Kernel Granger Component Analysis for Nonlinear Directed Component Discovery

Deena Francis
Technical University of Denmark
dfra@dtu.dk

Abstract

Granger Component Analysis (GCA) discovers latent components of multivariate time series that exhibit directed temporal dependence, but is limited to linear representations. We introduce Kernel Granger Component Analysis (KGCA), a nonlinear extension that learns directed latent components in a reproducing kernel Hilbert space. KGCA optimizes a ridge-regularized Granger predictive objective using explicit envelope-theorem gradients, and incorporates a time-reversal criterion to address driver-contamination ambiguity. We also describe a scalable random Fourier feature (RFF) variant that approximates the kernel map explicitly and avoids forming the full Gram matrix. To avoid spurious directionality from post hoc component selection, we use a restricted evaluation protocol that measures directed structure intrinsic to the learned representation. Empirically, KGCA and RFF-KGCA recover reliable nonlinear directed components, exhibit positive directionality gaps, and are assessed against independent-process controls. A controlled example clarifies the roles of kernelization and time reversal, while theory specifies conditions for latent recovery and consistency.

1 Introduction

Discovering directed interactions in high-dimensional time series is a central problem in neuroscience, econometrics, and climate science. Granger causality provides a principled notion of directed influence in time series by quantifying whether the past of one variable improves the prediction of another. However, classical Granger analysis operates on observed variables and does not address the discovery of latent driving processes. Conversely, component analysis methods such as Principal Component Analysis (PCA) Hotelling [1933] and Independent Component Analysis (ICA) Hyvarinen [1999], Hyvarinen and Oja [2000] aim to uncover latent representations, but are agnostic to temporal directionality and predictive asymmetry. As a result, standard component methods may yield representations that are statistically informative yet directionally ambiguous, while Granger methods may detect directed dependence without revealing interpretable latent structure. This gap motivates the development of methods that jointly *learn components* and *optimize directed predictability*, which is the central goal of GCA Dmochowski [2023] and its nonlinear extension proposed in this work. GCA identifies linear component projections, referred to as *driving* and *driven* components, by maximizing directed temporal predictability, i.e., the extent to which the past of one component improves prediction of another beyond its own past. This notion of directed dependence is defined at the level of latent temporal dynamics, while the observed time series may be nonlinear mixtures of these latent processes. Under linear observation models, GCA can recover such structure; however, in the presence of nonlinear interactions or nonlinear mixing, linear projections are insufficient to capture the underlying dependencies. KGCA addresses this limitation by lifting the GCA objective into a reproducing kernel Hilbert space, enabling nonlinear component discovery while preserving the same predictive notion of directionality. As illustrated in Figure 1, the latent directed relationship may not be visually apparent in the observed mixtures, yet KGCA is able to recover components

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

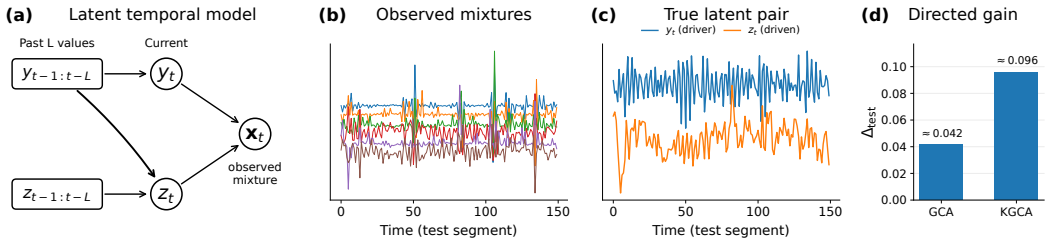


Figure 1: **Latent Granger-causal structure and component discovery.** **(a)** Schematic of the latent driver y_t, driven component z_t, and observed multivariate mixture $\mathbf{x}_t$. The subscript $t-1:t-L$ denotes the preceding L values, e.g., $y_{t-1:t-L} = (y_{t-1}, \ldots, y_{t-L})$. The relation $y \to z$ means that the past of y improves prediction of z_t beyond the past of z alone. **(b)** Observed mixtures from a test segment. **(c)** Ground-truth latent pair from the same segment, vertically offset for clarity. **(d)** Held-out directed gain Δ_{test} for GCA and KGCA; higher values indicate greater predictive improvement from including the driver history.

that exhibit strong directed temporal predictability on held-out data. Importantly, the method targets predictive directional structure rather than causal effects under intervention, and thus should be interpreted as identifying statistically directional relationships in latent space. We emphasize that directional dependence in our framework is defined at the level of latent temporal dynamics rather than at the level of the observed signals. Specifically, we seek latent component pairs (y_t, z_t) such that the past of y improves prediction of z beyond the past of z alone, which defines a Granger-type directional relationship in the latent transition model. The latent-to-observation emission model is distinct from the learned inference maps $x_t \mapsto (f(x_t), g(x_t))$. KGCA makes these component maps nonlinear while retaining linear lagged predictors of their outputs; predictive directionality is defined by the temporal comparison, not by instantaneous mixing alone.

Contributions. We make four contributions: (i) we introduce KGCA, a nonlinear extension of GCA that discovers latent components via nonlinear projections in a reproducing kernel Hilbert space (RKHS); (ii) we derive explicit envelope-theorem gradients and characterize time-reversal ambiguity and conditional recovery for the ridge-regularized component objective; (iii) we introduce an RFF-KGCA variant that optimizes the same directional objective in an explicit random-feature space, avoiding the full Gram matrix; and (iv) we propose and evaluate a restricted evaluation protocol for intrinsic directionality, demonstrating that KGCA-family methods learn directed components without post hoc pair selection.

2 Related work and background

Granger causality formalizes directed dependence via predictability: a process y_t is said to Granger-cause z_t (denoted $y \to z$) if $\mathbb{E}[\mathrm{Var}(z_t \mid \mathcal{F}_{t-1}^{(z)})] > \mathbb{E}[\mathrm{Var}(z_t \mid \mathcal{F}_{t-1}^{(y,z)})]$, where $\mathcal{F}_{t-1}^{(z)}$ and $\mathcal{F}_{t-1}^{(y,z)}$ denote the pasts of z alone and both y, z, respectively. In linear settings, this criterion is typically tested via comparisons of residual variance in vector autoregressive (VAR) models. Traditional approaches to Granger causality predominantly rely on vector autoregressive (VAR) models Shojaie and Fox [2022], Zaremba and Peters [2022], Weichwald et al. [2020] such as the factorized Bayesian vector autoregressive (FBVAR) model Zhao et al. [2024], which assume linear relationships among variables. While effective in many settings, this linearity assumption is often violated in complex real-world systems.

To relax this restriction, a range of nonlinear Granger causality methods have been proposed, which extend predictive modeling beyond linear dependencies. They include approaches based on kernel methods Marinazzo et al. [2008], Ancona et al. [2004], Gaussian process (GP) regression Amblard et al. [2012], Zaremba and Peters [2022], and neural networks Tank et al. [2021]. Kernel-based and GP-based Granger causality methods model nonlinear prediction functions, but are typically used for hypothesis testing between observed channels. Neural Granger causality methods further parameterize predictors with deep networks, but still operate on predefined variables and focus on sparsity or edge discovery rather than component discovery. These methods replace linear predictors with flexible nonlinear models, but typically assume a fixed set of observed variables and do not learn

latent driving signals. As a result, they are primarily used for hypothesis testing between observed channels rather than for discovering latent components with intrinsic directed structure.

Recent nonlinear Granger approaches also include attribution-aided industrial-variable discovery [Sui et al., 2025] and trend/seasonal decomposition followed by causal-network fusion [Yang et al., 2026]. These are relevant precedents for nonlinear prediction or decomposition of predefined series, but do not formulate the same joint search for nonlinear driver–driven projections. Recurrent switching linear dynamical systems (rSLDS) instead learn an expressive generative latent sequence model [Linderman et al., 2017]; we consider downstream directional-pair extraction from its inferred trajectories in Appendix E.2. Our claim is an extension of the GCA component-discovery objective, not the invention of nonlinear Granger causality or time reversal.

Unsupervised component methods such as PCA and ICA learn latent representations based on instantaneous structure, but ignore temporal directionality and predictive asymmetry. Specifically, PCA identifies orthogonal directions of maximal variance by solving an eigenproblem on the data covariance matrix, while ICA seeks statistically independent components by maximizing non-Gaussianity. When applied to multivariate time series x_t, these methods recover linear projections of the form $c_t = w^\top x_t$ that are optimal for variance or independence, but not for predictive power. As a result, PCA and ICA may mix causal and non-causal sources—even when nonlinear temporal structure is present—because they do not exploit past–future asymmetries.

GCA [Dmochowski, 2023] learns linear projections of a multivariate time series that maximize directed predictive gain. Given component time series y_t (driving) and z_t (driven), GCA quantifies directionality by comparing two predictive models for z_t. In the *reduced* predictor (Eqn. 1), z_t is regressed only on its own past,

$$z_t = \sum_{\ell=1}^{L} a_\ell z_{t-\ell} + \varepsilon_t^{(r)}, \tag{1}$$

where L is the lag order, $\{a_\ell\}_{\ell=1}^{L}$ are the coefficients of the autoregressive predictor based on the past of z_t, and $\varepsilon_t^{(r)}$ denotes the residual (innovation) term of the reduced model. In the *full* predictor (Eqn. 2), the past of y_t is included as an additional regressor,

$$z_t = \sum_{\ell=1}^{L} a_\ell z_{t-\ell} + \sum_{\ell=1}^{L} b_\ell y_{t-\ell} + \varepsilon_t^{(f)}. \tag{2}$$

If the full predictor yields a lower prediction error than the reduced predictor, then y is said to Granger-cause z in a predictive sense. GCA searches for linear projections $y = Xw$ and $z = Xv$ that maximize this reduction in predictive error, thereby discovering (linear) latent components that exhibit directed predictive relationships. KGCA generalizes this idea to nonlinear feature spaces while retaining the predictive interpretation.

KGCA differs fundamentally from prior nonlinear Granger causality methods by jointly learning latent components and their directed temporal *predictive* relationship. Rather than testing directed dependence between observed variables, KGCA optimizes a component-level objective that explicitly maximizes predictive gain due to one learned component over another. This places KGCA at the intersection of nonlinear Granger causality and representation learning. In contrast to existing nonlinear Granger approaches, which operate on predefined variables, KGCA learns the representations themselves and selects components that exhibit intrinsic predictive asymmetry.

3 Kernel Granger Component Analysis

We are given a multivariate time series $\{x_t \in \mathbb{R}^D\}_{t=1}^{T}$ and aim to discover pairs of latent component signals y_t, z_t that exhibit directed temporal dependence, with y driving z.

KGCA extends GCA by learning nonlinear component functions $y = f(x)$, $z = g(x)$ in an RKHS that maximize directed predictive gain, measured by the improvement in forecasting z_t from the past of y_t. This allows KGCA to uncover complex, nonlinear driving relationships between latent factors that classical linear GCA would miss. Algorithm 1 summarizes the KGCA workflow, while Algorithm 2 in Appendix A provides the full training procedure.

3.1 Kernel Components

To capture nonlinear structure in time series, KGCA generalizes linear components $y_t = w^\top x_t$ to nonlinear functions of the input x_t. Rather than specifying a parametric form for these functions, we adopt a nonparametric approach by representing components in an RKHS, where inner products correspond to kernel evaluations [Schölkopf and Smola, 2001]. This enables nonlinear component representations while retaining tractable convex inner ridge problems; the outer component optimization remains non-convex. Using the representer theorem [Schölkopf et al., 2001], we express the latent component functions in dual form in terms of the observed samples $\{x_i\}_{i=1}^T$, where each $x_i \in \mathbb{R}^D$ denotes the i-th observation in the time series:

$$y_t = \sum_{i=1}^{T} \alpha_i \, k(x_i, x_t), \quad z_t = \sum_{i=1}^{T} \beta_i \, k(x_i, x_t), \tag{3}$$

where $\alpha, \beta \in \mathbb{R}^T$ are kernel coefficients and k is a positive definite kernel function. This formulation allows KGCA to learn nonlinear projections $x_t \mapsto y_t, z_t$ optimized for directed temporal predictability.

3.2 Ridge Granger Objective

To quantify directed temporal dependence between two component signals y_t and z_t, we compare their predictive power using linear lagged predictors of nonlinear components. Specifically, we assess how much including the past of y improves prediction of z beyond using the past of z alone. We fix a lag order L and use ridge-regularized least squares to ensure numerical stability and prevent overfitting in low-sample or high-lag settings.

Let L denote the lag order and $N = T - L$. Define $z_+ = (z_{L+1}, \ldots, z_T)$ and lag matrices $Z_L, Y_L \in \mathbb{R}^{N \times L}$. We define lag matrices $Z_L, Y_L \in \mathbb{R}^{N \times L}$ by collecting past values of the component signals:

$$Z_L[t, \ell] = z_{t+L-\ell}, \qquad Y_L[t, \ell] = y_{t+L-\ell}, \tag{4}$$

for $t = 1, \ldots, N$, where $N = T - L$ is the number of valid training windows. That is, each row of Z_L contains the past L values of z_t, ordered from most recent to oldest, and similarly for Y_L. We define reduced and full ridge value functions

$$\Psi_r = \frac{1}{N} \min_h \big(\|z_+ - Z_L h\|^2 + \lambda \|h\|^2 \big), \tag{5}$$

$$\Psi_f = \frac{1}{N} \min_g \big(\|z_+ - [Z_L \; Y_L] g\|^2 + \lambda \|g\|^2 \big). \tag{6}$$

Here $h \in \mathbb{R}^L$ and $g \in \mathbb{R}^{2L}$ denote the ridge regression parameters for the reduced and full predictors, respectively, where each predicts z_t from lagged values of z alone (reduced) or from both y and z (full).

3.3 Time-Reversal Objective

Define $\tilde{y}_t = y_{T+1-t}$ and $\tilde{z}_t = z_{T+1-t}$. The forward score alone can leave the driver mixed with the driven component: in unregularized linear prediction, y and $y + cz$ give the same gain for predicting z (Appendix B.10). Following the time-reversal construction of GCA [Dmochowski, 2023], we minimize

$$J = \log \Psi_f + \log \tilde{\Psi}_f - \log \Psi_r - \log \tilde{\Psi}_r. \tag{7}$$

Here $\tilde{\Psi}_r$ predicts $\tilde{y}$ from its own history and $\tilde{\Psi}_f$ also includes the history of $\tilde{z}$: the roles are *swapped* in reversed time. Thus $-J$ sums forward $y \to z$ and reversed $\tilde{z} \to \tilde{y}$ gains. This additional criterion can distinguish contaminated drivers under reversed-time separation; it is not a universal causal-identifiability guarantee.

3.4 Optimization

The predictive loss involves an inner ridge minimization; differentiating through its solver is possible but unnecessary for the value-function gradient. We therefore apply the envelope theorem [Milgrom

Algorithm 1 KGCA: functional overview

Require: Series $\{x_t\}$, kernel k, lag L, ridge λ, pair budget P

1: Form the training-centered Gram matrix; initialize empty pair lists.
2: **for** $p = 1$ to P **do**
3: Optimize one pair using the forward and swapped reversed-time ridge losses (Algorithm 2).
4: Apply RKHS normalization and Gram–Schmidt orthogonalization against earlier pairs during optimization.
5: Store the driver and driven coefficients and their component trajectories.
6: **end for**
7: **return** P ordered pairs

and Segal, 2002], which allows computing gradients of the value function with respect to outer parameters—here, the kernel component coefficients—while treating the inner minimizer as fixed. As a result, gradient updates depend only on residuals and feature responses at the optimum, yielding stable and efficient optimization without backpropagating through the ridge solver Boyd and Vandenberghe [2004]. This avoids differentiating the ridge solution; dense-kernel storage remains a separate scalability constraint. The ridge value function admits closed-form envelope gradients via the envelope theorem. Let $u \in \mathbb{R}^N$ be the response vector, $X \in \mathbb{R}^{N\times d}$ the lagged design, $w \in \mathbb{R}^d$ its ridge minimizer, and $e = u - Xw$. Then the gradient of the value function is given by

$$\frac{\partial \Psi}{\partial u} = \frac{2}{N} e, \qquad \frac{\partial \Psi}{\partial X} = -\frac{2}{N} e w^\top. \tag{8}$$

KGCA alternates between computing ridge-regularized reduced and full predictive losses for kernel components (y, z), and updating their dual coefficients (α, β) via explicit envelope-theorem gradients. After each update, the coefficients are normalized to satisfy the RKHS constraints $\alpha^\top K\alpha = 1$, $\beta^\top K\beta = 1$. The time-reversal objective addresses the driver-contamination ambiguity (Section 3.3). KGCA extracts a fixed number P of component pairs, treated as a hyperparameter. For each pair, we use RKHS Gram–Schmidt orthogonalization to ensure that the learned components are approximately orthogonal in kernel space, preventing redundancy across pairs. This acts as an implicit deflation step without modifying the kernel matrix. We do not currently use an adaptive stopping rule, but observe that extracting $P = 2$–3 pairs suffices in practice. The output is P ordered pairs ($2P$ trajectories), not an unrestricted graph over shared latent nodes; cross-pair edges are not jointly estimated. Sensitivity to P is reported in Appendix E, Figure 4.

Scalable variant. We also implement an RFF-KGCA variant for scalability, replacing the dense Gram matrix with an explicit random Fourier feature matrix $\Phi(X) \in \mathbb{R}^{T\times M}$ and optimizing the same directional objective in primal feature space. For the RBF kernel, RFF-KGCA samples random frequencies ω_m and phases b_m and constructs an explicit feature map $\phi(x) \in \mathbb{R}^M$ such that $k(x, x') \approx \phi(x)^\top \phi(x')$ [Rahimi and Recht, 2007]. The component functions are then parameterized as $y_t = a^\top \phi(x_t)$ and $z_t = b^\top \phi(x_t)$, and the same KGCA objective is optimized in the primal feature space. This reduces representation storage from $O(T^2)$ to $O(TM)$; Appendix E.3 reports the Sim1 scalability ablation. A linear kernel restricts the component class to linear maps, polynomial features admit bounded-degree interactions, and the RBF kernel permits smooth nonlinear components. Our realizable ablation tests linear versus quadratic features; it is not an exhaustive kernel or random-feature-dimension sensitivity study.

3.5 Theoretical properties and scope of recovery

We distinguish optimization identities from statistical recovery. The latter concern a population component class and a global or near-global optimizer; they do not establish blind source identifiability under arbitrary nonlinear emissions. Complete statements and proofs are in Appendix B.

Non-redundant predictive signal. Let Y_t^-, Z_t^- denote lag-L histories, $\mu_f = \mathbb{E}[z_t \mid Z_t^-, Y_t^-]$, $\mu_r = \mathbb{E}[z_t \mid Z_t^-]$, and $\nu = \mathbb{E}(\mu_f - \mu_r)^2$. For optimal square-loss prediction over all square-integrable functions, with $0 < \Psi_f^\star < \infty$,

$$\Psi_r^\star = \Psi_f^\star + \nu, \qquad \Delta^\star = \log(1 + \nu/\Psi_f^\star). \tag{9}$$

Thus positive gain measures additional *conditional-mean* information. For a restricted ridge predictor, Appendix B.7 gives the corresponding lower bound after accounting for approximation and regularization error; the identity is not a claim of zero finite-sample gain under a null.

Conditional population and finite-sample recovery. Suppose the true components are realizable in a bounded-norm, centered, unit-variance RKHS class. Write $Q = -J$ and let $\mathcal{E}^\star$ be their sign-equivalence class. If candidates at population L_2 distance at least ϵ have score at least $\eta_\epsilon > 0$ below the true pair, every population maximizer lies in $\mathcal{E}^\star$. This separation is an *identifiability assumption*, not a consequence of kernelization. Under a bounded kernel, stationary geometric β-mixing, controlled covering entropy, and positive ridge risks, Appendix B.9 establishes a uniform error ε_N and

$$Q(f^\star, g^\star) - Q(\hat{f}, \hat{g}) \leq 2\varepsilon_N + \xi_N, \tag{10}$$

where ξ_N is empirical optimization error. Recovery within ϵ follows when the right side is below η_ϵ. The appendix makes the regularization scaling and normalization assumptions explicit; convergence of the implemented non-convex optimizer is not asserted.

What time reversal contributes. For unregularized linear prediction, replacing Y by $Y + cZ$ leaves the full predictive span, and hence the forward gain, unchanged. The additional score $\Delta(\tilde{Z} \to \widetilde{Y + cZ})$ can distinguish these drivers. Appendix B.10 proves the invariance and a conditional separation result, and verifies a unique uncontaminated optimum along this family for a Gaussian one-lag model. This is not full identifiability over arbitrary nonlinear components; fixed ridge penalties need not preserve the exact invariance.

4 Experiments

We evaluate directed components on simulated and real-world data. PCA and ICA provide variance- and independence-based decompositions, while linear GCA is the directly comparable directional-component baseline. We additionally discuss nonlinear encoder–predictor and rSLDS comparisons in Appendix E.2; these are not methods that natively output the same ordered pairs. The matched kernel/time-reversal ablation is reported separately with its own protocol.

Implementation details. Unless otherwise stated, KGCA uses a Gaussian radial basis function (RBF) kernel, $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$. The kernel bandwidth σ is selected using a median heuristic on the training data, and the ridge parameter λ is fixed across methods to ensure fair comparison. The number of extracted pairs P, lag order L, random-feature dimension M, and optimization budgets are experiment-specific; the recorded M, P, and L settings are provided in Appendix C, Table 4. To facilitate reproducibility, the full codebase can be found here `https://github.com/deenafrancis/KGCA`

Train/test protocol and fair comparison. We use a contiguous 70/30 train/evaluation split. Component representations are learned using the training segment and subsequently evaluated on the remaining segment. For benchmarks evaluated with refitted ridge probes, reduced and full ridge regressions are refitted on the evaluation-segment component trajectories, and their penalized objective values are used to compute directed gain and directionality gap. Thus, the reported evaluation-segment scores measure directed predictive dependence in representations learned from training data, rather than forecasting errors of frozen training-fitted predictors. To ensure a fair comparison across methods, the same lag order (L) and regularization parameters are used for all predictive models, and no information from the test set is used during component learning or selection. Evaluating directed component discovery requires care. Unsupervised baselines such as PCA and ICA can appear to exhibit directionality if component pairs are selected *after* observing test performance. Such post hoc selection reflects a best-case choice rather than intrinsically learned directed structure. We therefore adopt a restricted evaluation protocol in which PCA and ICA are evaluated on fixed component pairs, while KGCA is evaluated on the component pairs it learns. This restricted evaluation protocol is motivated by well-known pitfalls in causal inference and connectivity analysis, where post hoc selection of variable pairs can inflate apparent directionality or predictive gains Stokes and Purdon [2017], Barnett and Seth [2014]. By fixing component pairs for unsupervised baselines such as PCA and ICA, we isolate intrinsically learned directed structure rather than best-case post-selection effects.

4.1 Datasets

Simulated datasets. We evaluate KGCA on three synthetic benchmark families: *Sim1*, a low-dimensional nonlinear VAR system with nonlinear observations; *HD_NLDS*, a higher-dimensional nonlinear latent dynamical system with sparse directed interactions; and *Lorenz–96*, a coupled nonlinear chaotic system with known unidirectional forcing. Together, these benchmarks test KGCA under nonlinear latent dynamics, nonlinear observations, higher dimensionality, and non-VAR chaotic dynamics. Full generation details are provided in Appendix D; a separate exactly realizable example is given in Appendix E.1.

Climate application (ERSSTv5). We also evaluate KGCA on monthly sea surface temperature (SST) anomalies from the ERSSTv5 dataset (1950–2020) Huang et al. [2017]. This dataset provides globally gridded monthly SST estimates based on in-situ observations, and is commonly used to study large-scale climate variability. SST anomalies represent deviations from the long-term monthly climatology and capture patterns such as El Niño–Southern Oscillation (ENSO) and other ocean–atmosphere teleconnections. The data undergo standard climate preprocessing, including removal of the seasonal cycle and area-weighting by latitude to account for unequal grid cell sizes. EEG results are retained as a supplementary application in Appendix E.6.

4.2 Metrics

We report directed gain, directionality gap, and a circular-shift rejection rate. For the benchmarks using refitted ridge probes, "test" denotes the segment held out from component-map learning, not from fitting the evaluation regressions. The projected components are centered and scaled using their evaluation-segment means and standard deviations.

(1) Evaluation-segment directed gain. Let $N_{\mathrm{te}} = T_{\mathrm{te}} - L$ denote the number of valid lagged samples in the evaluation segment. Write z_{te}^{+} for the target values and $Z_{L,\mathrm{te}}, Y_{L,\mathrm{te}}$ for the corresponding lag matrices. We fit reduced and full ridge regressions on this segment and evaluate their penalized objectives on the same samples:

$$\Psi_r^{\mathrm{test}} = \frac{1}{N_{\mathrm{te}}} \min_a \left\{ \left\| z_{\mathrm{te}}^{+} - Z_{L,\mathrm{te}} a \right\|_2^2 + \lambda \|a\|_2^2 \right\}, \tag{11}$$

$$\Psi_f^{\mathrm{test}} = \frac{1}{N_{\mathrm{te}}} \min_{a,b} \left\{ \left\| z_{\mathrm{te}}^{+} - Z_{L,\mathrm{te}} a - Y_{L,\mathrm{te}} b \right\|_2^2 + \lambda \big(\|a\|_2^2 + \|b\|_2^2 \big) \right\}. \tag{12}$$

Here, $a, b \in \mathbb{R}^L$ weight the target's own history and the driver history, respectively; the reduced and full coefficients are optimized separately. The directed gain is

$$\Delta^{\mathrm{test}}(y \to z) = \log\big(\Psi_r^{\mathrm{test}} + \epsilon_{\mathrm{num}}\big) - \log\big(\Psi_f^{\mathrm{test}} + \epsilon_{\mathrm{num}}\big), \qquad \epsilon_{\mathrm{num}} = 10^{-12}.$$

This statistic measures the improvement in the fitted penalized objective when driver history is included. Because the full model contains the reduced model with the same penalty on shared coefficients, the gain is nonnegative up to numerical error. A positive value alone therefore does not establish out-of-sample forecasting improvement.

(2) Directionality gap. We quantify directional asymmetry by

$$\mathrm{Gap}^{\mathrm{test}}(y, z) = \Delta^{\mathrm{test}}(y \to z) - \Delta^{\mathrm{test}}(z \to y).$$

A positive gap indicates greater fitted predictive improvement from y to z than from z to y. Both gains use the original temporal order, with the component roles exchanged; this differs from the time-reversed training objective. Restricted results average the metrics over prespecified or returned pairs within each run, and then over runs where applicable.

(3) Circular-shift rejection rate. For each fixed or returned pair, we generate S surrogates by circularly shifting the driver by random offsets τ_s, giving $y_t^{(s)} = y_{t-\tau_s}$ with circular indexing. The reduced and full ridge regressions are refitted for each surrogate to obtain $\Delta_s = \Delta^{\mathrm{test}}(y^{(s)} \to z)$. Writing Δ_{obs} for the unshifted gain, we compute

$$p = \frac{1 + \sum_{s=1}^{S} \mathbf{1}\{\Delta_s \geq \Delta_{\mathrm{obs}}\}}{S + 1}.$$

Table 1: **Restricted evaluation without post hoc pair selection.** Values are mean ± SEM. Δ^{test} denotes evaluation-segment ridge-probe gain as defined in Section 4.2; ridge regressions are refitted and scored on that segment; $\text{Gap}_{\text{test}} = \Delta(y \to z) - \Delta(z \to y)$ measures directional asymmetry; Sig. rate is the within-run fraction of significant pairs under a circular-shift null ($p < 0.05$), averaged over runs. PCA and ICA are evaluated on fixed component pairs, while GCA and KGCA are evaluated on their learned pairs. Left: Sim1 benchmark. Right: higher-dimensional nonlinear latent dynamics benchmark (HD_NLDS).

(a) Sim1 dataset

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate↑
PCA	0.115 ± 0.018	-0.036 ± 0.029	0.55 ± 0.08
ICA	0.080 ± 0.015	-0.010 ± 0.035	0.50 ± 0.09
GCA	0.075 ± 0.011	-0.021 ± 0.017	0.45 ± 0.07
KGCA	0.261 ± 0.031	0.107 ± 0.028	0.80 ± 0.07
RFF-KGCA	$\mathbf{0.413 \pm 0.025}$	$\mathbf{0.232 \pm 0.031}$	$\mathbf{1.00 \pm 0.00}$

(b) HD_NLDS dataset

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate↑
PCA	0.012 ± 0.001	-0.006 ± 0.004	0.300 ± 0.092
ICA	0.008 ± 0.001	-0.001 ± 0.001	0.200 ± 0.054
GCA	0.007 ± 0.000	0.000 ± 0.001	0.067 ± 0.044
KGCA	0.104 ± 0.010	0.084 ± 0.011	$\mathbf{1.000 \pm 0.000}$
RFF-KGCA	0.171 ± 0.000	$\mathbf{0.167 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 2: **Null control and non-VAR nonlinear dynamics.** Left: null-control system with independent latent sources and no directed interactions. Right: coupled Lorenz–96 benchmark. The same restricted evaluation protocol is used as in Table 1.

(a) Null control

Method	$\Delta_{\text{test}} \uparrow$	Gap ↑	Sig. rate ↑
PCA	0.044 ± 0.003	0.005 ± 0.003	0.150 ± 0.045
ICA	0.038 ± 0.003	-0.010 ± 0.005	0.100 ± 0.043
GCA	0.038 ± 0.003	0.000 ± 0.004	0.167 ± 0.057
KGCA	0.043 ± 0.004	0.004 ± 0.004	0.167 ± 0.045
RFF-KGCA	0.041 ± 0.003	-0.002 ± 0.004	0.067 ± 0.031

(b) Lorenz–96 dataset

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate↑
PCA	0.044 ± 0.009	-0.002 ± 0.006	0.857 ± 0.143
ICA	0.009 ± 0.003	-0.000 ± 0.003	0.143 ± 0.143
GCA	0.007 ± 0.002	-0.008 ± 0.003	0.000 ± 0.000
KGCA	0.165 ± 0.012	0.035 ± 0.010	$\mathbf{1.000 \pm 0.000}$
RFF-KGCA	$\mathbf{0.353 \pm 0.082}$	$\mathbf{0.329 \pm 0.081}$	$\mathbf{1.00 \pm 0.00}$

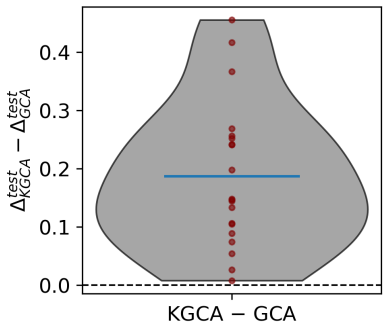
The reported significance rate is the fraction of pairs with $p < 0.05$, averaged over runs where applicable. Circular shifting preserves circular autocorrelation and disrupts temporal alignment, but does not provide a universally calibrated causal null. The test concerns the refitted gain statistic, not the directionality gap or an interventional causal effect.

Separate frozen-predictor evaluations. The realizable ablation and rolling-window analysis use reduced and full ridge predictors fitted on training components and held fixed during evaluation. Their forecast gain is the log ratio of evaluation-segment mean squared errors, without adding the ridge penalty to these errors and with the same numerical stabilization. Unlike the refitted-probe gain, this forecast gain can be negative. For the rolling-window analysis, circular-shift rejection rates are computed separately using refitted evaluation-segment ridge probes; they therefore do not test the frozen-predictor forecast gain (Appendix E.4).

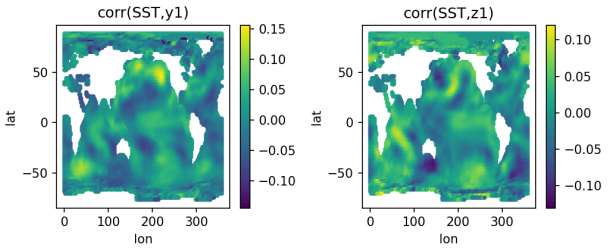
4.3 Results

Table 1 summarizes the main restricted evaluation on synthetic benchmarks. On Sim1, both KGCA and RFF-KGCA substantially outperform PCA, ICA, and linear GCA under restricted evaluation. Exact KGCA achieves $\Delta^{test} = 0.261 \pm 0.031$ with a positive directionality gap, while RFF-KGCA achieves the largest gain among the methods in Table 1(a), with $\Delta^{test} = 0.413 \pm 0.025$, $\text{Gap}^{test} = 0.232 \pm 0.031$, and significance rate 1.00 ± 0.00. KGCA and RFF-KGCA are also the only methods with a clearly positive directionality gap. This indicates that the proposed methods improve prediction and recover asymmetric component pairs. The higher significance rates further show that the recovered directed relationships are consistently detectable across independent runs. The same pattern holds in the HD_NLDS benchmark. Exact KGCA obtains $\Delta^{test} = 0.104 \pm 0.010$ and $\text{Gap}^{test} = 0.084 \pm 0.011$, while RFF-KGCA further improves to $\Delta^{test} = 0.171 \pm 0.000$ and $\text{Gap}^{test} = 0.167 \pm 0.000$. Both KGCA variants achieve significance rate 1.000 ± 0.000, whereas PCA, ICA, and linear GCA remain close to zero as shown in Table 1(b). These results suggest that the advantage of the proposed methods is not limited to the low-dimensional Sim1 setting, but persists in higher-dimensional nonlinear latent systems.

Realizable ablation and faithfulness. Appendix E.1 separates kernelization from time reversal in a matched 2×2 experiment. At $T = 1500$, $\gamma = 1$, quadratic-feature KGCA+TR achieves



(a) Per-run improvement in held-out directed predictive gain over linear GCA.



(b) Spatial correlation maps for the strongest KGCA component pair in ERSSTv5.

Figure 2: **Effect-size consistency and climate interpretability.** (a) Distribution of per-run improvements in held-out directed predictive gain (pair-mean) over linear GCA. Positive values dominate, indicating consistent gains rather than outlier-driven improvements. (b) Spatial correlation maps for the strongest KGCA component pair. The downstream pattern resembles an ENSO-like SST structure, while the upstream component exhibits a broader precursor pattern.

gain 1.570 ± 0.023 against the true-pair reference 1.609, with ordered squared-correlation recovery 0.992 ± 0.003; linear variants cannot recover the masked sources. Across settings, TR changes forward gain little but reduces driver contamination from 0.100 to 0.006. The separate KRR-probe evaluation (Table 5) gives $R^2 = 0.375$ for KGCA versus 0.186, 0.080, and 0.048 for GCA, PCA, and ICA. Probe recovery measures retained information, not one-to-one source identity.

Comparison with linear GCA. The direct comparison with linear GCA highlights the benefit of kernelization. Although GCA explicitly optimizes directed predictability, its linear representation is insufficient under nonlinear latent dynamics and nonlinear observations. On Sim1, linear GCA obtains $\Delta_{\text{test}} = 0.075 \pm 0.011$ and a negative directionality gap, whereas KGCA obtains substantially higher directed gain and a positive gap. Figure 2(a) further shows that the per-run improvement of KGCA over linear GCA is consistently positive, indicating that the aggregate improvement is not driven by a small number of favorable realizations.

Null control and non-VAR dynamics. Table 2(a) assesses independent latent sources without cross-lag coupling under the same observation construction. Mean directionality gaps are small, including 0.004 ± 0.004 for KGCA. However, its reported rejection fraction is 0.167 ± 0.045, above the nominal 0.05 level; the RFF rate is 0.067 ± 0.031. These results do not establish universal false-positive calibration, and a small mean gap should not be conflated with a calibrated surrogate test.

We also evaluate all methods on the coupled Lorenz–96 benchmark, a nonlinear chaotic system that is not of VAR form. As shown in Table 2(b), RFF-KGCA again achieves the strongest restricted performance, with $\Delta_{\text{test}} = 0.353 \pm 0.082$, $\text{Gap}_{\text{test}} = 0.329 \pm 0.081$, and Sig. rate 1.000 ± 0.000. Linear GCA and ICA remain near zero, while PCA achieves modest predictive gain but no positive directionality gap. These results provide evidence that RFF-KGCA and KGCA can recover directed predictive structure beyond the original VAR-based simulation setting.

Climate application (ERSSTv5). Table 3 reports restricted pair-mean results on ERSSTv5. KGCA achieves the highest directed predictive gain, while RFF-KGCA achieves the highest significance rate and positive directionality gap. Linear GCA has weaker predictive gain and a negative mean gap, indicating that its learned components do not exhibit a stable driver–driven orientation in this dataset. PCA and ICA show positive gaps, but lower predictive gain and lower significance rates than KGCA. Figure 2(b) visualizes the spatial correlation patterns for the strongest KGCA climate component pair. Both components exhibit coherent basin-scale structure characteristic of low-frequency SST variability. The downstream component resembles an ENSO-like tropical Pacific

Table 3: Restricted pair-mean evaluation on ERSSTv5. These are single-record point estimates; EEG results are in Table 11.

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate↑
PCA	0.073	0.046	0.33
ICA	0.037	0.025	0.33
GCA	0.033	-0.015	0.00
KGCA	**0.126**	0.020	0.67
RFF-KGCA	0.108	**0.092**	**1.0**

pattern, while the upstream component shows a broader precursor-like structure. This supports the interpretation that KGCA identifies an interpretable directed predictive mode in the climate data. We emphasize, however, that this should be interpreted as predictive directionality rather than physical or interventional causality.

Temporal reproducibility. In the additional 16-window climate analysis, KGCA has mean held-out gain 0.625 and gap 0.523, versus -0.001 and -0.008 for GCA. Reference-excluded pair-map similarity is 0.490 versus 0.278. Appendix E.4 gives the complete table, alignment rule and plots, including the full-record climatology and overlapping-window caveats. These descriptive results support greater map reproducibility than GCA, not an invariant physical mode or an independent-replicate significance claim.

5 Conclusion

We introduced Kernel Granger Component Analysis (KGCA), a nonlinear method for discovering ordered driver–driven latent component pairs in multivariate time series. KGCA combines RKHS component representations, envelope-theorem gradients, and a time-reversal criterion to optimize Granger-type predictive asymmetry directly in latent space. Across nonlinear synthetic benchmarks, including HD_NLDS and Lorenz–96, KGCA recovers directed predictive structure beyond the low-dimensional VAR setting, while independent-source controls reveal small mean gaps but do not establish uniformly calibrated rejection rates. Applications to climate and EEG data further demonstrate coherent predictive modes in real-world time series, which should be interpreted as exploratory Granger-type structure rather than interventional causality.

A realizable ablation separates representation from orientation, and recovery theory is conditional on realizability, separation, concentration and optimization assumptions. The RFF scalability ablation supports explicit random-feature optimization; broader feature-count, low-rank and streaming studies remain future work.

Acknowledgements and Disclosures

Funding. This research received no external funding.

Competing interests. The author declares no competing interests.

AI use statement. A generative AI tool was used to assist with manuscript drafting and editing, code development, and refining mathematical derivations and proofs. All text, code, mathematical derivations, and proofs were independently checked and verified by the author, who takes full responsibility for the accuracy and integrity of the manuscript, results, and conclusions.

References

Pierre-Olivier Amblard, Olivier JJ Michel, Cédric Richard, and Paul Honeine. A gaussian process regression approach for testing granger causality between time series data. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3360. IEEE, 2012.

Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(5):056221, 2004.

Lionel Barnett and Anil K Seth. The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):1, 2008.

Jacek Dmochowski. Granger components analysis: Unsupervised learning of latent temporal dependencies. *Advances in Neural Information Processing Systems*, 36:78168–78180, 2023.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Boyin Huang, Peter W Thorne, Victor F Banzon, Tim Boyer, Gennady Chepurin, John H Lawrimore, Matthew J Menne, Thomas M Smith, Russell S Vose, and Huai-Min Zhang. Extended reconstructed sea surface temperature, version 5 (ersstv5). *Journal of climate*, 30(20):8179–8205, 2017.

Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvarinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial intelligence and statistics*, pages 914–922. PMLR, 2017.

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical review letters*, 100(14):144103, 2008.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL `https://doi.org/10.7551/mitpress/3206.001.0001`.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. the MIT Press, 2001.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual review of statistics and its application*, 9:289–319, 2022.

Patrick A Stokes and Patrick L Purdon. A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proceedings of the national academy of sciences*, 114 (34):E7063–E7072, 2017.

Qingkai Sui, Yalin Wang, Chenliang Liu, Kai Wang, and Bei Sun. Attribution-aided nonlinear granger causality discovery method and its industrial application. *IEEE Transactions on Industrial Informatics*, 21(8):6147–6157, 2025.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Sebastian Weichwald, Martin E Jakobsen, Phillip B Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pages 27–36. PMLR, 2020.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Chunfeng Yang, Jiacheng Wei, Jianqing Li, Bin Liu, Régine Le Bouquin Jeannès, Liangcheng Qu, Kuiying Yin, Xiaojia Wang, and Wentao Xiang. Time series decomposition and fusion-based granger causality network for nonlinear causal inference: Application in emotion brain network analysis using eeg signals. *Biomedical Signal Processing and Control*, 116:109465, 2026.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Anna B Zaremba and Gareth W Peters. Statistical causality for multivariate nonlinear time series via gaussian process models. *Methodology and Computing in Applied Probability*, 24(4):2587–2632, 2022.

He Zhao, Vassili Kitsios, Terence J. O'Kane, and Edwin V. Bonilla. Bayesian vector autoregression with factorised granger-causal graphs, 2024.

Appendix

A Full algorithm

Algorithm 2 gives the full training procedure corresponding to the overview in Algorithm 1.

Algorithm 2 Kernel Granger Component Analysis (KGCA)

Require: Time series $\{x_t\}_{t=1}^{T}$, kernel k, lag order L, ridge parameter λ, number of pairs P, learning rate η, maximum iterations $T_{\max}$

Ensure: Dual coefficients $\{(\alpha^{(p)}, \beta^{(p)})\}_{p=1}^{P}$ (components: $y_p = K\alpha^{(p)},\ z_p = K\beta^{(p)}$)

Compute the Gram matrix K with entries $K_{ij} = k(x_i, x_j)$ and apply centering in RKHS.
Initialize empty lists: $\mathcal{A} \leftarrow [\,]$ (driver coefficients), $\mathcal{B} \leftarrow [\,]$ (driven coefficients)
for $p = 1$ to P **do**
  Initialize $\alpha^{(p)} \sim \mathcal{N}(0, I)$
  Normalize: $(\alpha^{(p)})^\top K \alpha^{(p)} = 1$
  Orthogonalize $\alpha^{(p)}$ in RKHS w.r.t. previously learned drivers:

$$\alpha^{(p)} \leftarrow \alpha^{(p)} - \sum_{j<p} \big[(\alpha^{(p)})^\top K \alpha^{(j)}\big] \alpha^{(j)}$$

  Repeat initialization, normalization, and orthogonalization for $\beta^{(p)}$ w.r.t. $\{\beta^{(j)}\}_{j<p}$
  for $t = 1$ to $T_{\max}$ **do**
    Compute kernel components: $y = K\alpha^{(p)}$, $z = K\beta^{(p)}$
    Compute time-reversed components: $\tilde{y}_t = y_{T+1-t}$, $\tilde{z}_t = z_{T+1-t}$
    Construct lag matrices and compute ridge predictive losses using Eqns. (5) – (6):
      - Reduced and full losses Ψ_r, Ψ_f from (y, z)
      - Reduced and full losses $\tilde{\Psi}_r, \tilde{\Psi}_f$ for predicting $\tilde{y}$ from its own past, with $\tilde{z}$ added in the full model
    Evaluate the KGCA objective:

$$J(\alpha^{(p)}, \beta^{(p)}) = \log \Psi_f + \log \tilde{\Psi}_f - \log \Psi_r - \log \tilde{\Psi}_r$$

    Compute gradients $\nabla_\alpha J$, $\nabla_\beta J$ using envelope-theorem expressions (Sections B.3 and B.4)
    Gradient update:

$$\alpha^{(p)} \leftarrow \alpha^{(p)} - \eta \nabla_\alpha J, \qquad \beta^{(p)} \leftarrow \beta^{(p)} - \eta \nabla_\beta J$$

    Orthogonalize $\alpha^{(p)}$ w.r.t. $\mathcal{A}$ in RKHS
    Orthogonalize $\beta^{(p)}$ w.r.t. $\mathcal{B}$ in RKHS
    Normalize: $(\alpha^{(p)})^\top K \alpha^{(p)} = 1$, $(\beta^{(p)})^\top K \beta^{(p)} = 1$
  end for
  Append $\alpha^{(p)}$ to $\mathcal{A}$, $\beta^{(p)}$ to $\mathcal{B}$
end for
return Dual coefficients $\{(\alpha^{(p)}, \beta^{(p)})\}_{p=1}^{P}$ (equivalently, $\mathcal{A}, \mathcal{B}$)

B Proofs and Derivations

B.1 Notation and ridge value function

Let $u \in \mathbb{R}^N$ denote a response vector and $X \in \mathbb{R}^{N \times d}$ a design matrix. For $\lambda > 0$, define the ridge *value function*

$$\Psi(X, u) := \frac{1}{N} \min_{w \in \mathbb{R}^d} \left(\|u - Xw\|_2^2 + \lambda \|w\|_2^2 \right). \tag{13}$$

Because $\lambda > 0$, the objective in Eqn. (13) is strongly convex and the minimizer

$$\hat{w}(X, u) = (X^\top X + \lambda I)^{-1} X^\top u \tag{14}$$

is unique. Define the ridge residual $e := u - X\hat{w}$.

B.2 Envelope-theorem gradients for ridge regression

Proposition 1 (Envelope gradients for ridge losses)**.** *Let $\Psi(X, u)$ be defined by Eqn.* (13)*, with minimizer $\hat{w}$ and residual $e = u - X\hat{w}$. Then Ψ is differentiable in (X, u) and its gradients are*

$$\nabla_u \Psi(X, u) = \frac{2}{N}\, e, \tag{15}$$

$$\nabla_X \Psi(X, u) = -\frac{2}{N}\, e\, \hat{w}^\top. \tag{16}$$

Proof. Define $F(w; X, u) = \frac{1}{N}\left(\|u - Xw\|_2^2 + \lambda\|w\|_2^2\right)$ so that $\Psi(X, u) = \min_w F(w; X, u)$. Since F is strongly convex in w (because $\lambda > 0$), $\hat{w}$ is unique and satisfies the first-order condition

$$\nabla_w F(\hat{w}; X, u) = \frac{2}{N}\left(-X^\top(u - X\hat{w}) + \lambda\hat{w}\right) = 0. \tag{17}$$

By the envelope theorem, the derivative of $\Psi(X, u)$ with respect to (X, u) equals the partial derivative of $F(w; X, u)$ with respect to (X, u) evaluated at $w = \hat{w}$, *holding $\hat{w}$ fixed*. Thus

$$\nabla_u \Psi = \nabla_u F(w; X, u)|_{w=\hat{w}} = \frac{2}{N}(u - X\hat{w}) = \frac{2}{N}e,$$

which proves Eqn. (15). Next, for each entry X_{ij},

$$\frac{\partial}{\partial X_{ij}}\|u - X\hat{w}\|_2^2 = -2\, e_i\, \hat{w}_j,$$

and the $\lambda\|\hat{w}\|_2^2$ term contributes no derivative when $\hat{w}$ is held fixed. Therefore

$$\nabla_X \Psi = -\frac{2}{N} e\, \hat{w}^\top,$$

which proves Eqn. (16). □

Remark. If one instead defines "residual power" as $\frac{1}{N}\|u - X\hat{w}\|_2^2$ *without* the ridge penalty, then differentiating requires additional terms through $\hat{w}(X, u)$ (implicit differentiation). Using Ψ as in Eqn. (13) avoids this complication and yields explicit gradients.

B.3 KGCA operators and explicit gradients w.r.t. dual coefficients

Kernel components. Given samples $\{x_t\}_{t=1}^T$, define the Gram matrix $K \in \mathbb{R}^{T\times T}$ with $K_{ij} = k(x_i, x_j)$. Kernel components are represented as

$$y = K\alpha, \qquad z = K\beta,$$

for $\alpha, \beta \in \mathbb{R}^T$.

Lag selection operators. Fix lag order L and let $N = T - L$. Let $S_+ \in \mathbb{R}^{N\times T}$ be the "future selector" such that $S_+ s = s_{L+1:T}$. For $\ell = 1, \ldots, L$, let $S_\ell \in \mathbb{R}^{N\times T}$ select lagged values $S_\ell s = s_{L+1-\ell:T-\ell}$. Define

$$D_+ := S_+ K, \qquad D_\ell := S_\ell K.$$

Then

$$z_+ = S_+ z = D_+\beta, \qquad Z_L = [D_1\beta, \ldots, D_L\beta], \qquad Y_L = [D_1\alpha, \ldots, D_L\alpha].$$

Reduced and full ridge losses. Define the reduced and full ridge value functions

$$\Psi_r(\beta) := \Psi(Z_L, z_+), \qquad \Psi_f(\alpha, \beta) := \Psi([Z_L\ Y_L], z_+),$$

where $\Psi(\cdot, \cdot)$ is the ridge value function (Eqn. 13).

Proposition 2 (Gradients of reduced and full losses)**.** *We now express the gradients of the reduced and full ridge losses $\Psi_r(\beta)$ and $\Psi_f(\alpha, \beta)$ with respect to the kernel dual coefficients (α, β), which are the building blocks of the KGCA objective $J(\alpha, \beta)$. Let $(\hat{w}_r, e_r)$ denote the ridge solution/residual*

for $\Psi_r = \Psi(Z_L, z_+)$ and $(\hat{w}_f, e_f)$ for $\Psi_f = \Psi([Z_L\ Y_L], z_+)$. Partition $\hat{w}_f = [\hat{w}_f^{(Z)}; \hat{w}_f^{(Y)}]$ conformably with $[Z_L\ Y_L]$. Then

$$\nabla_\beta \Psi_r = D_+^\top \left(\frac{2}{N} e_r \right) + \sum_{\ell=1}^{L} D_\ell^\top \left(-\frac{2}{N} e_r\, \hat{w}_{r,\ell} \right), \tag{18}$$

$$\nabla_\beta \Psi_f = D_+^\top \left(\frac{2}{N} e_f \right) + \sum_{\ell=1}^{L} D_\ell^\top \left(-\frac{2}{N} e_f\, \hat{w}_{f,\ell}^{(Z)} \right), \tag{19}$$

$$\nabla_\alpha \Psi_f = \sum_{\ell=1}^{L} D_\ell^\top \left(-\frac{2}{N} e_f\, \hat{w}_{f,\ell}^{(Y)} \right), \tag{20}$$

where $\hat{w}_{r,\ell}$ is the ℓ-th entry of $\hat{w}_r$, and $\hat{w}_{f,\ell}^{(Z)}$ (resp. $\hat{w}_{f,\ell}^{(Y)}$) is the ℓ-th entry of $\hat{w}_f^{(Z)}$ (resp. $\hat{w}_f^{(Y)}$).

Proof. From Proposition 1,

$$\nabla_{z_+} \Psi = \frac{2}{N} e, \qquad \nabla_X \Psi = -\frac{2}{N} e\, \hat{w}^\top.$$

For Ψ_r, $z_+ = D_+\beta$ and $Z_L(:, \ell) = D_\ell \beta$ are linear in β. Apply the chain rule:

$$\nabla_\beta \Psi_r = D_+^\top \nabla_{z_+} \Psi_r + \sum_{\ell=1}^{L} D_\ell^\top \left(\nabla_{Z_L} \Psi_r \right)_{(:,\ell)}.$$

Since $\nabla_{Z_L} \Psi_r = -\frac{2}{N} e_r \hat{w}_r^\top$, the ℓ-th column equals $-\frac{2}{N} e_r\, \hat{w}_{r,\ell}$, giving Eqn. (18). The full-loss gradients follow identically, noting that $[Z_L\ Y_L]$ splits into Z-columns depending on β and Y-columns depending on α. □

B.4 Time reversal and objective gradients

Let $R \in \mathbb{R}^{T \times T}$ be the reversal operator, $(Rs)_t = s_{T+1-t}$. Define reversed components $\tilde{y} = Ry$, $\tilde{z} = Rz$. Equivalently, if $y = K\alpha$, then $\tilde{y} = (RK)\alpha$, and similarly for z. These reversed losses correspond exactly to $\tilde{\Psi}_r(\alpha)$ and $\tilde{\Psi}_f(\alpha, \beta)$ appearing in the KGCA objective $J(\alpha, \beta)$ in Eqn. (7).

Construct reversed operators $\tilde{D}_+ = S_+ RK$ and $\tilde{D}_\ell = S_\ell RK$ and define

$$\tilde{\Psi}_r(\alpha) := \Psi(\tilde{Y}_L, \tilde{y}_+), \qquad \tilde{\Psi}_f(\alpha, \beta) := \Psi([\tilde{Y}_L\ \tilde{Z}_L], \tilde{y}_+),$$

where $\tilde{y}_+ = \tilde{D}_+ \alpha$, $\tilde{Y}_L = [\tilde{D}_1 \alpha, \ldots, \tilde{D}_L \alpha]$, and $\tilde{Z}_L = [\tilde{D}_1 \beta, \ldots, \tilde{D}_L \beta]$.

The KGCA objective is

$$J(\alpha, \beta) = \log \Psi_f(\alpha, \beta) + \log \tilde{\Psi}_f(\alpha, \beta) - \log \Psi_r(\beta) - \log \tilde{\Psi}_r(\alpha). \tag{21}$$

Proposition 3 (Gradients of J)**.** *Assume all four ridge losses are positive. Then*

$$\nabla_\alpha J = \frac{1}{\Psi_f} \nabla_\alpha \Psi_f + \frac{1}{\tilde{\Psi}_f} \nabla_\alpha \tilde{\Psi}_f - \frac{1}{\tilde{\Psi}_r} \nabla_\alpha \tilde{\Psi}_r, \tag{22}$$

$$\nabla_\beta J = \frac{1}{\Psi_f} \nabla_\beta \Psi_f - \frac{1}{\Psi_r} \nabla_\beta \Psi_r + \frac{1}{\tilde{\Psi}_f} \nabla_\beta \tilde{\Psi}_f, \tag{23}$$

where the component gradients are given by Proposition 2 (and its reversed analogue).

Proof. Immediate from differentiating Eqn. (21) and using $\nabla \log a = \nabla a / a$. □

B.5 Null directionality

If neither lag history changes the other component's conditional mean, the two *Bayes population* gains are zero by Proposition 5; both directional gaps are then zero. Conditional independence in both directions is sufficient for this statement. It does not imply zero expected empirical log gain, zero fixed-ridge gain, or calibrated circular-shift rejection rates: finite-sample estimation, restricted predictor classes, regularization, and learned mixtures must be considered separately.

B.6 Invariance to RKHS isometries

Proposition 4 (Kernel representation invariance)**.** *Let $\phi(\cdot)$ be a feature map inducing kernel k, and let U be any orthonormal operator on the RKHS. Replacing ϕ by $U\phi$ leaves k unchanged and does not affect KGCA objectives or solutions in terms of the induced component time series y, z.*

Proof sketch. All quantities in KGCA depend on the RKHS only through inner products $\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$ and hence through K. Orthonormal transformations preserve these inner products, leaving the Gram matrix and therefore the objective unchanged. □

B.7 Population prediction and the directed-gain bound

Fix L and a stationary pair with $\mathbb{E}z_t^2 < \infty$. Define $Z_t^- = (z_{t-1}, \ldots, z_{t-L})$ and $Y_t^- = (y_{t-1}, \ldots, y_{t-L})$. In this subsection only, $\Psi_r^\star$ and $\Psi_f^\star$ are the infima of population squared prediction error over *all square-integrable predictors* using Z_t^- and (Z_t^-, Y_t^-), respectively. They are not finite-sample ridge values.

Proposition 5 (Population gain identity and ridge lower bound)**.** *Let $\mu_r = \mathbb{E}[z_t \mid Z_t^-]$, $\mu_f = \mathbb{E}[z_t \mid Z_t^-, Y_t^-]$, $\nu = \mathbb{E}(\mu_f - \mu_r)^2$, and $0 < \Psi_f^\star < \infty$. Then*

$$\Psi_r^\star = \Psi_f^\star + \nu, \qquad \Delta^\star = \log \frac{\Psi_r^\star}{\Psi_f^\star} = \log\left(1 + \frac{\nu}{\Psi_f^\star}\right). \tag{24}$$

For population ridge objectives defined as prediction risk plus $\tau\|w\|^2$, suppose the full class contains a predictor of risk at most $\Psi_f^\star + e_f$ and squared coefficient norm at most B_f^2. Then

$$\Delta_\tau \geq \log \frac{\Psi_f^\star + \nu}{\Psi_f^\star + e_f + \tau B_f^2}. \tag{25}$$

In particular, $\nu > e_f + \tau B_f^2$ suffices for strictly positive ridge gain.

Proof. Conditional expectations minimize squared prediction risk. Decompose $z_t - \mu_r = (z_t - \mu_f) + (\mu_f - \mu_r)$. The cross term has expectation zero because $\mu_f - \mu_r$ is measurable given the full history and $\mathbb{E}[z_t - \mu_f \mid Z_t^-, Y_t^-] = 0$. Taking expected squares gives the risk identity; taking the log ratio gives Eqn. (24). Every reduced ridge predictor has risk at least $\Psi_r^\star$ and a nonnegative penalty. Evaluating the full ridge objective at the assumed candidate gives $\Psi_{f,\tau} \leq \Psi_f^\star + e_f + \tau B_f^2$. Dividing these inequalities proves Eqn. (25). □

For $z_t = a^\top Z_t^- + b^\top Y_t^- + \varepsilon_t$, with $\mathbb{E}[\varepsilon_t \mid Z_t^-, Y_t^-] = 0$ and $\mathbb{E}\varepsilon_t^2 = \sigma^2 > 0$, the identity becomes

$$\Delta^\star = \log\left(1 + \frac{\mathbb{E}[\mathrm{Var}(b^\top Y_t^- \mid Z_t^-)]}{\sigma^2}\right). \tag{26}$$

It concerns prediction of the conditional mean; additional information about conditional variance alone need not produce positive square-loss gain. Neither a positive gain nor this bound identifies which observed nonlinear functions equal the true sources.

B.8 Conditional population recovery

A generative interpretation is $x_t = h(s_t)$, where a stationary latent process contains a target pair $s_t^{(i)}, s_t^{(j)}$. More general emissions can be considered, but exact realizability below requires the target coordinates to be deterministic functions of x_t; for example, an injective noiseless emission with suitable coordinate inverses suffices for decodability. Neither injectivity alone nor temporal coupling establishes the objective-separation assumption. Define

$$\mathcal{F}_B = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B,\ \mathbb{E}f(x_t) = 0,\ \mathbb{E}f(x_t)^2 = 1\}. \tag{27}$$

For fixed $\tau > 0$, let $Q_\tau(f, g)$ be the sum of forward $f \to g$ and reversed-time $\tilde{g} \to \tilde{f}$ population log ridge-risk ratios. Independent sign changes leave this score unchanged. With realizable $f^\star, g^\star \in \mathcal{F}_B$, define $\mathcal{E}^\star = \{(af^\star, bg^\star) : a, b \in \{-1, 1\}\}$ and

$$d((f, g), \mathcal{E}^\star) = \min_{(u,v)\in\mathcal{E}^\star} \left(\|f - u\|_{L_2(P_X)}^2 + \|g - v\|_{L_2(P_X)}^2\right)^{1/2}. \tag{28}$$

Theorem 6 (Recovery conditional on objective separation)**.** *Suppose for every $\epsilon > 0$ with a nonempty complement there is $\eta_\epsilon > 0$ such that*

$$d((f,g),\mathcal{E}^\star) \geq \epsilon \implies Q_\tau(f,g) \leq Q_\tau(f^\star,g^\star) - \eta_\epsilon. \tag{29}$$

Every global maximizer over $\mathcal{F}_B^2$ equals a member of $\mathcal{E}^\star$ almost surely. Moreover, a candidate whose score is less than η_ϵ below the true score lies within distance ϵ.

Proof. The true pair is feasible. A maximizer at positive distance d_0 from $\mathcal{E}^\star$ would, by Eqn. (29) with $\epsilon = d_0/2$, score strictly below this feasible pair, a contradiction. The same argument proves the near-optimal statement. Zero distance implies almost-sure equality because the sign class is finite and hence closed in $L_2(P_X)$. □

If the search class is compact in $L_2(P_X)$, Q_τ is continuous, and its maximizers are precisely $\mathcal{E}^\star$, then the complement of any open ϵ-neighborhood is compact. Its attained maximum is strictly below the global maximum; their difference gives η_ϵ. This observation does not derive uniqueness from a structural data-generating model: Eqn. (29) remains the substantive identifiability assumption. The theorem is a conditional recovery statement, not a general source-identification or latent-graph theorem. The unit-variance population normalization is also an analytical idealization distinct from the finite-sample RKHS unit-norm constraint.

B.9 Finite-sample objective control and consistency

We give an explicit sufficient concentration condition rather than leave the effective sample size undefined. Let $N = T - L$, $d = 2L$, and $M_0 = B\kappa$, where $k(x,x) \leq \kappa^2$; thus $|f(x)| \leq M_0$ on $\mathcal{F}_B$. Assume measurability/separability, stationary absolute regularity $\beta(k) \leq c_0 e^{-c_1 k}$, and finite uniform covering entropy $H(h) = \log \mathcal{N}(h, \mathcal{F}_B, \|\cdot\|_\infty)$ for the radii used below. The entropy is required to grow slowly enough for the stated error to vanish; mere finiteness at each radius is not a rate assertion.

Use the *normalized* ridge convention $\Psi_\tau = \min_w \{\mathbb{E}(u_t - A_t^\top w)^2 + \tau\|w\|^2\}$ and its empirical analogue. Assume all four population risks are uniformly at least $\rho > 0$. In our sum-of-squares convention (Eqn. 13), $\tau = \lambda/N$; fixed τ therefore corresponds to $\lambda = N\tau$.

Theorem 7 (A finite-sample conditional consistency bound)**.** *Choose an integer $q > L$ with $N\beta(q - L) \leq \delta/2$, set $m = \lfloor N/q \rfloor \geq 1$ and $J_0 = 4(d+1)^2$, and let*

$$u_N = 4M_0 h + M_0^2 \sqrt{\frac{2}{m}\left(2H(h) + \log\frac{4qJ_0}{\delta}\right)}, \tag{30}$$

$$C_\tau = 1 + \frac{2dM_0^2}{\tau} + \frac{d^2 M_0^4}{\tau^2}, \qquad \varepsilon_N = \frac{8C_\tau u_N}{\rho}. \tag{31}$$

Here u_N denotes the right side of Eqn. (30)*. If $C_\tau u_N \leq \rho/2$, then, with probability at least $1 - \delta$,*

$$\sup_{f,g \in \mathcal{F}_B} |\widehat{Q}_{N,\tau}(f,g) - Q_\tau(f,g)| \leq \varepsilon_N. \tag{32}$$

For an approximate optimizer with empirical suboptimality at most ξ_N,

$$Q_\tau(f^\star,g^\star) - Q_\tau(\hat{f},\hat{g}) \leq 2\varepsilon_N + \xi_N. \tag{33}$$

Under Theorem 6, $2\varepsilon_N + \xi_N < \eta_\epsilon$ implies distance less than ϵ from $\mathcal{E}^\star$. Consequently, choices with $\varepsilon_N \to 0$ and $\xi_N \to 0$ yield consistency in probability.

Proof. Every required moment is an expectation of a product of two component evaluations in a lag block $(x_{t-L}, \ldots, x_t)$. Approximate each of f, g by an h-net member. A product changes by at most $2M_0 h$, and its empirical-minus-population difference changes by at most $4M_0 h$.

Partition the N lag blocks by their index modulo q. Blocks within each subsequence are separated by at least $q - L$ original time steps. The independent-block coupling for absolute regularity gives total coupling-failure probability at most $N\beta(q - L)$ across these subsequences [Yu, 1994]. On the coupled independent subsequences, each product lies in $[-M_0^2, M_0^2]$. Hoeffding's inequality and a union bound over at most $qJ_0 \exp(2H(h))$ choices give the square-root term in Eqn. (30), with

failure probability at most $\delta/2$. Each subsequence has at least m observations, and the full average is a weighted average of its subsequence averages. Including the coupling error and net approximation bounds every scalar moment deviation by u_N, simultaneously for forward and reversed losses, with probability at least $1-\delta$. Reversed blocks use the same observations in reversed order, so the same argument applies.

For a generic ridge loss put $a = \mathbb{E}u_t^2$, $b = \mathbb{E}(A_t u_t)$, and $\Sigma = \mathbb{E}(A_t A_t^\top)$. Completing the square yields $\Psi_\tau = a - b^\top(\Sigma + \tau I)^{-1}b$. On the moment event, $|\hat{a} - a| \le u_N$, $\|\hat{b} - b\| \le \sqrt{d}\, u_N$, $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \le d u_N$, and $\|b\|, \|\hat{b}\| \le \sqrt{d} M_0^2$. Both regularized inverses have norm at most $1/\tau$. The resolvent identity bounds their difference by $d u_N/\tau^2$; expanding the two quadratic forms therefore gives $|\widehat{\Psi}_\tau - \Psi_\tau| \le C_\tau u_N$. The empirical risks are consequently at least $\rho/2$. The log function is $2/\rho$-Lipschitz on this interval; summing the four log errors proves Eqn. (32).

Finally, insert the empirical scores between the true and fitted population scores. The two uniform errors contribute $2\varepsilon_N$, and approximate maximization contributes ξ_N, proving Eqn. (33). A candidate at distance at least ϵ would lose at least η_ϵ by separation, contradicting the bound when $2\varepsilon_N + \xi_N < \eta_\epsilon$. For fixed δ, geometric mixing permits $q = O(L + \log(N/\delta))$ and hence m of order N/q. Taking $h \to 0$ with $H(h) = o(m)$ gives vanishing error for fixed τ, ρ. Since δ is arbitrary, the distance converges in probability. □

Relation to the implemented optimizer. The bound is for a specified population-normalized class and a near-global optimizer. For a changing normalized penalty τ_N, its explicit factors C_{τ_N} and the risk floor must still permit $\varepsilon_N \to 0$, and the population target may change. Thus fixed sum-scale λ, empirical normalization, and local gradient optimization do not automatically satisfy the consistency conclusion. Nor does this bounded-kernel theorem directly cover unbounded Gaussian observations with an unbounded polynomial kernel; that example is used below for representability and a directly calculable oracle, not as an application of the concentration theorem.

B.10 Driver contamination and the role of time reversal

Proposition 8 (Unregularized invariance and reversed-time separation)**.** *Let $U_t = Y_t + cZ_t$ and $V_t = Z_t$. For optimal unregularized linear lag prediction with positive risks, $F(c) := \Delta(U \to V) = F(0)$. If $R(c) := \Delta(\tilde{V} \to \tilde{U})$ satisfies $R(0) > R(c)$ for every $c \neq 0$, then $F(c) + R(c)$ has its unique maximum at $c = 0$.*

Proof. The reduced target and history are unchanged. Moreover,

$$[V^-, U^-] = [Z^-, Y^-] \begin{bmatrix} I & cI \\ 0 & I \end{bmatrix}, \tag{34}$$

and the block matrix is invertible. Thus both full predictor spaces coincide, proving $F(c) = F(0)$. Then $(F(0) + R(0)) - (F(c) + R(c)) = R(0) - R(c) > 0$. □

The equality need not hold for a fixed Euclidean ridge penalty because the shear changes coefficient norms. Reversing time without swapping the driver and target also leaves the analogous forward predictive span invariant; it is the *swapped reversed-time* term that matters here.

A model in which reversed-time separation can be verified. For the realizable example let $Y_t \sim \mathcal{N}(0, a)$ independently over time and $Z_t = \gamma Y_{t-1} + \varepsilon_t$, with independent $\varepsilon_t \sim \mathcal{N}(0, v)$, $a, v > 0$, and $\gamma \neq 0$. Put $b = \gamma^2 a + v$, $e = \gamma^2 a/b \in (0, 1)$, $r = c^2 b/a$, and use $L = 1$. Forward gain is $F(c) = \log(b/v)$. For reverse-time prediction of U_t from future values, $\mathrm{Var}(U_t) = A = a + c^2 b$, $\mathrm{Cov}(U_t, U_{t+1}) = c\gamma a$, and $\mathrm{Cov}(U_t, Z_{t+1}) = \gamma a$. The full future span is that of (Y_{t+1}, Z_{t+1}), with independent entries. Consequently,

$$\Psi_r^{\mathrm{rev}} = A - \frac{c^2\gamma^2 a^2}{A}, \qquad \Psi_f^{\mathrm{rev}} = A - \frac{\gamma^2 a^2}{b}, \qquad R(c) = \log\left(1 + \frac{e}{(1+r)(1+r-e)}\right). \tag{35}$$

The denominator is strictly increasing for $r \ge 0$, so R is uniquely maximized at $c = 0$. This verifies the proposition's separation assumption on the contamination family for this model; it does not prove uniqueness over the entire nonlinear component class or for general autoregressive/chaotic systems.

C Implementation Details

Kernel centering. All Gram matrices are centered with respect to the training data. When projecting test points, the cross-kernel matrix must be centered using the same training-set means to ensure that test components are expressed in the same RKHS coordinate system. Failure to apply consistent centering can introduce systematic bias in test-time projections.

RKHS normalization. After each gradient update, we enforce the normalization constraints $\alpha^\top K\alpha = 1$ and $\beta^\top K\beta = 1$. These constraints prevent trivial scaling solutions of the log-ratio objective and ensure that component magnitudes remain comparable across iterations and across extracted pairs.

Numerical stability. All ridge-regularized prediction problems are solved using linear system solvers (e.g. Cholesky or QR decomposition) rather than explicit matrix inversion. The ridge parameter $\lambda > 0$ guarantees strong convexity of the prediction objective, leading to well-conditioned solutions and stable envelope-theorem gradients in practice.

Kernel choice. The original KGCA benchmarks use a Gaussian RBF kernel, $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$. The bandwidth σ is selected using a median heuristic computed on the training data. The realizable ablation instead uses explicit degree-two features. No exhaustive RBF-bandwidth or kernel-family robustness claim is made here.

Regularization. The ridge parameter λ is fixed across all methods and experiments to ensure a fair comparison of predictive gains. The realizable ablation reports its own prediction and coefficient penalties in Appendix E.1.

Experiment-specific configurations. We distinguish the number of extracted driver–driven pairs P from the number of random Fourier features M. For the configurations documented in Table 4, RFF-KGCA uses $M = 256$ for the Sim1 $P = 2$ sensitivity run, the full-record climate run, and the additional HD_NLDS and Lorenz–96 comparisons, and $M = 128$ for the BCI IV-2a dataset. All five configurations use a training fraction of 0.7 and record a ridge setting of 10^{-2}; the pair count and lag order are experiment-specific.

Table 4: **Recorded experiment settings.** M is the RFF feature count, P the number of directed pairs, and L the lag order.

Experiment	M	P	L
Sim1: $P = 2$ sensitivity run	256	2	3
Climate	256	3	4
BCI IV-2a	128	2	8
HD_NLDS	256	3	5
Lorenz–96	256	1	5

D Simulated data generation

Simulated datasets. We evaluate KGCA on three synthetic benchmark families.

(1) Sim1. The first benchmark is a low-dimensional nonlinear VAR system with nonlinear observations. The latent state contains directed driver–driven interactions together with nuisance latent variables, while observations are generated through nonlinear mixtures of the latent processes.

(2) HD_NLDS. The second benchmark is a higher-dimensional nonlinear latent dynamical system with $K = 20$ latent variables, $D = 50$ observed channels, and $T = 3000$ time points. The system contains sparse nonlinear latent dynamics with three embedded directed interactions, $(s_1 \to s_2)$, $(s_3 \to s_4)$, and $(s_5 \to s_6)$, while the remaining latent variables act as nuisance processes.

(3) Lorenz–96. The third benchmark is a coupled Lorenz–96 system consisting of two 10-dimensional chaotic subsystems, where the upstream subsystem unidirectionally drives the downstream subsystem,

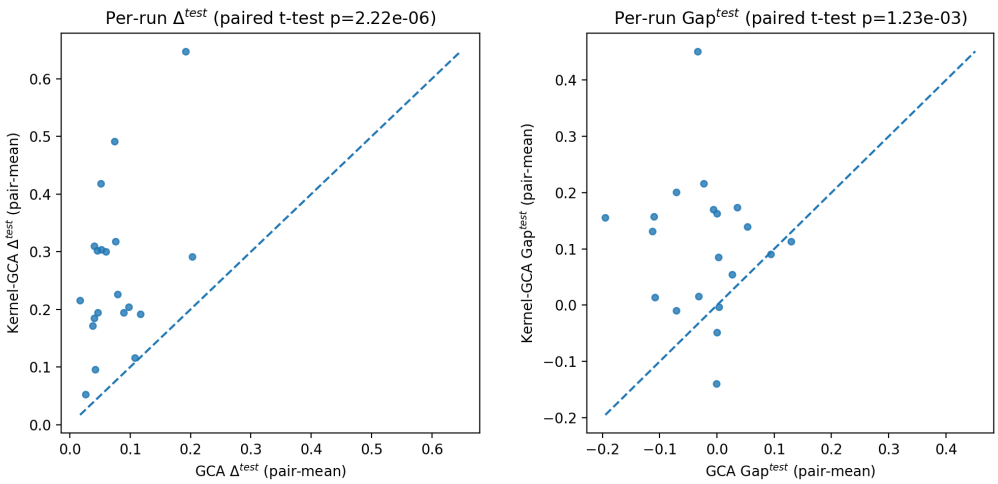


Figure 3: **Left:** Per-run comparison of held-out directed predictive gain (pair-mean) between KGCA and linear GCA. Each point corresponds to one run; the dashed line indicates equal performance. KGCA outperforms GCA in most runs. **Right:** Per-run comparison of directionality gap (pair-mean) between KGCA and linear GCA. KGCA achieves more positive and consistent directionality gaps across runs.

Table 5: Nonlinear recoverability using KRR probes and edge metrics (mean $\pm$ standard deviation over 20 runs). These probe-based metrics differ from the squared-correlation recovery metric used in the realizable ablation.

Method	KRR Probe $R^2 \uparrow$	KRR Edge AUC $\uparrow$	KRR Top-k F1 $\uparrow$
KGCA	$\mathbf{0.375 \pm 0.040}$	$\mathbf{0.756 \pm 0.028}$	$\mathbf{0.550 \pm 0.034}$
GCA	0.186 ± 0.056	0.688 ± 0.043	0.475 ± 0.062
PCA	0.080 ± 0.057	0.744 ± 0.031	0.525 ± 0.048
ICA	0.048 ± 0.063	0.750 ± 0.035	0.450 ± 0.071

yielding a known directed relationship. We generate $D = 50$ observations over $T = 2500$ time points using random linear or nonlinear mixtures of the latent states with additive noise.

Together, these benchmarks test KGCA under nonlinear latent dynamics, nonlinear observations, higher dimensionality, and non-VAR chaotic dynamics.

E Additional results

Figure 3 shows the per-run comparisons between KGCA and linear GCA for held-out predictive gain and directionality gap, respectively. Each point corresponds to one run, with dashed lines indicating equal performance. KGCA consistently achieves higher predictive gain and larger directionality gaps than linear GCA across runs, supporting the aggregate results reported in Table 1.

Number of component pairs P and sensitivity analysis. P controls the extraction budget. Figure 4 reports the existing Sim1 sensitivity analysis. Average pair gain decreases across the tested budgets for the kernel variants, which is consistent with weaker residual structure but does not prove a monotone ordering of individual held-out gains. For practical retention, a separate validation set and a prespecified surrogate/multiple-testing rule can be used; the reported test sensitivity curves are descriptive and are not used to choose P. Sequential extraction does not constitute joint estimation of all cross-pair edges.

Interpreting the faithfulness metrics. A probe regressor predicts each ground-truth latent trajectory from the learned representation; KRR is a nonlinear readout, not an additional component-discovery model. Table 5 therefore measures retained source information rather than exact identity of

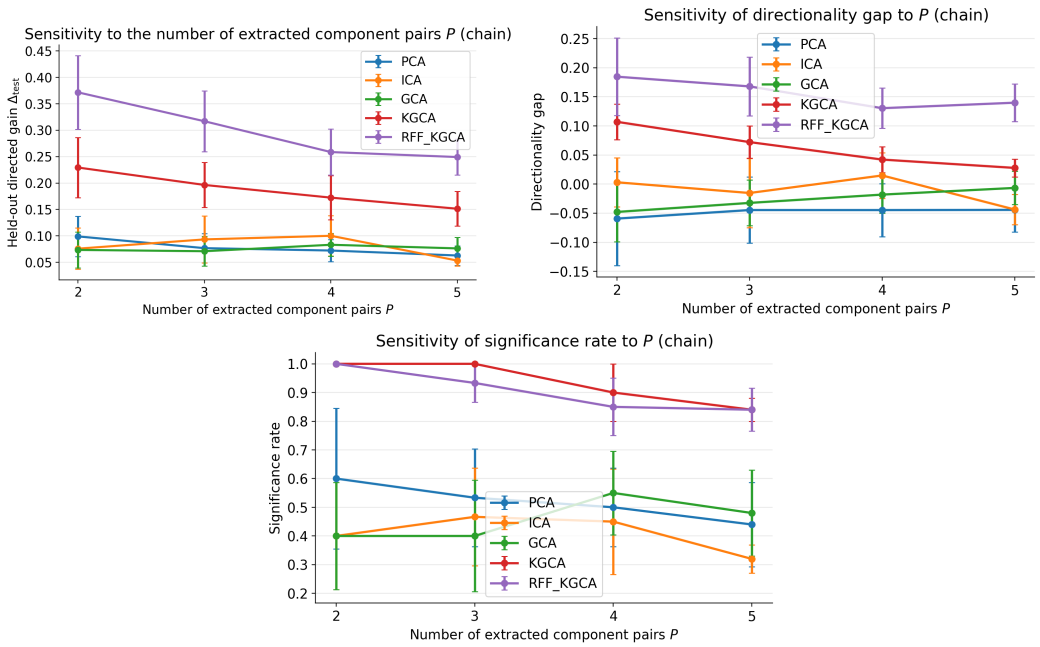


Figure 4: Sensitivity analysis to P for all methods on Sim1 dataset.

an individual component. KGCA has the largest mean probe R^2 and edge AUC, although the AUC difference from ICA is small relative to the reported dispersion. Graph scores require an explicit correspondence to ground-truth nodes: arbitrary learned column indices must not be identified with source indices.

E.1 Realizable example and kernel/time-reversal ablation

This experiment provides an attainable *true-pair reference*, not a universal upper bound over all learned nonlinear projections. After a burn-in of 200 observations, generate

$$y_t = \varepsilon_t^y, \qquad z_t = \gamma y_{t-1} + \varepsilon_t^z, \qquad x_t = (r_t, r_t y_t, q_t, q_t z_t), \tag{36}$$

where all innovations and masks are mutually independent over time, $\varepsilon_t^y \sim \mathcal{N}(0,1)$, $\varepsilon_t^z \sim \mathcal{N}(0, 0.5^2)$, and r_t, q_t are independent equiprobable signs. The true components are $y_t = x_{1,t}x_{2,t}$ and $z_t = x_{3,t}x_{4,t}$. These functions lie in the degree-two polynomial feature span, whereas every individual observed channel has zero population correlation with either contemporaneous latent. Moreover, all nonzero-lag observation cross-covariances vanish under the independent masks, so linear lag prediction of any linear component pair has zero population gain. The oracle for the specified true pair is

$$\Delta^\star(y \to z) = \log(1 + 4\gamma^2). \tag{37}$$

This example establishes representability and a known gain, not automatic convergence of the non-convex component optimizer.

We use $T \in \{500, 1000, 1500\}$, $\gamma \in \{0.25, 0.5, 1\}$, and ten independent replicates per setting. Each variant extracts one pair with $L = 1$, a contiguous 70/30 train/test split, five restarts, 500 Adam iterations, learning rate 0.01, prediction ridge 0.01, and coefficient penalty 10^{-4}. The common ablation optimizer uses either linear features or explicit polynomial features of degree at most two (including a constant); feature scaling is fitted on training data. Representation parameters and prediction coefficients are fitted on training data; the full/reduced predictors are frozen for test evaluation. TR adds the swapped reversed-time term; no ground-truth matching enters the training objective.

Recovery here means squared Pearson correlation, not probe prediction accuracy. Ordered recovery is $\{\mathrm{corr}^2(\hat{y}, y) + \mathrm{corr}^2(\hat{z}, z)\}/2$, with roles fixed. Driver contamination is $\mathrm{corr}^2(\hat{y}, z)$; this is interpretable in this example because y_t and z_t are contemporaneously independent. Low contamination alone is uninformative when latent recovery fails.

Table 6: Matched 2×2 ablation at $T = 1500$, $\gamma = 1$ (mean $\pm$ SEM over ten runs). The true-pair oracle gain is 1.609. Recovery and contamination are squared correlations as defined in the text in appendix Section E.1.

Representation / objective	Test gain	Ordered recovery	Driver contamination
Linear, forward only	-0.010 ± 0.004	0.003 ± 0.001	0.003 ± 0.001
Linear + TR	-0.010 ± 0.004	0.002 ± 0.001	0.004 ± 0.001
Polynomial, forward only	1.570 ± 0.023	0.955 ± 0.011	0.087 ± 0.028
Polynomial + TR	1.570 ± 0.023	0.992 ± 0.003	0.005 ± 0.003

Table 7: **Additional nonlinear-baseline comparisons.** Restricted evaluation uses fixed ordered PCA/ICA pairs and returned ordered pairs for the remaining methods. HD_NLDS has one completed run ($P = 3$), so point estimates are reported without SEM; Lorenz–96 has nine runs ($P = 1$), reported as mean $\pm$ SEM across runs. Sig. rate is the per-run fraction of evaluated pairs declared significant by the benchmark, averaged across runs. Bold indicates the largest reported value within each dataset.

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate
HD_NLDS: linear observations ($R = 1$)			
PCA	0.012	0.002	0.333
ICA	0.004	-0.006	0.000
Linear GCA	0.005	-0.001	0.000
Nonlinear encoder	1.53×10^{-4}	3.66×10^{-5}	0.000
rSLDS	0.112	0.082	$\mathbf{1.000}$
KGCA	$\mathbf{0.171}$	0.144	$\mathbf{1.000}$
RFF-KGCA	0.170	$\mathbf{0.160}$	$\mathbf{1.000}$
Lorenz–96: nonlinear observations ($R = 9$)			
PCA	0.042 ± 0.007	-0.003 ± 0.005	0.778 ± 0.147
ICA	0.010 ± 0.002	-0.001 ± 0.003	0.111 ± 0.111
Linear GCA	0.007 ± 0.001	-0.007 ± 0.003	0.000 ± 0.000
Nonlinear encoder	0.011 ± 0.004	0.001 ± 0.001	0.333 ± 0.167
rSLDS	0.188 ± 0.023	0.083 ± 0.026	$\mathbf{1.000 \pm 0.000}$
KGCA	0.154 ± 0.014	0.033 ± 0.008	$\mathbf{1.000 \pm 0.000}$
RFF-KGCA	$\mathbf{0.353 \pm 0.027}$	$\mathbf{0.329 \pm 0.027}$	$\mathbf{1.000 \pm 0.000}$

Table 6 and Figure 5 show that polynomial features closely approach the true-pair oracle, whereas time reversal cannot repair a misspecified linear component class. Across the nine settings, adding TR to polynomial KGCA leaves mean forward gain essentially unchanged (0.775 for each), increases mean ordered recovery from 0.902 to 0.952 and driver recovery from 0.852 to 0.953, and reduces mean contamination from 0.100 to 0.006. These are equally weighted averages over settings, not independent-replicate confidence estimates. The principal effect of TR is therefore improved driver purity, rather than an increase in forward gain.

Figure 6 presents driver contamination (left) and the scores along $U = Y + cZ$ (right). The explicit sweep over $c \in [-2, 2]$ has forward-score variation approximately 1.4×10^{-6}, while the combined score peaks at $c = -0.05$ near the unmixed value zero (scores approximately 1.337 and 1.335, respectively). This is consistent with sampling and grid error; it is not an empirical proof of exact recovery for all nonlinear data-generating processes.

E.2 Additional nonlinear latent-model comparisons

We compare the nonlinear encoder–predictor and the rSLDS baseline [Linderman et al., 2017] with PCA, ICA, linear GCA, KGCA, and RFF-KGCA in Table 7. The two cohorts are HD_NLDS with linear observations ($T = 3000$, $D = 50$) and Lorenz–96 with nonlinear observations ($T = 2500$, $D = 50$). The rSLDS results refer to the directional-pair extraction pipeline, rather than its native sequence-model likelihood. These are additional experiment cohorts; their results are not pooled with main-table runs obtained under different configurations.

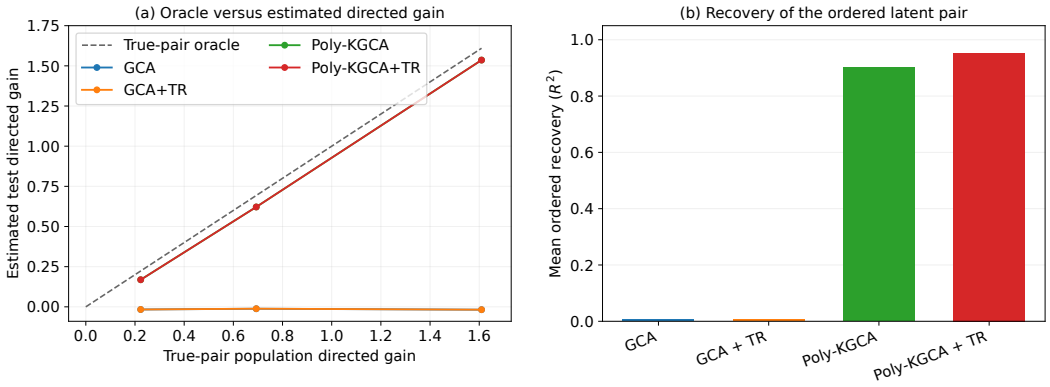


Figure 5: Realizable nonlinear example: estimated versus oracle gain (left) and ordered latent recovery (right), summarized over the ablation settings. The oracle is the population score of the true pair, not an upper bound for every possible representation. Polynomial features recover the nonlinear sources, whereas linear features cannot represent them.

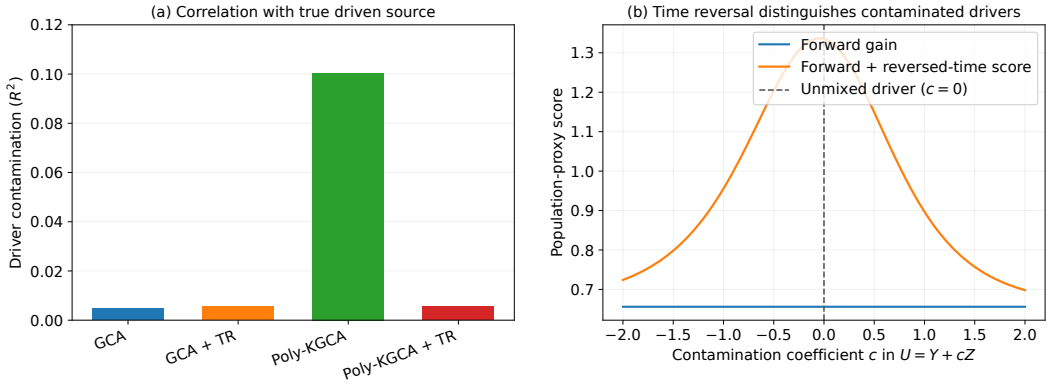


Figure 6: Driver-contamination analysis: driver contamination (left) and scores along $U = Y + cZ$ (right). Time reversal improves driver purity in this example. The sweep illustrates Proposition 8, whose exact invariance is an unregularized population statement. Low contamination alone is not recovery when the linear methods fail to recover either source.

Recorded configuration. Both cohorts use a training fraction of 0.7, lag order $L = 5$, and ridge parameter 10^{-2}. HD_NLDS uses $P = 3$ evaluated pairs and Lorenz–96 uses $P = 1$; PCA/ICA extract six components in both cases. We used 200 KGCA optimization iterations, learning rate 0.02, two GCA restarts, and 256 RFF features. For the nonlinear encoder, 500 epochs, learning rate 10^{-3}, and encoder/predictor hidden widths of 64 were used. The rSLDS configuration uses four discrete states, a five-dimensional continuous latent state, 100 fitting iterations, 50 initialization iterations, 50 posterior-inference iterations, and 20 pair-extraction restarts; its optional validation-based model selection is disabled. The circular-shift procedure is configured with 200 surrogates and a minimum shift of 50 time points.

Results. On Lorenz–96, RFF-KGCA has the largest mean directed gain (0.353 ± 0.027) and directionality gap (0.329 ± 0.027). The strongest external nonlinear baseline, rSLDS, obtains 0.188 ± 0.023 and 0.083 ± 0.026, respectively, exceeding exact KGCA (0.154 ± 0.014 and 0.033 ± 0.008). All three have a reported significance rate of 1.000. Thus, RFF-KGCA leads on the mean directional metrics in this cohort, but exact KGCA does not dominate rSLDS. On HD_NLDS, exact KGCA has the largest recorded gain (0.171), while RFF-KGCA has the largest gap (0.160); both exceed rSLDS in this single realization.

Table 8: Sim1 scalability measurements. Times and memory are approximate; slopes summarize the tested sequence lengths.

Method	Time at $T = 1500$ (s)	Peak memory (MB)	Time slope	Memory slope
Dense KGCA	969	156	2.42	1.99
RFF-KGCA	37	61	1.56	1.36

Table 9: Climate rolling-window results. Errors are descriptive $s/\sqrt{n}$ across $n = 16$ *overlapping* windows (15 for maps). Sig. rate is the reported circular-shift rejection fraction across pairs, averaged over windows.

Method	Test gain	Gap	Sig. rate	Pair-map similarity
GCA	-0.001 ± 0.004	-0.008 ± 0.009	0.333 ± 0.091	0.278 ± 0.018
KGCA	0.625 ± 0.081	0.523 ± 0.061	0.938 ± 0.045	0.490 ± 0.030

E.3 RFF-KGCA scalability ablation

For an RBF kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$, independent frequencies $\omega_m \sim \mathcal{N}(0, \sigma^{-2}I)$ and phases $b_m \sim \mathrm{Unif}[0, 2\pi]$ give features $\phi_m(x) = \sqrt{2/M}\cos(\omega_m^\top x + b_m)$ [Rahimi and Recht, 2007]. RFF-KGCA optimizes the same *form* of directional objective in this approximate feature space; a finite random map is not identical to the exact kernel. Storing its feature matrix costs $O(TM)$ rather than $O(T^2)$ for the dense Gram matrix. These are representation-storage costs, not a complete bound on optimizer memory or runtime.

The Sim1 scalability measurements cover $T \in \{200, 500, 1000, 1500\}$. Table 8 retains the reported endpoints and empirical log–log slopes. At $T = 1500$, the runtime reduction is approximately 26-fold and peak memory is lower. Empirical slopes over four sequence lengths are not asymptotic complexity guarantees. This study varies sequence length, not the random-feature count.

E.4 Temporal reproducibility in the climate record

We supplement the single-period climate example with 16 windows of length 360 time points. The regular step is 120; the final window is right-aligned with the end of the available record (zero-based indices 1703–2062; stop index 2063 is exclusive), so this additional analysis covers a longer record than the 1950–2020 example in the main dataset description. Window centers are plotted as time indices. Within each window the evaluation procedure fits components and ridge predictors on its initial training segment and computes forecast-error ratios on the remaining segment. The rolling-window implementation first removes the monthly climatology over the loaded record; subsequent standardization is training-window based.

To compare maps, for each method we take its first window as reference and match ordered pairs by maximizing the average absolute spatial correlation of their driver and driven maps. Driver/target roles are not swapped. Spatial correlation maps are computed from the window's observations and component trajectories, so this is retrospective map reproducibility, separate from held-out prediction. We exclude the first window's unit self-similarity from the map average. Matching removes sign/permutation ambiguities for description; it does not establish uniqueness or semantic identity of a climate mode. The results are shown in Table 9.

Figure 7 shows that KGCA has positive held-out gain and gap in all 16 windows, compared with 6 and 5 windows for GCA. The gain varies substantially, from approximately 0.002 to 1.066. Figure 8 shows the corresponding spatial-map reproducibility relative to each method's first window; excluding the reference self-comparison, pair-map similarity averages 0.490 for KGCA versus 0.278 for GCA. These observations support *relative temporal reproducibility*, not invariance of a single physical mode. A positive gap can coexist with weak forward gain when the reverse predictor generalizes worse; the gap should not be interpreted in isolation. The window overlap precludes treating these points as independent evidence for a between-method significance test. The circular-shift rejection rate tests its own surrogate statistic, not the sign of the gap or an interventional causal effect.

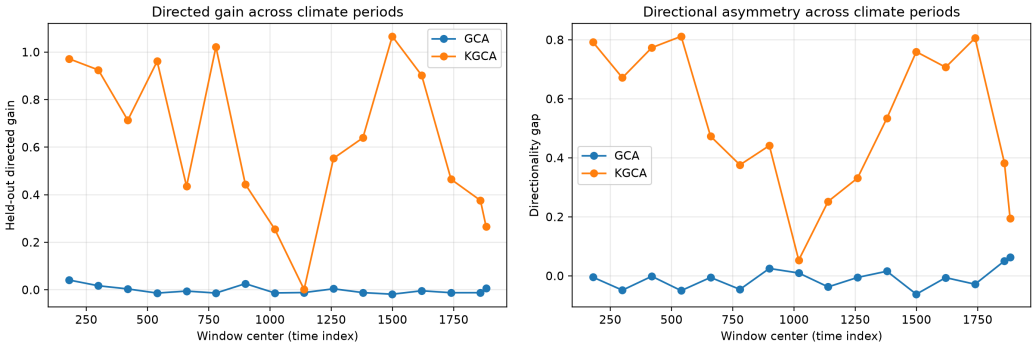


Figure 7: Climate-period predictive gain (left) and directional asymmetry (right). The 360-point windows overlap, so these are descriptive temporal profiles rather than independent replication. A positive gap must be interpreted alongside forward gain, especially when reverse prediction generalizes poorly.

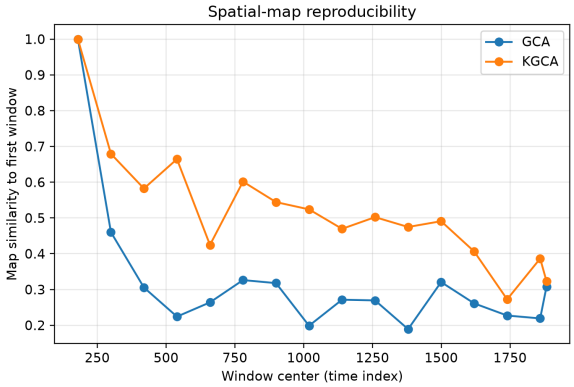


Figure 8: Ordered spatial-map reproducibility relative to each method's first window. Its self-similarity is one by construction and is excluded from Table 9. Pair matching handles sign and permutation ambiguities, not physical identity; the declining similarities motivate cautious, exploratory interpretation.

Table 10: **Climate application (ERSSTv5)**: best-pair directed predictive results (descriptive post-selection analysis).

Method	Δ^{test} (best)↑	Gap^{test} (best)↑	p (shift null)
PCA	0.181	0.125	0.005
ICA	0.049	0.041	0.040
GCA	0.037	-0.010	0.169
KGCA	0.226	0.219	0.035

E.5 Climate data results

Table 10 reports climate results using a best-pair evaluation, in which the single component pair with the largest held-out directed predictive gain is shown for each method. Under this evaluation, KGCA identifies a component pair with higher directed predictive gain and a positive directionality gap compared to linear GCA, indicating stronger asymmetric predictability for at least one learned pair. PCA and ICA also exhibit isolated directional effects for their strongest pairs; however, these effects rely on post hoc selection and are not representative of consistent directionality across components.

We emphasize that this best-pair analysis is included for completeness and interpretive comparison with prior climate and connectivity studies, where reporting the strongest mode is common. The main text therefore focuses on restricted, pair-mean evaluation to assess intrinsic directionality, while the

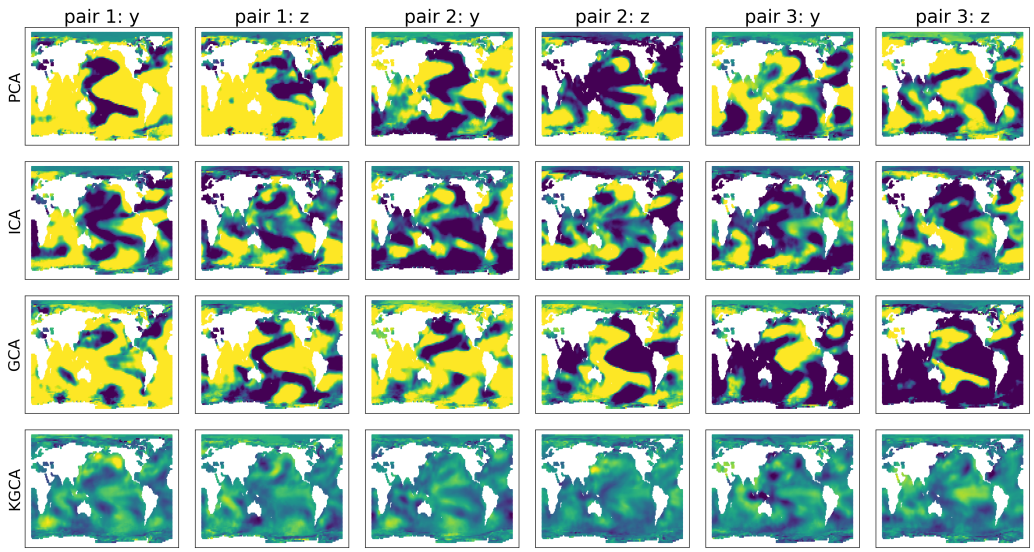


Figure 9: Qualitative comparison of SST correlation maps across methods under restricted evaluation. Each row corresponds to a method (PCA, ICA, GCA, KGCA), and columns show fixed component pairs (y_p, z_p) with no post-hoc selection. All maps use a shared color scale and are computed on the training period. KGCA yields stable, coherent basin-scale patterns with clear directional differentiation, while PCA, ICA, and linear GCA exhibit noisier or less consistent spatial structure.

Table 11: Restricted pair-mean evaluation on the BCI IV-2a EEG dataset.

Method	$\Delta_{\text{test}} \uparrow$	$\text{Gap}_{\text{test}} \uparrow$	Sig. rate$\uparrow$
PCA	0.050	0.008	0.00
ICA	0.144	**0.078**	**0.50**
GCA	0.061	0.007	0.00
KGCA	0.163	0.060	**0.50**
RFF-KGCA	**0.208**	-0.169	**0.50**

best-pair results reported here illustrate the maximal directed structure that can be recovered by each method.

Figure 9 shows correlation maps between SST anomalies and component time series for all methods using fixed pairs. While PCA and ICA recover high-variance spatial patterns that are noisy and unstable across pairs, and linear GCA yields partially structured but inconsistent modes, KGCA produces smooth, coherent basin-scale patterns that are stable across extracted pairs. Moreover, the y and z components learned by KGCA exhibit distinct spatial footprints, consistent with directed low-frequency climate variability.

E.6 BCI IV-2a EEG data results

EEG application (BCI IV-2a). We additionally evaluate KGCA on multichannel electroencephalography (EEG) recordings from the BCI Competition IV dataset (Dataset 2a) [Brunner et al., 2008]. This dataset consists of 22-channel scalp EEG recordings collected during motor imagery tasks, and is commonly used as a benchmark for analyzing neural dynamics. EEG signals exhibit structured temporal dependencies and nonlinear interactions across channels, making them a suitable testbed for directed component discovery.

Table 11 summarizes restricted pair-mean results on the BCI IV-2a EEG dataset. RFF-KGCA achieves the largest held-out directed predictive gain, but its directionality gap is negative in this run, suggesting that its strongest predictive structure is not consistently aligned with the learned driver–driven orientation. ICA obtains the largest gap and matches KGCA's significance rate, but RFF-KGCA has the highest predictive gain among all methods. Linear GCA and PCA show substantially weaker

Table 12: BCI EEG dataset. best-pair directed predictive results (descriptive post-selection analysis).

Method	Δ^{test} (best)↑	Gap^{test} (best)↑	p (shift null)
PCA	0.053	0.010	0.398
ICA	0.209	0.145	0.005
GCA	0.087	-0.004	0.114
KGCA	0.250	0.148	0.015

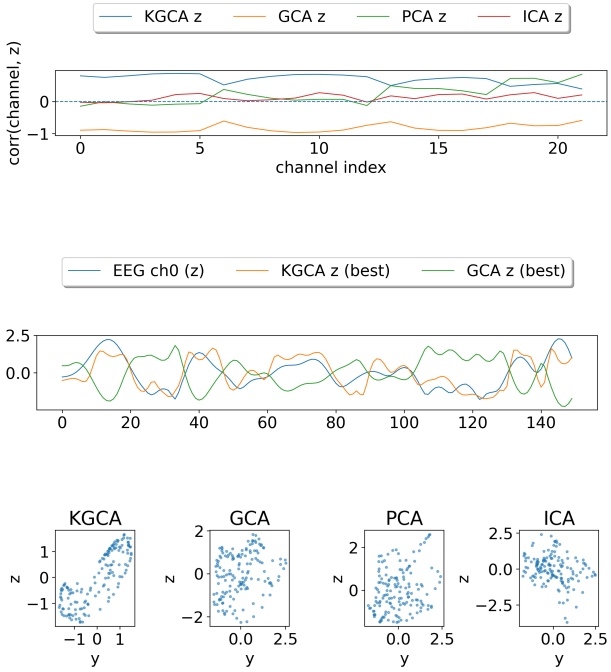


Figure 10: Qualitative analysis on the BCI IV-2a EEG dataset. (**Top**) Channel correlation footprints for the best downstream component. (**Middle**) Time-series overlays of extracted components. (**Bottom**) Joint distributions of component pairs. KGCA yields more coherent and nonlinear structure than linear GCA and unsupervised baselines.

gains and near-zero gaps. Overall, these results suggest that KGCA can recover structured directed predictive components in real high-dimensional time series, while remaining most appropriate as an exploratory tool for hypothesis generation rather than as a standalone causal discovery method. Table 12 reports results on the BCI motor imagery dataset under a best-pair evaluation, where the single component pair with the largest held-out directed predictive gain is shown for each method. Under this analysis, KGCA identifies a component pair with high directed predictive gain and a positive directionality gap, significant under a circular-shift null. ICA also exhibits a strong best-pair result on this dataset, reflecting the presence of isolated directional structure captured by unsupervised components. In contrast, linear GCA yields weaker and non-significant directionality for its strongest pair, while PCA shows minimal directed asymmetry. As with the climate analysis, these best-pair results are included for completeness and comparison with prior EEG studies.

Figure 10 provides qualitative comparisons of the extracted components on the BCI IV-2a dataset. KGCA yields a downstream component with a coherent, distributed channel correlation footprint, while PCA and ICA exhibit weaker and less structured patterns. Time-series overlays show that KGCA components capture smooth envelope dynamics aligned with the observed EEG signal. Scatter plots further indicate that KGCA recovers nonlinear structure between component pairs that is not captured by linear or variance-based baselines.

F Discussion

Predictive directionality versus causality. As with Granger causality more broadly, the directed relationships identified by KGCA reflect *predictive asymmetry* rather than interventional or mechanistic causation Granger [1969], Barnett and Seth [2014]. Formally, for two component time series y_t and z_t, Granger causality from y to z is defined by a strict reduction in optimal prediction risk, $\mathbb{E}\left[\ell\left(z_t, \hat{z}_t^{(r)}\right)\right] > \mathbb{E}\left[\ell\left(z_t, \hat{z}_t^{(f)}\right)\right]$, where $\hat{z}_t^{(r)}$ denotes the optimal predictor based only on the past of z, and $\hat{z}_t^{(f)}$ denotes the optimal predictor based on the joint past of (z, y). KGCA optimizes this predictive criterion at the level of learned latent components rather than observed variables. Consequently, the learned directionality should be interpreted as a property of temporal predictability in the representation space, not as evidence of physical causation. This distinction is particularly important in scientific applications, where predictive influence does not necessarily imply a direct underlying mechanism.

Why kernelization matters. By lifting the latent component search into a reproducing kernel Hilbert space, KGCA effectively optimizes over nonlinear functions of the observations while preserving linear predictive structure in feature space. From a risk-minimization perspective, kernelization enlarges the hypothesis class of candidate components (y, z) over which the predictive gain $\Delta(y \to z) = \log \Psi_r - \log \Psi_f$ is maximized. The consistently higher held-out predictive gains and positive directionality gaps observed in nonlinear simulations (Table 1, 2, Figure 2) indicate that this enlarged function class captures directed dependencies that are inaccessible to linear projections. As with all kernel methods, the expressive power of KGCA depends on the choice of kernel and its associated hyperparameters. An overly restrictive kernel may fail to capture relevant nonlinear structure, while an overly flexible kernel may introduce variance and reduce statistical reliability [Rasmussen and Williams, 2005, Gretton et al., 2012]. The realizable experiment establishes a linear-versus-quadratic contrast; it does not establish robustness across all kernels. Principled kernel selection, bandwidth tuning and systematic RFF feature-count sensitivity remain future work.

Intrinsic directionality and restricted evaluation. A central contribution of this work is the emphasis on restricted evaluation for unsupervised baselines. PCA and ICA do not explicitly optimize temporal directionality. Selecting their component pairs using test performance can inflate apparent directed gain; restricted evaluation avoids this particular source of selection bias. In contrast, KGCA directly optimizes a directionality objective during training, yielding component pairs whose predictive asymmetry persists under restricted evaluation. The higher significance rates observed under a circular-shift null (Table 1) provide a descriptive check on the returned pairs; the null-control rates in Table 2(a) limit claims of calibration. Without a restricted protocol, unsupervised component methods can appear competitive by selecting component pairs that maximize test-set predictive gain. Our evaluation protocol explicitly avoids this pitfall, ensuring that performance differences reflect learned structure rather than favorable selection.

Pairwise scope and recovery. KGCA returns a collection of ordered pairs, not a joint causal graph over identifiable shared nodes. Reconstruction/probe accuracy, directional gain and spatial-map similarity measure different properties; none alone establishes a physical mechanism. The recovery results explicitly assume realizability and population separation, and their statistical guarantees do not imply convergence of a local optimizer or remove the observed null-calibration limitations.

Computational considerations. In its exact form, KGCA scales quadratically with the number of samples due to the kernel Gram matrix, limiting its applicability to moderate-length time series. This computational cost is shared by many kernel-based methods, but is particularly relevant for long climate or neural recordings. Although exact KGCA is limited by Gram matrix construction, scalable kernel approximations such as Nyström methods [Williams and Seeger, 2000] offer a clear path toward faster variants. Our RFF-KGCA experiments show that this approximation can retain or improve directional performance on the synthetic benchmarks while avoiding the dense $T \times T$ Gram matrix. This suggests that the main scalability bottleneck arises from the exact kernel representation rather than from the directional objective itself. Stochastic optimization and streaming formulations may further extend KGCA to large-scale settings without sacrificing interpretability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: KGCA as a nonlinear extension of GCA for discovering directed latent components in an RKHS, the use of a ridge-regularized Granger predictive objective with a time-reversal criterion, and a restricted evaluation protocol. The claims are limited to predictive, Granger-type directionality rather than interventional causality, and this scope is clarified in the introduction and discussion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations including the interpretation of KGCA as predictive directionality rather than causal intervention, dependence on kernel choice and hyperparameters, computational scaling of exact kernel methods, and the role of the number of extracted component pairs P. The discussion also identifies scalable kernel approximations and stronger causal validation as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical properties are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental setup, including the train/test split, lag order, regularization, number of component pairs, evaluation metrics, synthetic data generation, and baseline protocols. The synthetic datasets are generated procedurally, and the paper reports results over multiple independent runs where applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a code repository for reproducing the experiments. The synthetic benchmarks are generated by the provided code, and the real-data experiments use publicly available datasets cited in the paper, including ERSSTv5 and BCI Competition IV-2a.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the key experimental details needed to understand the results, including the contiguous 70/30 train/test split, lag order, ridge parameter, kernel choice, median bandwidth heuristic, number of component pairs, optimization procedure, and evaluation metrics. Additional implementation details are provided in the appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports mean ± SEM over independent runs for the synthetic experiments and uses a circular-shift null test to assess significance of directed predictive gain. The tables define the reported uncertainty and significance rate, and additional paired comparisons are reported in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper discusses computational scaling qualitatively and notes that exact KGCA scales quadratically with the number of samples due to the kernel Gram matrix. However, it does not provide a complete hardware, runtime, memory, or total-compute description for each experiment.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The work develops a methodological tool for predictive directional representation learning, uses synthetic data and public datasets, and explicitly cautions that KGCA should not be interpreted as establishing interventional causality.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive uses of KGCA for exploratory analysis of high-dimensional time series in domains such as climate science and neuroscience. It also discusses the main potential misuse: over-interpreting predictive directionality as causal mechanism, and explicitly states that the learned relationships should be treated as hypothesis-generating rather than definitive causal claims.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose any such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing datasets and related software/assets used, including ERSSTv5 and BCI Competition IV-2a.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases new code for Kernel GCA, synthetic benchmark generation, and experimental evaluation. The code is provided through a code repository and includes the implementation needed to reproduce the main experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new human-subject data collection. The EEG experiment uses an existing public benchmark dataset, and no new participant instructions, compensation, or recruitment procedures were introduced by this work.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new human-subject research. The EEG analysis uses an existing public, de-identified benchmark dataset, so no new IRB approval or equivalent human-subjects review was required for this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An AI use statement is included in the paper. The author takes responsibility for verifying the final scientific content.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.