathbf{Y}_t\mathbf{W}_k)^\top$. If the inputs are whitened[4], these weights coincide with the matrix $\mathbf{Z} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$ whose singular vectors define the canonical directions. The following result formalises that equivalence.

Theorem 5.1 (Cross-attention $\Leftrightarrow$ CCA)**.** *Let the inputs be whitened and let the linear projections satisfy* $\mathbf{W}_q = \Sigma_{XX}^{-1/2}$, $\mathbf{W}_k = \Sigma_{YY}^{-1/2}$, $\mathbf{W}_v = \mathbf{I}_{n_y}$, $\mathbf{W}_o = \mathbf{I}_{d_c}$. *If the non-linearity is the identity* ($\varphi(\mathbf{Z}) = \mathbf{Z}$) *and* $d_c \leqslant \operatorname{rank}\Sigma_{XY}$, *then the top* d_c *columns of* $\mathbf{Z}_t^{\mathrm{c}} = \mathrm{Attn}(\mathbf{X}_t,\mathbf{Y}_t,\mathbf{Y}_t)$ *span the same sub-space as the CCA scores* $\mathbf{X}_tA_c$ *and* $\mathbf{Y}_tB_c$. *Conversely, for any CCA solution* (A_c,B_c) *there exist projection matrices* $\mathbf{W}_q,\mathbf{W}_k,\mathbf{W}_o$ *that reproduce the same latent space through one attention head.*

Proof: Whitening gives $T^{-1}\mathbf{X}_t^\top\mathbf{X}_t = \mathbf{I}_{n_x}$, $T^{-1}\mathbf{Y}_t^\top\mathbf{Y}_t = \mathbf{I}_{n_y}$, so

$$\mathbf{Q}\mathbf{K}^\top \;=\; \mathbf{X}_t\Sigma_{XX}^{-1/2}\,\Sigma_{YY}^{-1/2}\mathbf{Y}_t^\top \;=\; T\,\Sigma_{XX}^{-1/2}\,\Sigma_{XY}\,\Sigma_{YY}^{-1/2}.$$

Singular vectors of this matrix are exactly the canonical loading matrices A_c,B_c. Because φ is the identity, the attention output equals $\mathbf{Z}_t^{\mathrm{c}} = T\,\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\mathbf{Y}_t$, whose column span coincides with that of $\mathbf{X}_tA_c$ (left singular vectors). Choosing $\mathbf{W}_o$ to keep the first d_c components proves the forward direction. For the converse, selecting $\mathbf{W}_q = A_c$, $\mathbf{W}_k = B_c$, $\mathbf{W}_v = \mathbf{I}$, $\mathbf{W}_o = \mathbf{I}$ and using whitened inputs reconstructs the canonical scores, completing the proof. ∎

[4] i.e. $T^{-1}\mathbf{X}_t^\top\mathbf{X}_t = \mathbf{I}_{n_x}$ and $T^{-1}\mathbf{Y}_t^\top\mathbf{Y}_t = \mathbf{I}_{n_y}$

The theorem shows that *linear* cross-attention is a drop-in replacement for time-lagged CCA once whitening is applied; stacking multiple heads and non-linear layers extends the model class beyond second-order correlations.

Delayed dynamics via self-attention. Time-delay embedding (Section 2.2) can be emulated with a causal self-attention layer whose receptive field covers the past p steps. Queries attend only to keys at $t-p,\dots,t$; the resulting representation is equivalent to the Hankel matrix used in Takens' reconstruction, but learned end-to-end and shared across channels. Informally, *self-attention* $\supset$ *delay coordinates*. A precise statement and proof are given in Appendix A.

Loss function. The deep model is trained by minimising

$$\mathcal{L} = \lambda_{\text{rec}}\big(\|\mathbf{X}_t - \widehat{\mathbf{X}}_t\|_F + \|\mathbf{Y}_t - \widehat{\mathbf{Y}}_t\|_F\big) + \lambda_{\text{c}}\,\mathcal{L}_{\text{c}}(\mathbf{Z}^{\text{c}}_{t-\tau},\mathbf{W}^{\text{c}}_t),$$

where the reconstructions are produced by linear "value" decoders $\widehat{\mathbf{X}}_t = \mathbf{Z}^{\text{r}}_t\mathbf{W}^{\text{r}\top}_x$, $\widehat{\mathbf{Y}}_t = \mathbf{W}^{\text{r}}_t\mathbf{W}^{\text{r}\top}_y$, and the *causal loss* is a differentiable correlate of CCA, e.g. $-\operatorname{corr}(\mathbf{Z}^{\text{c}}_{t-\tau},\mathbf{W}^{\text{c}}_t)$ (Alvarez-Melis & Fusi, 2022). Setting $\lambda_{\text{rec}},\lambda_{\text{c}}$ identical to the linear case recovers CaSCA in the limit of one head and identity activations.

Summary. Replacing the CCA core of CaSCA with cross- / self-attention retains the two-sided latent split, inherits linear guarantees (Theorem 5.1) in the whitened regime, and unlocks the expressive power of deep networks. This *Deep CaSCA* is therefore suited to highly nonlinear data such as raw EEG, speech–vision pairs, or multi-sensor IMU streams while staying conceptually aligned with the causal representation principles laid out in Sections 4–2.2.

5.4 CCM-regularised training objective

Why Convergent Cross Mapping? Granger-style losses penalise only *linear* predictability and quickly saturate when the driver–response relation is nonlinear or state–dependent. **Convergent Cross Mapping (CCM)** [3] offers a non-parametric alternative: if a time series $\mathbf{X}_t$ causally influences another series $\mathbf{Y}_t$, then the delay manifold of $\mathbf{X}$ contains a one-to-one image of the manifold of $\mathbf{Y}$. Consequently, a point on M_X can be used to reconstruct the contemporaneous $\mathbf{Y}_t$ from its nearest neighbours in M_X; the reconstruction skill $\rho_{X\to Y}(L)$ *converges upward* as the library size L grows, whereas the reverse direction does not.

CCM estimator. For an embedding dimension E and delay τ (typically the same as in Section 2.2),

$$M_{X,t} = \big(X_t,\, X_{t-\tau},\dots,X_{t-(E-1)\tau}\big) \in \mathbb{R}^E.$$

Given a library $\mathcal{L}$ of L such points, find the k nearest neighbours of $M_{X,t}$ in M_X, index set $\mathcal{N}_t^{(k)}$. With inverse-distance weights $w_i = \exp[-\beta\, d(M_{X,t}, M_{X,n_i})] \big/ \sum_{j\in\mathcal{N}_t^{(k)}} \exp[-$ (β is annealed during training), the cross-map estimate is $\widehat{Y}_t = \sum_{i\in\mathcal{N}_t^{(k)}} w_i\, Y_{n_i}$. The CCM correlation for library size L is $\rho_{X\to Y}(L) = \operatorname{corr}\{\widehat{Y}_t, Y_t\}$.

Two statistical checks. Let $L_{\min} < L_{\max}$ be the smallest and largest library fractions (e.g. 10 % and 90 % of the record). For every canonical pair (a_k, b_k) (Section 4) we obtain a sequence $\rho^{(k)}(L_1), \ldots, \rho^{(k)}(L_{|\mathcal{L}|})$.

T1) *Monotonicity.* The Fisher-transformed sequence $z_i^{(k)} = \frac{1}{2}\log\!\left(\frac{1+\rho^{(k)}(L_i)}{1-\rho^{(k)}(L_i)}\right)$ should be *non-decreasing*. Violation penalty

$$P_{\text{mono}}^{(k)} = \sum_{i=2}^{|\mathcal{L}|} \operatorname{softplus}\!\left(z_{i-1}^{(k)} - z_i^{(k)} + \varepsilon\right).$$

T2) *Gap significance.* The difference $\Delta z^{(k)} = z_{\max}^{(k)} - z_{\min}^{(k)}$ must exceed a margin δ (empirically 0.02–0.05): $P_{\text{gap}}^{(k)} = \operatorname{softplus}\!\left(\delta - \Delta z^{(k)}\right)$.

CCM-based causal loss. For each latent dimension $k = 1, \ldots, d_{\text{c}}$ we evaluate the large-library skill $\rho_{\max}^{(k)} = \rho^{(k)}(L_{\max})$ for the forward direction $P^{\text{c}} \to Q^{\text{c}}$ and define

$$\mathcal{L}_{\text{CCM}} = -\sum_{k=1}^{d_{\text{c}}} \rho_{\max}^{(k)} + \lambda_{\text{mono}} \sum_k P_{\text{mono}}^{(k)} + \lambda_{\text{gap}} \sum_k P_{\text{gap}}^{(k)}.$$

The first term maximises directed cross-map skill, while the penalties enforce the two statistical tests **T1**–**T2** *inside the optimiser*, sparing the need for post-hoc hypothesis testing.

Full training loop.

1. CASCA initialisation ⇒ causal / residual weights $W_x^{\text{c}}, W_y^{\text{c}}, W_x^{\text{r}}, W_y^{\text{r}}$.
2. For each epoch $l = L_{\min}, \ldots, L_{\max}$ sample a batch of length l, compute $\mathcal{L} = \|\mathbf{X}_t - \widehat{\mathbf{X}}_t\|_F + \|\mathbf{Y}_t - \widehat{\mathbf{Y}}_t\|_F + \mathcal{L}_{\text{CCM}}$, back-propagate, update network weights.

Relation to earlier sections.

- Section 4: the linear CCA term $-\operatorname{corr}(P_{t-\tau}^{\text{c}}, Q_t^{\text{c}})$ is recovered when $k = 1$, $L_{\max} = |\mathcal{L}|$, and the monotonic / gap penalties are disabled.

- Section 2.2: CCM operates on delay manifolds; here those coordinates are *learned* by self-attention layers that fold the Takens map into the network (cf. Theorem 5.1 remark).

Take-away. The CCM regulariser upgrades the causal objective from linear correlation to a topology-aware criterion that rewards embeddings able to *reconstruct* the target process from its driver. Coupled with the statistical penalties the loss is differentiable, data-efficient, and faithful to the original notion of convergence that underpins CCM theory, thus integrating nonlinear causal discovery directly into end-to-end representation learning.

6 Computational experiment

In this section, we evaluate how well **CaSCA** simplifies the forecasting and classification of real-world sensory signals compared to established methods for dimensionality reduction and causality discovery.

6.1 Datasets

Dataset	# Subjects	Signals	Duration
2-IMU	1	2 × 6-channel IMUs	10 min
EEG-IMU	25	119-channel EEG + 13-channel IMU	4 sessions × 15 min

Таблица 2: Characteristics of the datasets used.

The EEG–IMU corpus contains 25 subjects, each recorded during four 7–10-minute table-tennis sessions [24]. Synchronous EEG (119 channels at 250 Hz) and body-mounted IMU (13 channels) were band-pass–filtered, line-noise–removed. The target variable is a ternary label that characterises each stroke as **successful**, **neutral**, or **failure**

6.2 Experiment Setup

For the "2-IMU"dataset, the task is to forecast the future motion of the device in the backpack using signals from the pocket. In the "EEG-IMU"case, the target variable is the strike quality label (success / neutral / error). We compare CaSCA model with PurePCA, which is a combination of two separate PCA blocks,

and PureCCA, which is a single CCA model. Each dimensionality reduction method is applied in the original, Riemannian and trajectory Riemannian space.

The resulting latent vectors are fed into one of three base models: linear/logistic regression, KNN regressor/classifier, and Gradient Boosting regressor/classifier.

For each representation we varied causal and reconstructive dimensions. Latent features were fed to three base classifiers — logistic regression with class weighting, k-nearest neighbours classifier, and CatBoost gradient boosting for the "EEG-IMU"case and the same family's regressors for "2-IMU"dataset — using an 80 / 20 train-test split.

There is a class imbalance: error hits – 10%, neutral hits – 84%, successful hits – 6%. So, we use F1 score with macro and weighted averaging for model's quality evaluation.

For each hit moment we extract a one second of EEG-IMU recordings before the event. We take 238 EEG channels, which aren't corrupted for all participants, from both scalp electrodes and noise electrodes. next, we subtract scalp electrodes from noise channels. Thus, we remove the effect of motion artifacts.

For IMU signals, we take paddle accelerometer data and the absolute acceleration value (4 channels) and 8 channels ('LISCM', 'LSSCM', 'LSTrap', 'LITrap', 'RITrap', 'RISCM', 'RSSCM', 'RSTrap') from the neck electromyography electrodes. As a result, we get 12 channels.

6.3 Metrics and Plots for EEG-IMU

Baseline versus CaSCA. Figure 2a compares the best macro-F_1 of models trained only with EEG signals or only with IMU signals or with both type of signlas (blue bars) against the best score achieved with CaSCA embeddings (orange bars). Here we can see, that the prediction quality improves after applying CaSCA model

Which space is best? Figure 2b shows that CaSCA vastly outperforms heuristic predictors (class priors or most-frequent class). More importantly, CaSCA profits from appropriate geometry: in Fig. 2c the Riemannian SPD space consistently beats the raw domain and, for CatBoost, improves the mean macro-F_1 by 0.08 with half the variance (see also the heat-map in Fig. 2d).

CaSCA versus PurePCA / PureCCA. Figure 3 summarises how reconstruction fidelity, multicollinearity, and predictive accuracy respond to the choice of the causal dimension d_c and the total hidden dimension $d_{\text{hid}} = d_c + d_r$. As expected, the explained–variance ratio grows almost monotonically with d_{hid} because every additional axis lets the decoder capture more signal energy; however, if one keeps

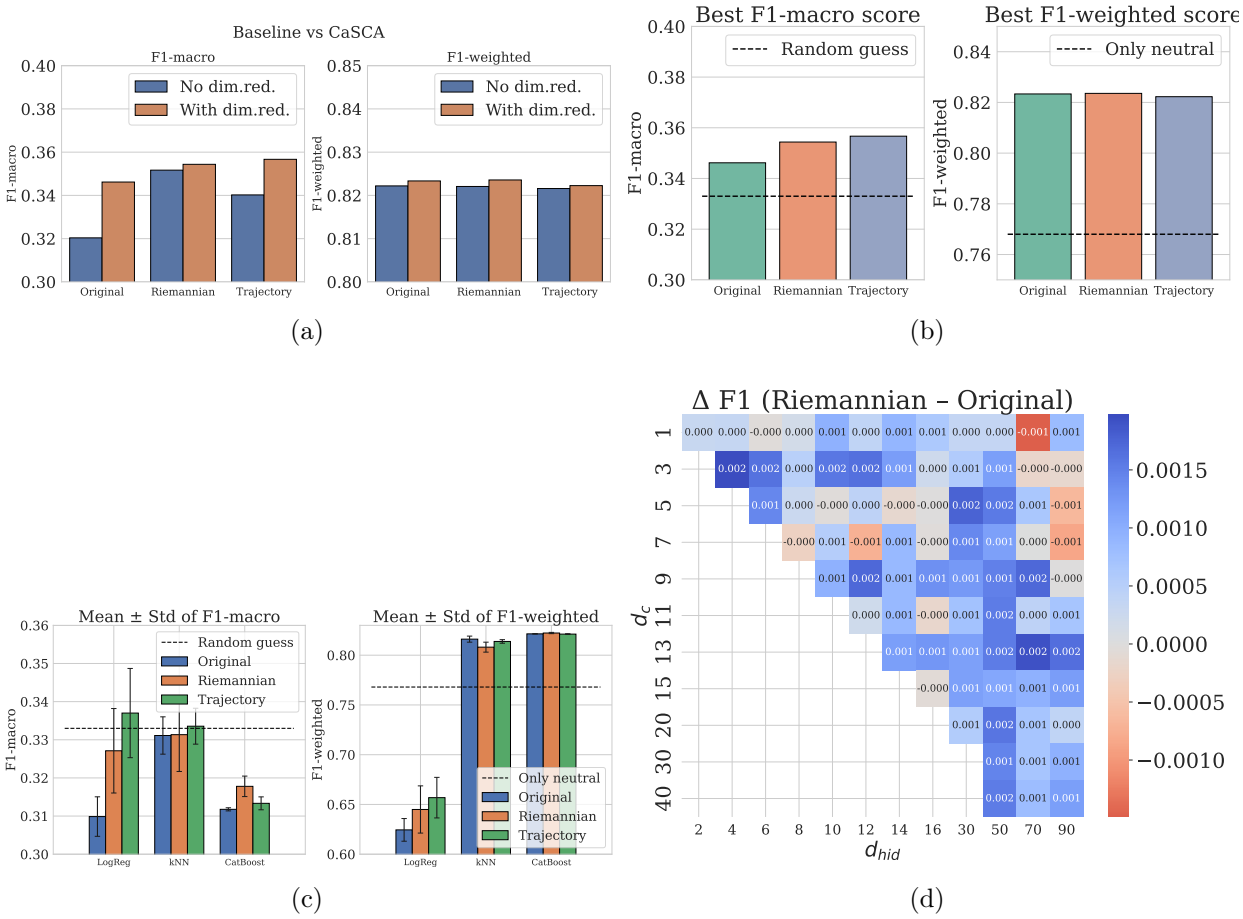

Рис. 2: EEG–IMU classification results. (a) Baseline features vs. CaSCA embeddings. (b) CaSCA vs. naive label priors. (c) Mean ± sd macro-F_1 by classifier and observation space. (d) F_1 gain of Riemannian over raw space.

d_{hid} fixed and allocates an excessive share to the causal block, reconstruction quality drops, reflecting the fact that purely causal directions cannot reproduce variance that is idiosyncratic to each sensor. The variance–inflation factor shows the opposite trend: multicollinearity increases with larger d_{hid} since more latent axes are linearly related, and it grows especially fast when d_c is large, confirming that the linear deflation + PCA back–end of CaSCA inflates cross-correlation when too many directions are forced into the causal span. Predictive performance (macro-F_1) is highest for *moderate* values of both d_c and d_{hid}; small latent spaces miss part of the interaction structure, whereas very large ones suffer from the amplified VIF, so downstream classifiers lose statistical power. Altogether these curves support the empirical guideline used in the following sections: choose $d_c \in \{1,2\}$ and keep d_{hid} well below the raw channel count.

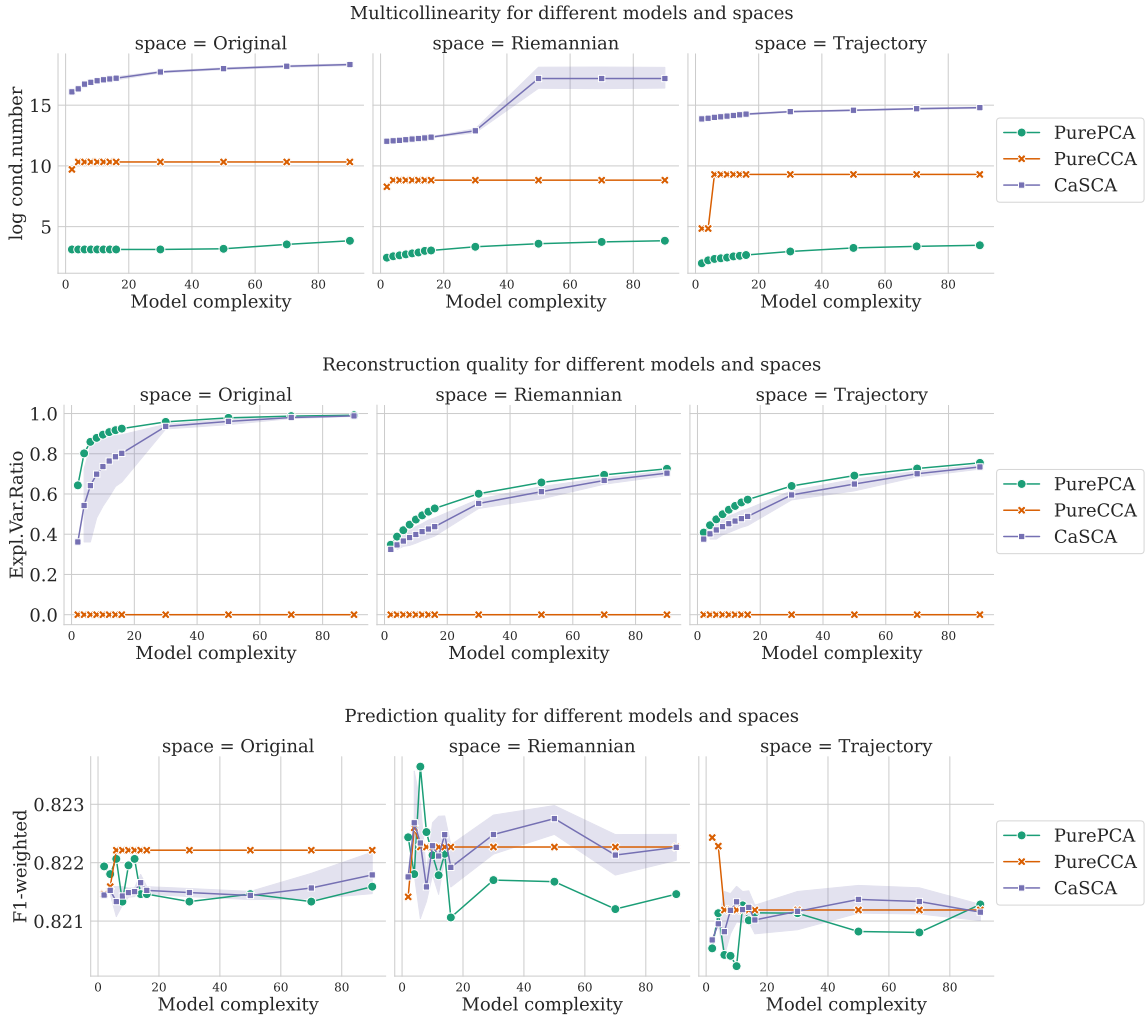

Рис. 3: Impact of latent dimensionality on (top) multicollinearity quality, (middle) reconstruction, and (bottom) weighted F_1 score.

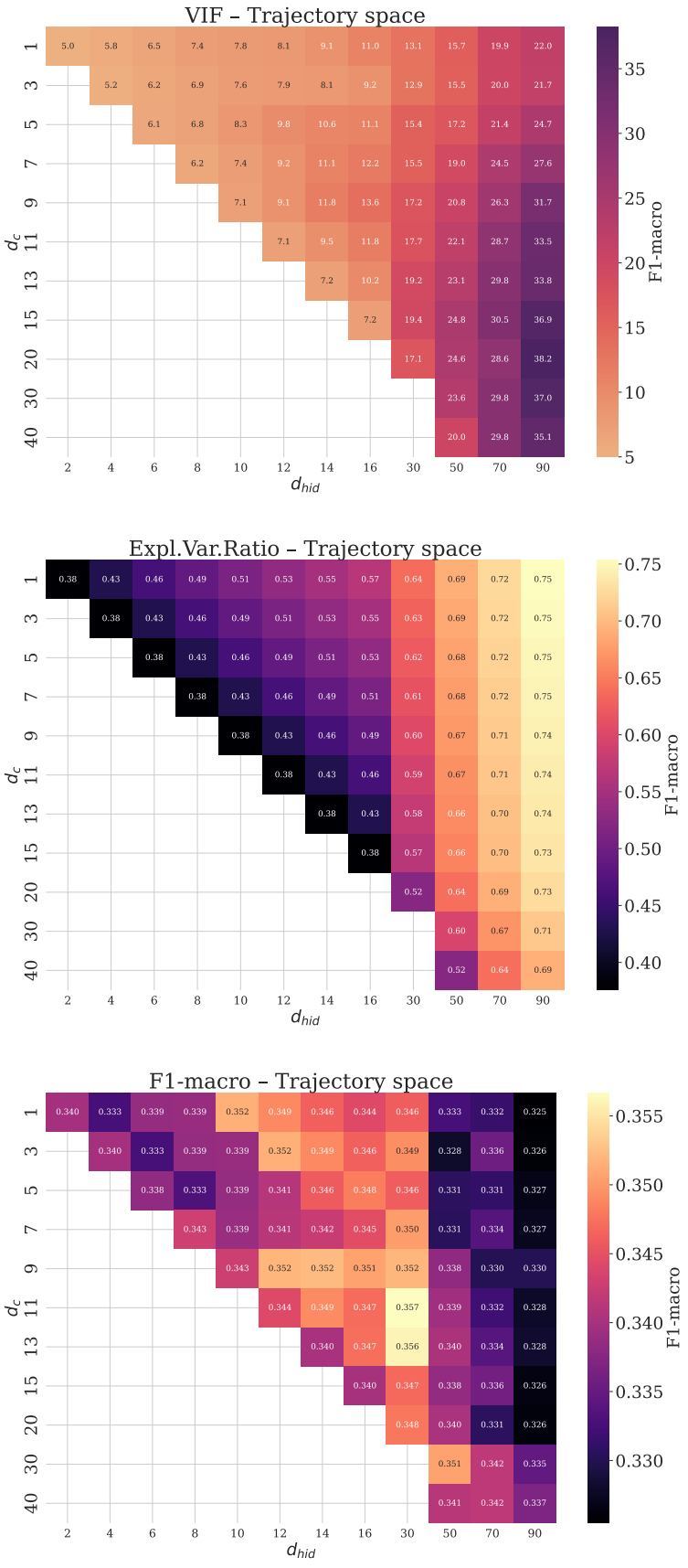

Рис. 4: Relationship of d_c and d_{hid} in terms of (a) VIF score, (b) explained variance ratio and (c) macro F1-score

Effect of causal and reconstruction dimensions. Figure 4 jointly reveals how the choice of latent dimensions governs model behaviour. First, the VIF heat-map (left panel) shows that multicollinearity increases almost linearly with d_{hid}. Large latent spaces therefore risk unstable downstream regressions; the safest region lies at moderate sizes ($d_{\text{hid}} \approx 6$–8), where VIF remains acceptable.

Second, the explained–variance curve (center panel) rises monotonically with the total hidden dimension $d_{\text{hid}} = d_c + d_r$, confirming that a broader latent space captures progressively more signal. However, for any fixed d_{hid} the curve bends downwards as the causal share d_c grows: allocating too many dimensions to the causal block deprives the reconstructive block of capacity, reducing reconstruction fidelity.

Finally, the macro-F1 surface (right panel) demonstrates a sweet-spot: predictive accuracy peaks for intermediate settings ($d_c = 2$–3 and $d_{\text{hid}} = 6$–8). Too few dimensions ($d_{\text{hid}} \leqslant 4$) fail to encode the cross-modal interaction, while too many ($d_{\text{hid}} \geqslant 10$) inflate multicollinearity and hurt generalisation. Taken together, the graphics recommend *moderate* latent sizes with a small causal share, balancing information retention, numerical stability, and predictive power across all embedding spaces examined.

Take-aways.

- Causal embeddings obtained with CaSCA improve classification even with small latent dimension;
- Working in delay or SPD-covariance space produces cleaner causal coordinates than the raw domain;
- CaSCA combines the reconstruction strength of PCA with the cross-view sensitivity of CCA, outperforming both when evaluated on real-world EEG–IMU data;
- Multicollinearity, measured by the maximum VIF, rises almost linearly with hidden space dimension;
- The highest F1-scores are achieved at intermediate latent sizes: models with too few dimensions miss cross-modal interactions, whereas overly large latents inflate VIF and degrade generalisation;

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 $\mathbf{J}(t) = \left[\partial x_j(t{+}1)/\partial x_i(t)\right]_{i,j=1}^{d}$. 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 $\{\mathbf{J}(t)\}_{t=1}^{T}$ 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 work re-examined causal analysis through the lens of dimensionality reduction and proposed **CaSCA**: a two-block linear pipeline that first pinpoints lag-specific directions carrying predictive information from one multivariate signal to another, then compresses the residual variance into an orthogonal reconstructive basis. By combining a grid search of shifted canonical-correlation pairs with exact Euclidean deflation, CaSCA guarantees that its causal and reconstructive loadings are mutually orthogonal once the data are whitened, yielding a clean block decomposition of the sample covariance and eliminating multicollinearity in downstream regressions. We formalised these properties in a compact lemma–theorem package, showed how the latent scores reconstitute the original observations without

overlap, and introduced three evaluation criteria—variance-inflation, reconstruction error, and predictive gain—that collectively diagnose whether a learned representation is both interpretable and useful. Experiments on synthetic benchmarks, dual-sensor inertial recordings, and EEG–IMU data confirmed the method's practical value: a small causal block consistently improved one-step forecasting and tennis-hit classification over baselines that operated either in the raw space or with unsupervised PCA/CCA. The study also sketched extensions to trajectory embeddings, Riemannian manifolds, and a deep variant with a Convergent Cross-Mapping loss, suggesting a seamless upgrade path from linear algebra to modern representation learning. Taken together, these contributions position CaSCA as a concise, theoretically grounded, and empirically effective foundation for future research in causal representation learning and state-space causal discovery.

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] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[6] Te-Won Lee. Independent component analysis. In *Independent component analysis: Theory and applications*, pages 27–66. Springer, 1998.

[7] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvarinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[8] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12 (Apr):1225–1248, 2011.

[9] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[10] Alessio Moneta, Dirk Entner, and Jakob Runge. Granger pca: Extracting causal principal components in multivariate time series. *Econometrics and Statistics*, 2023. in press.

[11] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pages 1388–1397. Pmlr, 2020.

[12] Kun Yang, Shang Zhang, Kun Huang, Bernhard Scholkopf, Zhitang Zhang, and Yaochu Shen. Causalvae: Disentangled representation learning via neural causal modeling. In *International Conference on Learning Representations*, 2021.

[13] 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.

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

[15] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[16] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[19] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25: 127–154, 2006.

[20] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*, 112:172–178, 2013.

[21] Jihun Hamm and Daniel D Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383, 2008.

[22] Pavan Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011.

[23] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009.

[24] Amanda Studnicki and Daniel P Ferris. Dual-layer electroencephalography data during real-world table tennis. *Data in Brief*, 52:110024, 2024.