

A Supplementary Material

A.1 Notation

Symbol	Description
ξ_{ai}	An observation of time-series a at time index i
X_a	Phase-space reconstruction of time-series ξ_a
X	Matrix of all $X.$, excluding X_a
Z	Matrix of all $X.$ in the system
$\mathbf{y}$	Vector of response variable
n	Number of observations
m	Maximum lag order
τ	Specified lag
n_t	Number of time-series in the system
$F_{a \rightarrow b}$	F-test statistic for determining $\xi_a \xrightarrow{GC} \xi_b$
SSR_R, SSR_u	Sum of squared residuals for restricted, unrestricted models
p_r, p_u	Number of parameters in restricted, unrestricted models
$\hat{\mathbf{y}}, \hat{\mathbf{y}}'$	Predictions for restricted, unrestricted models
$\hat{\beta}, \{\hat{\beta}', \hat{\gamma}'\}$	Estimated regression coeff. restricted, unrestricted models
P, P'	Projection matrices restricted, unrestricted models
δ	KGC Granger causality index
δ_f	KGC filtered Granger causality index
$\mathbf{e}$	Prediction error
r_i	i -th Pearson correlation coefficient
n_r	Number of correlation coefficients
n_f	Number of filtered correlation coefficients
K, K'	Kernel matrices for restricted, unrestricted models
$\mathbf{v}_i, \mathbf{v}'_i$	i -th eigenvector of the kernel matrices for restricted, unrestricted models
H, H'	Hypothesis space for restricted, unrestricted models
$H^\perp$	Orthogonal complement of H
$\mathbf{f}(\cdot), \mathbf{g}(\cdot)$	Nonlinear functions transforming covariates
c_f, c_g	LsNGC number of cluster centroids
V, U	LsNGC cluster centers
$\phi(\mathbf{x})$	Feature map for KPCR
$\mathbf{a}_j, \lambda_j$	Eigenvector/eigenvalues for KPCR
$\tilde{K}$	Nyström approximation kernel
l	Lengthscale for SE kernel
τ^2	SE kernel variance
$GP(\cdot, k(\cdot, \cdot))$	Gaussian process with mean and kernel function
ω	Telescoping values used in SIC
$\mathcal{P}(X_{at})$	Parents of X_{at}
S	Set of conditioning variables PCMCI

Table 4 Notation and descriptions.

A.2 Proofs

Proof of Proposition 1

Proof Kernel PCR solves a regression problem where the covariates are given by the data projected onto the principal components Eq. (31), such that the full covariate matrix can be defined as $\tilde{K}(X, X)A\Lambda^{-1/2}$, where $\tilde{K}(X, X) \in \mathbb{R}^{n \times n}$ is the full kernel matrix, $A = [\mathbf{a}_1, \dots, \mathbf{a}_s] \in \mathbb{R}^{n \times s}$ is the matrix defined by the s leading eigenvectors of the kernel, and $\Lambda = \text{diag}(\lambda_i) \in \mathbb{R}^{s \times s}, i = 1, \dots, s$ is a diagonal matrix of the s leading eigenvalues. The function $\mathbf{f}(X)$ can thus be defined as $\mathbf{f}(X) = \tilde{K}(X, X)A\Lambda^{-1/2}$, such that the regression model in Eq. (34) is equivalent to

$$\hat{\mathbf{y}} = \tilde{K}(X, X)A\Lambda^{-1/2}\hat{\beta}.$$

The estimated coefficients are given by least-squares:

$$\hat{\beta} = \left((\tilde{K}(X, X)A\Lambda^{-1/2})^\top \tilde{K}(X, X)A\Lambda^{-1/2} \right)^{-1} (\tilde{K}(X, X)A\Lambda^{-1/2})^\top \mathbf{y},$$

such that the regression model can be written as a projection of the observations $\mathbf{y}$ as it is done in the KGC method, where the orthogonal projection matrix is now equivalent to

$$P = \tilde{K}(X, X) A \Lambda^{-1/2} \left((\tilde{K}(X, X) A \Lambda^{-1/2})^\top \tilde{K}(X, X) A \Lambda^{-1/2} \right)^{-1} (\tilde{K}(X, X) A \Lambda^{-1/2})^\top.$$

In the KGC method, the projection matrix is defined by the n_v eigenvectors of K with non-zero eigenvalue, i.e. $P = \sum_{i=1}^{n_v} \mathbf{a}_i \mathbf{a}_i^\top = A A^\top$, which will coincide with establishing $s = n_v$ in Kernel PCR. By using the eigenvector-eigenvalue decomposition of the kernel matrix $\tilde{K}(X, X) = A \Lambda A^\top$, the projection matrix from Kernel PCR is seen to be equivalent to the projection matrix in the KGC method

$$P = A \Lambda A^\top A \Lambda^{-1/2} ((\Lambda^{-1/2})^\top A^\top (A \Lambda A^\top)^\top A \Lambda A^\top A \Lambda^{-1/2})^{-1} (\Lambda^{-1/2})^\top A^\top (A \Lambda A^\top)^\top = A A^\top,$$

i.e. the regression problem solved in KGC is equivalent to that solved in Kernel PCR. The same rationale follows for the unrestricted regression model, where Eq. (35) is equivalent to $\hat{\mathbf{y}}' = \tilde{K}(Z, Z) A' \Lambda'^{-1/2} \hat{\beta}'$. $\square$

Proof of Proposition 2

Proof The lsNGC method uses c_f cluster centroids, identified via k-means clustering, within the normalized Radial Basis Functions in Eq. (24), reducing the dimensionality of the covariates $c_f < (n_t - 1)m$. This results in a covariate matrix with dimensionality $n \times c_f$, where each element of the matrix is defined by the nonlinear transformation $f_j(X_i)$ for $j = 1, \dots, c_f$ and $i = 1, \dots, n$.

Assume that the number of inducing points used in the Nyström approximation $\tilde{K}$ is equivalent to the number of cluster centroids used in the lsNGC method, $n_J = c_f$. Using the proposed kernel Eq. (36), we can write the components of $\tilde{K}$ as $K(X, X_J) = D_X^{-1} K_0(X, X_J) D_J^{-1}$ and $K(X_J, X_J) = D_J^{-1} K_0(X_J, X_J) D_J^{-1}$, where $K_0(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / l^2}$ is the standard SE kernel, D_X is an n -dimensional diagonal matrix of the normalisation terms $\sum_{k=1}^{n_J} e^{-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / l^2}$ and D_J is an n_J -dimensional diagonal matrix of the normalisation terms $\sum_{k=1}^{n_J} e^{-\|\mathbf{x}_k - \mathbf{x}_j\|^2 / l^2}$. As such, $\tilde{K}$ can be written as

$$\tilde{K} = D_X^{-1} K_0(X, X_J) K_0^{-1}(X_J, X_J) K_0(X_J, X) D_X^{-1},$$

such that its feature vectors can be written as $\phi(X) = D_X^{-1} K_0(X, X_J) K_0^{-1/2}(X_J, X_J)$. It can be seen that $D_X^{-1} K_0(X, X_J) D_J^{-1}$ is equivalent to $\mathbf{f}(X)$ in the lsNGC framework. Thus, the covariates $\phi(X) V$ of a KPCR with Nyström approximation approach, where V is the matrix of $s = n_J$ leading eigenvectors of the covariance matrix $\phi(X)^\top \phi(X)$, is equivalent to the approach used in lsNGC, up to the factor $D_J K_0^{-1/2}(X_J, X_J) V$. However, since this factor is multiplied by β it is absorbed into the parameters during optimisation:

$$\begin{aligned} \tilde{\beta} &= (D_J K_0^{-1/2}(X_J, X_J) V)^{-1} (\mathbf{f}(X)^\top \mathbf{f}(X))^{-1} (D_J K_0^{-1/2}(X_J, X_J) V)^{-\top} (D_J K_0^{-1/2}(X_J, X_J) V)^\top \mathbf{f}(X)^\top \mathbf{y} \\ &= (D_J K_0^{-1/2}(X_J, X_J) V)^{-1} (\mathbf{f}(X)^\top \mathbf{f}(X))^{-1} \mathbf{f}(X)^\top \mathbf{y}, \end{aligned}$$

and so the relationship $\hat{\beta} = (D_J K_0^{-1/2}(X_J, X_J) V) \tilde{\beta}$ proves the equivalence between lsNGC and Kernel PCR. $\square$

Proof of Contemporaneous Algorithm

Assumptions

We consider the phase-space reconstruction of a time series system, $X = [X_{a(t-m)}, \dots, X_{a(t-1)}, X_{at}, X_{b(t-m)}, \dots, X_{b(t-1)}, X_{bt}, \dots, X_{n_t(t-m)}, \dots, X_{n_t t}]^\top$, which adheres to the underlying Structural Causal Model (SCM)

$$X_{bt} = f(\mathcal{P}(X_{bt})) + \varepsilon_{bt}, \varepsilon_{bt} \sim \mathcal{N}(0, \sigma^2).$$

The assumed SCM is a linear or nonlinear function of the (lagged and contemporaneous) parents $\mathcal{P}(X_{bt})$ of the variable with additive Gaussian noise that is independent across time points and between time series.

The assumptions made regarding causal graph identifiability (Spirtes et al. 2000; Pearl 2000; Runge 2020) include

1. *Causal Sufficiency*: All variables are observed within the system X , i.e. there are no hidden confounders.
2. *Causal Faithfulness*:
 - (a) *Adjacency Faithfulness*: The conditional dependence relations implied by the edges present in the underlying causal graph hold in the joint probability distribution.

- (b) **Standard Faithfulness:** The conditional independence relations implied by the d -separation structures in the underlying causal graph hold in the joint probability distribution.
3. **Causal Markov Condition:** Each variable X_{bt} is independent of all non-descendants given its parents $\mathcal{P}(X_{bt})$.

Regarding the time series, stationarity of time-series is assumed, although non-stationary time-series can be evaluated with GP-based methods by modifying the kernel function (Cui and Djuric 2024). It is also assumed that the correct lag order m is selected for the model.

Furthermore, the proposed method assumes the absence of contemporaneous collider triples, i.e. $X_{at} \rightarrow X_{ct} \leftarrow X_{bt}$. It will be shown that spurious lagged parents due to conditioning on colliders within a lagged triple, i.e. $X_{a(t-\tau)} \rightarrow X_{ct} \leftarrow X_{bt}$, are identifiable and may be filtered, but the proposed method precludes consistent graph identification when fully contemporaneous colliders exist.

Lemma 1. *The Granger non-Causal relationship $X_a \not\rightarrow X_b$ identified by the GP_{SIC} method implies the conditional independence relationship $X_a \perp\!\!\!\perp X_b | X_{(t-m)}$, where $X_{(t-m)}$ includes all variables conditioned on in the GP_{SIC} model for each $b = 1, \dots, n_t$, up to lag m .*

Proof The connection between linear Granger Causality and the concept of transfer entropy can be understood based on statistical risk minimisation (Marinazzo et al. 2008). Consider the stochastic variables X_b and y , where both variables are recorded across $t = 1, \dots, n$ time instances. The regression function which minimises the risk functional $R[f] = \int p(X_b, y) (y - f(X_b))^2 dy dX_b$ is the conditional expectation of the current observation given its past values, $f^*(X_b) = \int p(y|X_b) y dy$. The inclusion of another variable X_a , which may be multivariate and include observations at time instant t that are contemporaneous to the observation y , yields the solution $g^*(X_b, X_a) = \int p(y|X_b, X_a) y dy$. The statement that X_a does not Granger cause X_b indicates that the inclusion of X_a in the regression function does not improve the prediction of y , such that $p(y|X_b) = p(y|X_a, X_b)$ and thus $f^*(X_b) = g^*(X_b, X_a)$. Transfer entropy evaluates whether $p(y|X_b) = p(y|X_b, X_a)$ via the assessment of the conditional mutual information between X_a and X_b ,

$$\begin{aligned} I(y; X_a | X_b) &= \int \int \int p(y, X_b, X_a) \log \left(\frac{p(y|X_b, X_a)}{p(y|X_b)} \right) dy dX_b dX_a, \\ &= \int \int \int p(y, X_b, X_a) \log \left(\frac{p(y, X_a | X_b)}{p(y|X_b) p(X_a | X_b)} \right) dy dX_b dX_a, \end{aligned}$$

which may be considered equivalent to the evaluation of the conditional independence relationship $y \perp\!\!\!\perp X_a | X_b$, which tests $p(y, X_a | X_b) = p(y|X_b) p(X_a | X_b)$ and thus whether $TE_{X_a \rightarrow X_b} = 0$ (Runge 2018). As Granger non-causality implies conditional independence, Granger causality implies conditional dependence.

As linear Granger Causality is equivalent to Kernel Granger Causality with a linear kernel (Marinazzo et al. 2008), Kernel Granger Causality with a suitable choice of nonlinear kernel generalises GC to the non-linear setting (Marinazzo et al. 2008). The regression equations $f^*(X_b)$ and $g^*(X_b, X_a)$ remain the solutions to the minimisation of the risk functional, and so a nonlinear GC relationship identified via KGC can also be interpreted in terms of conditional independence relations. When using kernel (ridge) regression, the solution to the risk-minimising function at one realisation, $f^*(X_{bt^*})$, with the addition of noise has the analytical form $y^*(X_{bt^*}) = k(X_{bt^*})^\top (K(X_b, X_b) + \lambda I_n)^{-1} \mathbf{y}$, where X_{bt^*} is a new test point, k and K are kernel functions, λ is the regularisation parameter, and $\mathbf{y}$ is the vector of n observations.

As the Bayesian approach to kernel regression, the solution to the risk-minimising function a GP regression model coincides with that in kernel regression, i.e. the expected value of the GP posterior predictive distribution,

$$\mathbb{E}[y^*(X_{bt^*}) | X_b] = k(X_{bt^*})^\top (K(X_b, X_b) + \sigma^2 I_n)^{-1} \mathbf{y},$$

where σ^2 is the variance of the noise. Under the assumption that the correct hyperparameters are learned when maximising the marginal-likelihood, which are faithful to the underlying SCM, we can draw an equivalence between a Granger non-causal relationship identified by GP_{SIC} and a conditional independence relationship between variables. $\square$

Lemma 2. *The set of estimated adjacencies $\hat{\mathcal{A}}(X_{bt})$, for each $b = 1, \dots, n_t$, identified as graph G' is a superset of the true adjacencies, containing exactly $\hat{\mathcal{A}}(X_{bt}) = \mathcal{P}^-(X_{bt}), \mathcal{F}^-(X_{bt}), \mathcal{A}_t(X_{bt})$, where $\mathcal{A}_t(X_{bt})$ are the true contemporaneous adjacencies and $\mathcal{P}^-(X_{bt})$ are the true lagged parents of X_{bt} . The remaining lagged conditions $X_{a(t-\tau)} \in \mathcal{F}^-(X_{bt})$ are, at most, the lagged parents of a collider that is contemporaneous to X_{bt} , i.e. $X_{a(t-\tau)} \rightarrow X_{ct} \leftarrow X_{bt}$.*

Proof Under Lemma 1 and the Adjacency faithfulness assumption, the adjacencies $X_{a(t-\tau)} - X_{bt}$ for $\tau \geq 0$ in G' returned by GP_{SIC} imply the corresponding conditional dependence relations $X_{a(t-\tau)} \not\perp\!\!\!\perp X_{bt} | S$, where $S = X \setminus X_{bt}$, such that S is a maximal conditioning set. We show for an arbitrary pair of variables $X_{a(t-\tau)}, X_{bt} \in X$ with $\tau \geq 0$:

$$(a) X_{a(t-\tau)} - X_{bt} \notin G' \implies X_{a(t-\tau)} \notin \mathcal{P}^-(X_{bt}), \mathcal{F}^-(X_{bt}), \mathcal{A}_t(X_{bt})$$

Assume we observe the absence of a lagged edge in G' , i.e. $\tau > 0$ and $X_{a(t-\tau)} \not\rightarrow X_{bt}$, resulting from Granger non-causality/the conditional independence relation $X_{a(t-\tau)} \perp\!\!\!\perp X_{bt} | S$. This implies $X_{a(t-\tau)} \notin \mathcal{P}^-(X_{bt})$ by Causal Markov condition. Furthermore, due to the maximal conditioning set S , we may condition on contemporaneous colliders, i.e. $X_{a(t-\tau)} \rightarrow X_{ct} \leftarrow X_{bt}$, where it may also be $c = a$. This opens the pathway from $X_{a(t-\tau)} \rightarrow X_{bt}$, such that $X_{a(t-\tau)} \in \mathcal{F}^-(X_{bt})$, corresponding to a spurious dependency $X_{a(t-\tau)} \not\perp\!\!\!\perp X_{bt} | S$. However, since we observe the contrary conditional independence relation, this implies that $X_{a(t-\tau)} \notin \mathcal{F}^-(X_{bt})$.

The remaining case to consider is the contemporaneous non-adjacencies. If GP_{SIC} returns a Granger non-causal relationship, this implies the conditional independence relation $X_{a(t-\tau)} \perp\!\!\!\perp X_{bt} | S$, and by adjacency faithfulness the absence of a contemporaneous edge in G' , i.e. $\tau = 0$ and $X_{a(t-\tau)} \not\rightarrow X_{bt}$. Thus, the lack of a contemporaneous adjacency returned in G' indicates that $X_{a(t-\tau)}$ is truly conditionally independent/not adjacent to X_{bt} , $X_{a(t-\tau)} \notin \mathcal{A}_t(X_{bt})$.

$$(b) X_{a(t-\tau)} - X_{bt} \in G' \implies X_{a(t-\tau)} \in \mathcal{P}^-(X_{bt}), \mathcal{F}^-(X_{bt}), \mathcal{A}_t(X_{bt})$$

By (a), the presence of an edge $X_{a(t-\tau)} - X_{bt} \in G'$ implies that $X_{a(t-\tau)}$ is in a superset of $\mathcal{P}^-(X_{bt})$, $\mathcal{F}^-(X_{bt})$, $\mathcal{A}_t(X_{bt})$. Under the assumptions of standard faithfulness and no fully instantaneous colliders in the SCM, we can proceed with the rules of d -separation. We begin simply with direct dependencies, i.e. $X_{a(t-\tau)} - X_{bt}$, for $\tau \geq 0$, are unblocked. This accounts for $X_{a(t-\tau)} \in \mathcal{P}^-(X_{bt})$ for $\tau > 0$. Similarly, for $\tau = 0$, a true parent or child relationship is unblocked, s.t. $X_{a(t-\tau)} \in \mathcal{A}_t(X_{bt})$.

Then, the crux of this first step is that conditioning on a collider (or descendant of a collider) unblocks the path between parent nodes. When conditioning on all contemporaneous variables, we may condition on a child of X_{bt} , $\mathcal{D}_t(X_{bt})$, which is a collider between a lagged node and another contemporaneous node, $X_{a(t-\tau)} \rightarrow \mathcal{D}_t(X_{bt}) \leftarrow X_{bt}$, opening the path $X_{a(t-\tau)} \rightarrow X_{bt}$. This is exactly the case where $X_{a(t-\tau)} \in \mathcal{F}^-(X_{bt})$. Conditioning on a descendant of a collider also opens the path $X_{a(t-\tau)} \rightarrow X_{bt}$, but as we condition on all nodes except X_{bt} , this effect is already considered by conditioning on the collider itself.

Now, to show that no other adjacencies are estimated, we consider that if a path p contains a mediator or confounder as an internal node on the path, then this path is blocked, even if a collider is also conditioned on. Time orientation restricts the structure of the causal graph such that for $\tau > 0$, any path $X_{a(t-\tau)} \rightarrow \dots \rightarrow \mathcal{D}_t(X_{bt}) \leftarrow X_{bt}$ is blocked by conditioning on the internal variables preceding $\mathcal{D}_t(X_{bt})$. The internal nodes are guaranteed mediators as all lagged links are oriented forward in time. The other possibility is a path such as $X_{a(t-\tau)} \rightarrow \mathcal{D}_t(X_{bt}) \leftarrow \dots \leftarrow X_{bt}$, where $\mathcal{D}_t(X_{bt})$ is a collider that is a non-child descendant of X_{bt} . The path $X_{a(t-\tau)} \rightarrow X_{bt}$ is still blocked by the internal variables between X_{bt} and $\mathcal{D}_t(X_{bt})$. As such, only true lagged parents or unmediated collider pathways are unblocked, corresponding to a conditional dependency $X_{a(t-\tau)} \not\perp\!\!\!\perp X_{bt} | S$.

Finally, we consider contemporaneous conditions. As we have assumed the absence of colliders within an instantaneous time point, no false pathways may be unblocked by $X_{at} \rightarrow \mathcal{D}_t(X_{bt}) \leftarrow X_{bt}$. This ensures that the only unblocked paths are direct adjacencies, as non-adjacent contemporaneous variables are mediated as a result of the maximal conditioning set, and there is no pathway backward through lagged conditions due to time-orientation. $\square$

Proof of Theorem 1

Proof By Lemma 2, we have shown that the adjacencies recovered in graph G' are a superset of the true adjacencies, where the extra conditions are at most $\mathcal{F}^-(X_{bt})$. The contemporaneous adjacencies were shown to be consistently estimated in G' , and so these are retained in the final graph G''' .

The extra conditions are filtered by fitting a second graph, G'' , which is estimated by conditioning only on the lagged parents identified in G' , i.e. G'' retests all lagged adjacencies $X_{a(t-\tau)} \rightarrow X_{bt}$ by assessing $X_{a(t-\tau)} \perp\!\!\!\perp X_{bt} | S$, where $S = \mathcal{P}^-(X_{bt}), \mathcal{F}^-(X_{bt})$. We show that any lagged adjacency retained in G'' is in the true parent set $\mathcal{P}^-(X_{bt})$, s.t. the final graph G''' is asymptotically consistent with the true adjacencies, $\mathcal{A}(X_{bt}) = \mathcal{P}^-(X_{bt}), \mathcal{A}_t(X_{bt})$. We show for an arbitrary pair of variables $X_{a(t-\tau)}, X_{bt} \in X$ with $\tau > 0$:

$$(a) X_{a(t-\tau)} - X_{bt} \notin G'' \implies X_{a(t-\tau)} \notin \mathcal{P}^-(X_{bt})$$

This follows similarly from Lemma 2 (a), in that the conditional independence relation $X_{a(t-\tau)} \perp\!\!\!\perp X_{bt} | S$ implies the absence of a lagged adjacency in G'' . By the Causal Markov condition, $X_{a(t-\tau)} \notin \mathcal{P}^-(X_{bt})$.

$$(b) X_{a(t-\tau)} - X_{bt} \in G'' \implies X_{a(t-\tau)} \in \mathcal{P}^-(X_{bt})$$

By (a), the edges in G'' are a superset of $\mathcal{P}^-(X_{bt})$. As we retest only lagged edges from G' , the only possible extra conditions are $\mathcal{F}^-(X_{bt})$. By Lemma 2, a condition in $\mathcal{F}^-(X_{bt})$ is only opened due to conditioning on a contemporaneous collider. Thus, by removing the contemporaneous conditions from S , we block these false dependencies. It follows that any edge left in G'' must be in the set of true parents $\mathcal{P}^-(X_{bt})$.

The contemporaneous adjacencies from G' and the lagged parents from G'' yield an estimated graph G''' that is consistent with the adjacencies in the underlying SCM. $\square$

Proof of Theorem 2

Proof Due to time-ordering constraints, any lagged edge is automatically oriented forward in time, i.e. $X_{a(t-\tau)} \rightarrow X_{bt}$. Given the assumption of stationarity, contemporaneous orientations are assumed to be consistent across $t = 1, \dots, n$.

The first orientation step orients colliders based on false edges that are identified in graph G' . Lemma 2 shows that the lagged parents identified in graph G' are at most the true parents $\mathcal{P}^-(X_{bt})$ or the false parents $\mathcal{F}^-(X_{bt})$ that are a result of conditioning on a collider, i.e. $X_{a(t-\tau)} \rightarrow X_{ct} \leftarrow X_{bt}$ for $\tau > 0$ and it may be $c = a$. These false parents are filtered in G'' , s.t. only the true parents remain in G''' . Under consistent identification of the true parents and false parents induced by the aforementioned collider structure, for an unshielded lagged triple $X_{a(t-\tau)} \rightarrow X_{ct} - X_{bt}$ where $X_{a(t-\tau)} \rightarrow X_{bt} \in G'$ and $X_{a(t-\tau)} \rightarrow X_{bt} \notin G'''$, we can identify X_{ct} as a collider, and orient $X_{ct} \leftarrow X_{bt}$. Under our assumptions and Lemma 2, all false parents due to colliders appear in G' and no additional false parents appear. Thus, any unshielded lagged triple $X_{a(t-\tau)} \rightarrow X_{ct} - X_{bt}$ in G''' , where $X_{a(t-\tau)} \rightarrow X_{bt} \notin G'$ and $X_{a(t-\tau)} \rightarrow X_{bt} \notin G'''$, implies X_{ct} is a non-collider, such that we can orient $X_{ct} \rightarrow X_{bt}$.

We note one case where orientation is not identifiable given this schema. In the case described by $X_{a(t-\tau)} \rightarrow X_{ct} - X_{bt}$ and $X_{a(t-\tau)} \rightarrow X_{dt} - X_{bt}$, where $c \neq d$, a false edge from $X_{a(t-\tau)} \rightarrow X_{bt}$ has ambiguous source. This is because it may be that either one or both lagged triples are colliders. Hence, we exclude these lagged triples from being used for orientation.

Finally, in the assumed SCM, there are no contemporaneous colliders. This allows us to orient any partially oriented fully contemporaneous triples $X_{at} \rightarrow X_{ct} - X_{bt}$ as $X_{ct} \rightarrow X_{bt}$ to avoid colliders. If the assumption of no contemporaneous colliders is relaxed by adding more tests to filter out false pathways, this orientation rule can no longer be applied.

The arrows oriented by the GP_{SIC} algorithm are asymptotically correct, but not necessarily complete. As we use a maximal conditioning set S , we cannot orient based on conditioning sets, and so lose power in some cases. For example, much of the power of this method comes from orienting contemporaneous edges based on lagged triples, including the use of autocorrelated edges. If there are not many lagged edges or strong, identifiable autocorrelation, we expect this method to lose power. $\square$

A.3 Supplementary Results Tables & Figures

Tables 5 and 6 show the average F1 score and standard deviation across 100 MC runs for each simulated experiment and for $n = \{250, 500\}$, coinciding with Figure 2. In Figure 5, we report the comparison between the algorithms performed by using the Bayesian Wilcoxon signed-rank test for the case $n_t = 250$. We include Figure 8 to show the slight degradation in performance of the GP_{SIC} method when using Student's t noise with 2 degrees of freedom across a selection of the simulated experiments, although we note that the method still performs comparatively to the other causal discovery methods. This method can be extended to accommodate non-Gaussian noise in further work.

Table 5 Average F1 score and standard deviation (reported in parentheses) across 100 MC simulations for the five models: GP_{SIC} , lsNGC, KGC, KPCR, and PCMCI. The name of the time series system and true lag order m are listed. Each simulation was tested on $n + m = 250$ samples.

Time Series System	GP_{SIC}	lsNGC	KGC	KPCR	PCMCI
1-way Logistic	0.98 (0.07)	0.99 (0.07)	0.99 (0.05)	0.98 (0.09)	0.24 (0.42)
2-way Logistic	0.69 (0.09)	0.71 (0.11)	0.68 (0.07)	0.70 (0.10)	0.41 (0.33)
Stochastic Linear	0.93 (0.14)	0.95 (0.12)	0.95 (0.12)	0.94 (0.13)	0.97 (0.10)
Stochastic Nonlinear	0.96 (0.14)	0.81 (0.39)	0.99 (0.05)	1.00 (0.03)	0.64 (0.27)
3 Fan-in	0.99 (0.05)	0.94 (0.12)	0.96 (0.10)	0.96 (0.11)	0.94 (0.11)
3 Fan-out	0.98 (0.07)	0.95 (0.11)	0.98 (0.07)	0.98 (0.07)	0.85 (0.15)
Confounder	0.97 (0.07)	0.94 (0.10)	0.97 (0.08)	0.96 (0.09)	0.95 (0.09)
Mediator	0.99 (0.05)	0.98 (0.07)	0.99 (0.04)	0.98 (0.06)	0.96 (0.08)
Redundant Collider	0.99 (0.04)	0.14 (0.28)	0.02 (0.11)	0.03 (0.13)	0.76 (0.27)
Synergistic Collider	0.97 (0.07)	0.70 (0.31)	0.98 (0.05)	0.99 (0.03)	0.92 (0.12)
Sync. (1-way intermediate)	0.96 (0.14)	0.93 (0.18)	0.90 (0.21)	0.90 (0.22)	0.48 (0.40)
Sync. (1-way strong)	0.67 (0.46)	0.24 (0.31)	0.24 (0.34)	0.26 (0.34)	0.62 (0.21)
Sync. (2-way strong)	0.96 (0.09)	0.92 (0.14)	0.92 (0.15)	0.89 (0.19)	0.82 (0.21)
Moran Effect	0.79 (0.05)	0.55 (0.11)	0.57 (0.08)	0.69 (0.10)	0.65 (0.07)
5 Linear	0.95 (0.06)	0.68 (0.09)	0.86 (0.10)	0.74 (0.12)	0.84 (0.08)
5 Nonlinear	0.97 (0.05)	0.52 (0.05)	0.96 (0.06)	0.99 (0.03)	0.77 (0.12)
8 Nonlinear	0.90 (0.13)	0.63 (0.14)	0.84 (0.21)	0.81 (0.23)	0.78 (0.11)
20 Nonlinear	0.73 (0.06)	0.50 (0.04)	0.47 (0.07)	0.65 (0.06)	0.69 (0.05)
30 Nonlinear	0.68 (0.06)	0.32 (0.05)	0.65 (0.14)	0.68 (0.08)	0.63 (0.03)

Table 6 Average F1 score and standard deviation (reported in parentheses) across 100 MC simulations for the five models: GP_{SIC} , lsNGC, KGC, KPCR, and PCMCI. The name of the time series system and true lag order m are listed. Each simulation was tested on $n + m = 500$ samples.

Time Series System	GP_{SIC}	lsNGC	KGC	KPCR	PCMCI
1-way Logistic	1.00 (0.00)	0.99 (0.07)	1.00 (0.03)	1.00 (0.03)	0.27 (0.44)
2-way Logistic	0.67 (0.00)	0.77 (0.15)	0.74 (0.14)	0.72 (0.12)	0.51 (0.32)
Stochastic Linear	0.97 (0.10)	0.97 (0.09)	1.00 (0.03)	0.99 (0.05)	0.99 (0.03)
Stochastic Nonlinear	0.95 (0.17)	0.80 (0.40)	0.99 (0.05)	1.00 (0.03)	0.68 (0.09)
3 Fan-in	1.00 (0.00)	0.84 (0.13)	0.71 (0.11)	0.95 (0.09)	0.97 (0.08)
3 Fan-out	1.00 (0.00)	0.66 (0.12)	0.61 (0.06)	0.97 (0.07)	0.73 (0.10)
Confounder	0.99 (0.04)	0.97 (0.08)	0.99 (0.03)	1.00 (0.03)	0.98 (0.04)
Mediator	0.99 (0.03)	0.97 (0.08)	0.99 (0.03)	0.99 (0.04)	0.96 (0.08)
Redundant Collider	0.99 (0.03)	0.41 (0.34)	0.00 (0.00)	0.01 (0.09)	0.76 (0.27)
Synergistic Collider	0.98 (0.06)	0.70 (0.36)	0.98 (0.06)	1.00 (0.02)	0.95 (0.09)
Sync. (1-way intermediate)	1.00 (0.00)	0.84 (0.19)	0.66 (0.22)	0.88 (0.18)	0.16 (0.19)
Sync. (1-way strong)	0.55 (0.50)	0.31 (0.04)	0.44 (0.10)	0.62 (0.10)	0.36 (0.06)
Sync. (2-way strong)	1.00 (0.00)	0.69 (0.13)	0.52 (0.03)	0.93 (0.14)	0.78 (0.13)
Moran Effect	0.82 (0.04)	0.58 (0.09)	0.74 (0.08)	0.80 (0.08)	0.68 (0.10)
5 Linear	0.95 (0.06)	0.59 (0.06)	0.98 (0.04)	0.96 (0.07)	0.82 (0.09)
5 Nonlinear	0.97 (0.05)	0.48 (0.04)	0.99 (0.03)	1.00 (0.02)	0.82 (0.09)
8 Nonlinear	1.00 (0.00)	0.62 (0.09)	0.99 (0.02)	1.00 (0.01)	0.81 (0.08)
20 Nonlinear	0.75 (0.08)	0.41 (0.03)	0.86 (0.04)	0.86 (0.04)	0.72 (0.04)
30 Nonlinear	0.74 (0.06)	0.24 (0.03)	0.95 (0.02)	0.82 (0.04)	0.65 (0.03)

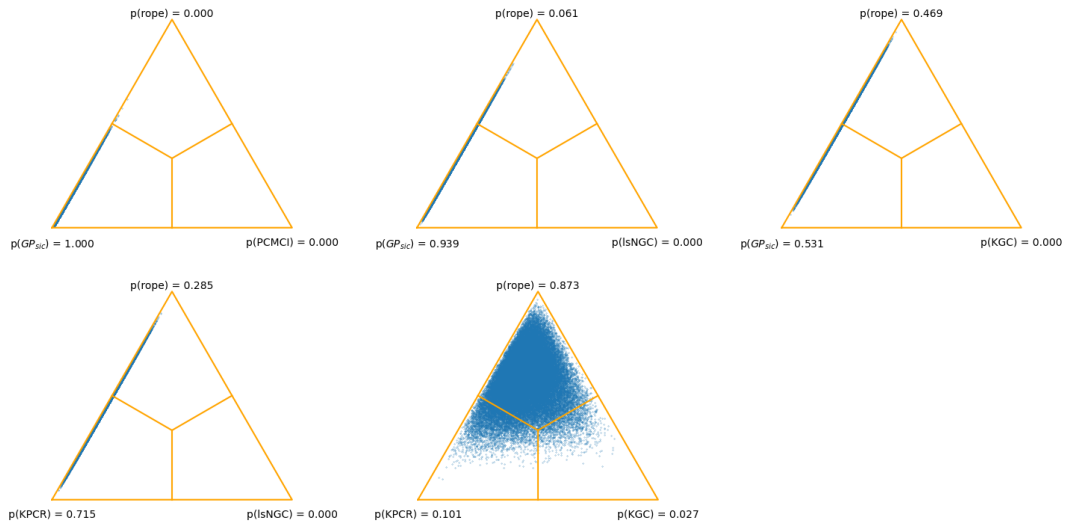


Fig. 5 $n = 250$ – For each pair of causal discovery methods (C_1, C_2) (e.g., $C_1 = \text{GP}_{SIC}$ and $C_2 = \text{PCMCI}$), we report the posterior probability vector that $C_1 > C_2$, $C_1 \equiv C_2$ and $C_1 < C_2$ denoted respectively as $[p(C_1), p(\text{rope}), p(C_2)]$, obtained via the Bayesian Wilcoxon signed-rank test. The cloud of points in the figure corresponds to samples from this posterior distribution: each point is a probability vector $[p(C_1), p(\text{rope}), p(C_2)]$ which we plot in the probability simplex.

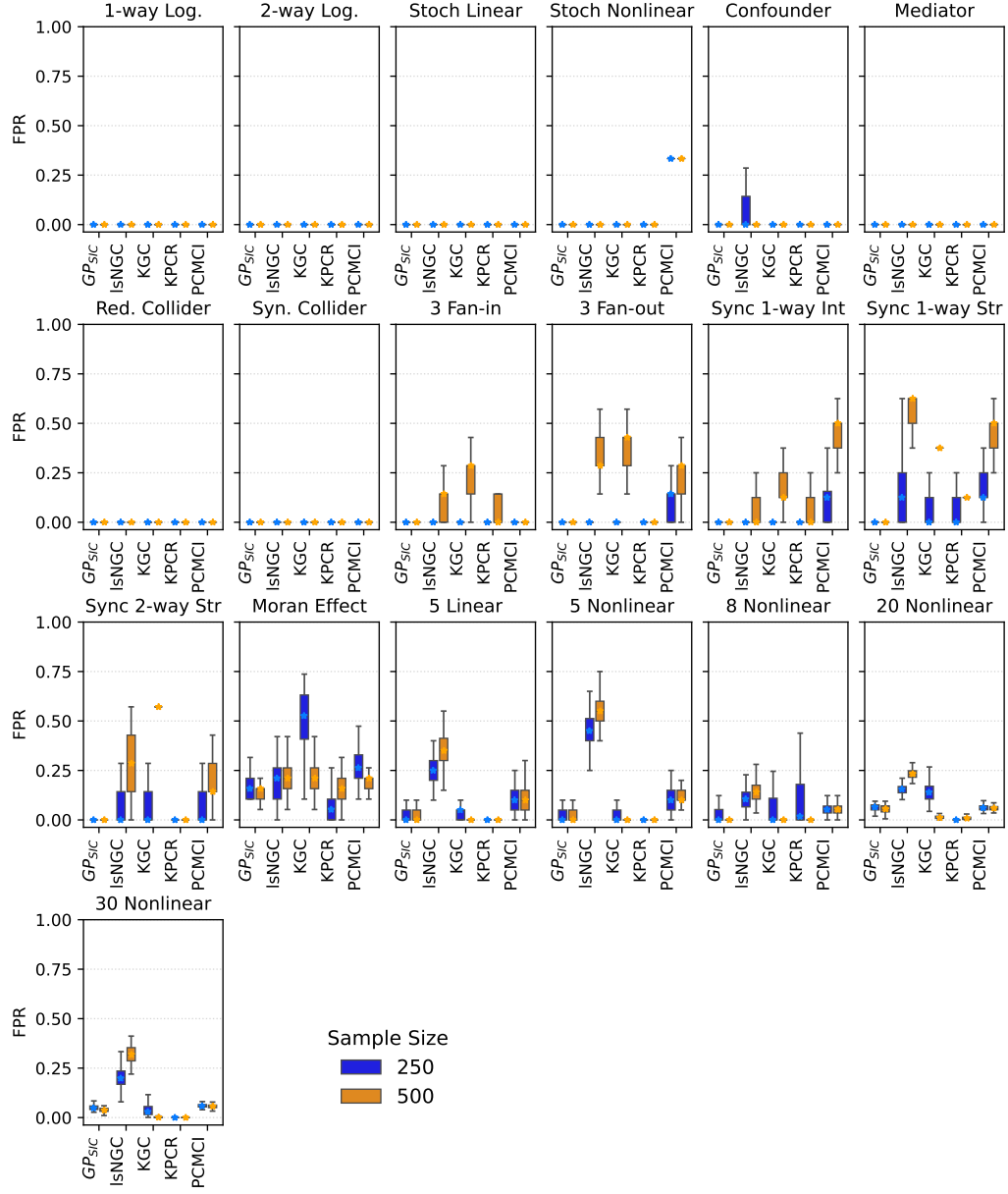


Fig. 6 False positive rate for five nonlinear causal discovery methods: GP_{SIC} , IsNGC, KGC, KPCR method, and PCMCI. 100 MC experiments were run for each simulated dataset for $n = 250, 500$. The median is marked with a *.

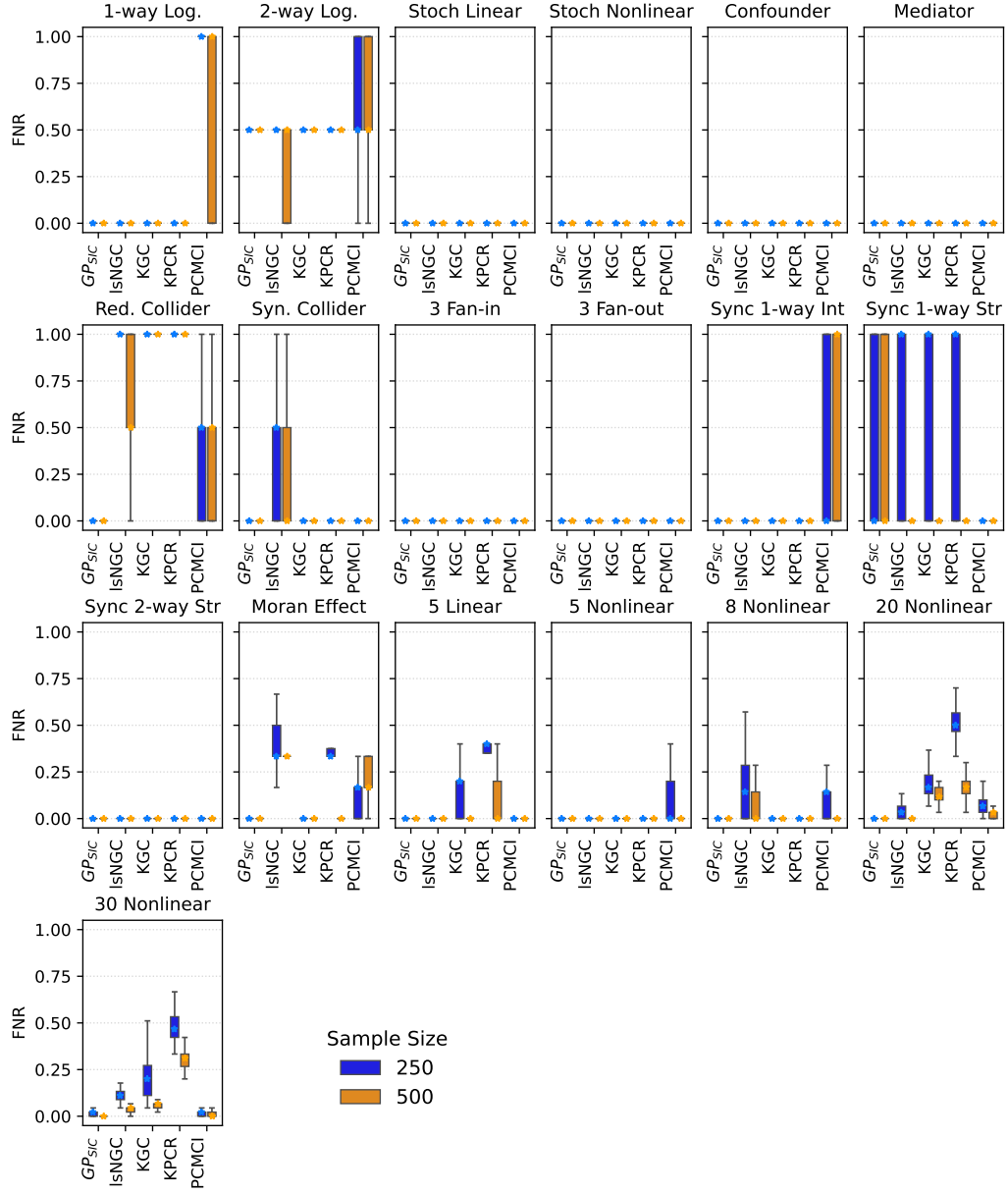


Fig. 7 False negative rate for five nonlinear causal discovery methods: GP_{SIC} , IsNGC, KGC, KPCR method, and PCMCI. 100 MC experiments were run for each simulated dataset for $n = 250, 500$. The median is marked with a $\star$.

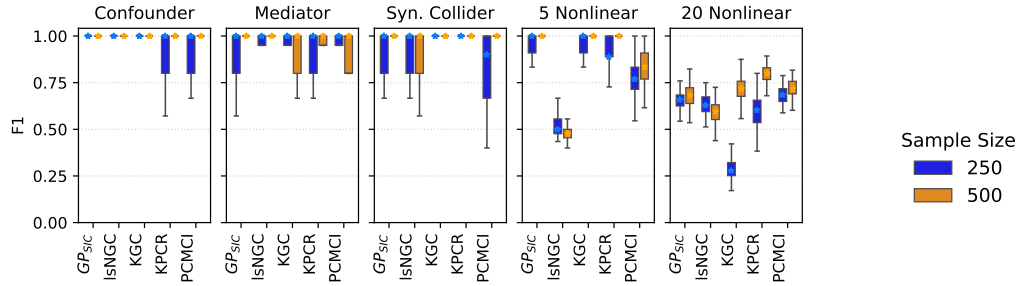


Fig. 8 F1 score for five nonlinear causal discovery methods, where the data was generated using Student's t noise with 2 degrees of freedom. 100 MC experiments were run for each simulated dataset for $n = 250, 500$. The median is marked with a $\star$.

A.4 Contemporaneous Algorithm Experiments

For contemporaneous adjacency and orientation identification, the PCMCi+ performs better across the experiments when autocorrelation strength a is lower, while the GP_{SIC} method performs better for higher autocorrelation. As a GC method the GP_{SIC} contemporaneous algorithm relies on a maximal conditioning set, and as such will have lower power in comparison to methods that use reduced conditioning sets. For contemporaneous orientation, this is especially true in cases where there is not strong, identifiable autocorrelation, as the GP_{SIC} method relies on orienting contemporaneous edges based on correctly identified lagged triples, which may include autocorrelated edges. When there is weak autocorrelation and autocorrelated edges go undetected, this limits the GP_{SIC} -based algorithm's ability to orient contemporaneous edges. However, the GP_{SIC} -based algorithm is fully defined within the GC framework, as it performs only $2n_t$ tests, compared to PCMCi+, which performs $O(n_t^2)$ tests.

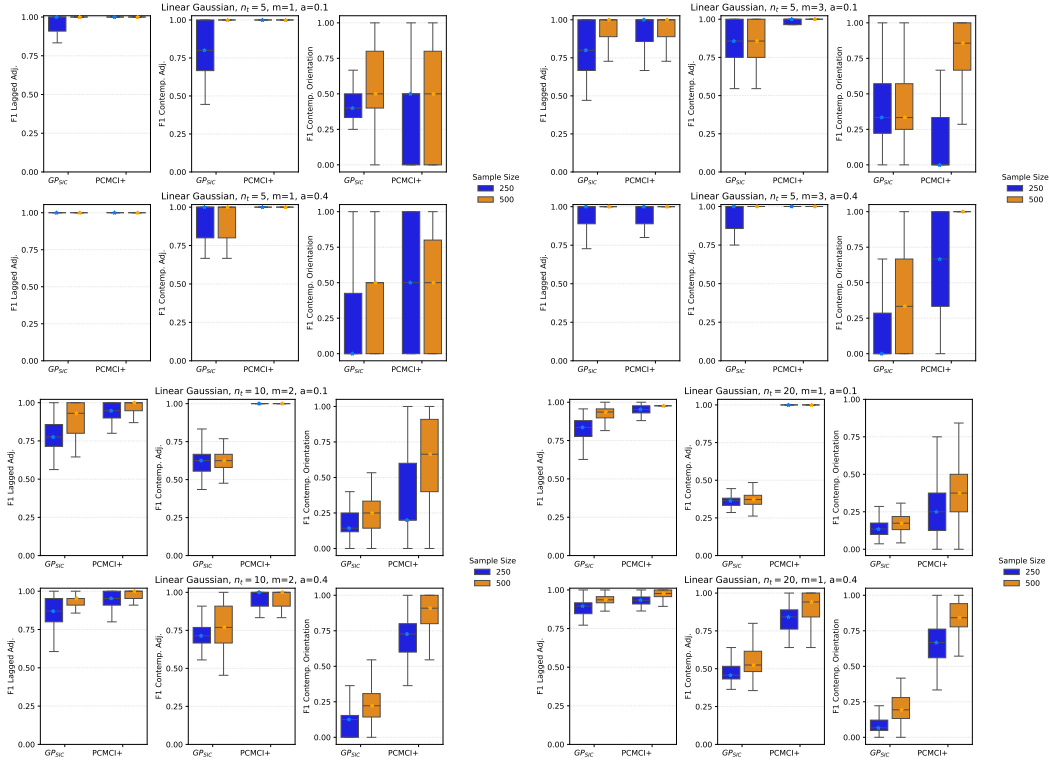


Fig. 9 F1 scores for the contemporaneous linear, Gaussian noise experiments.

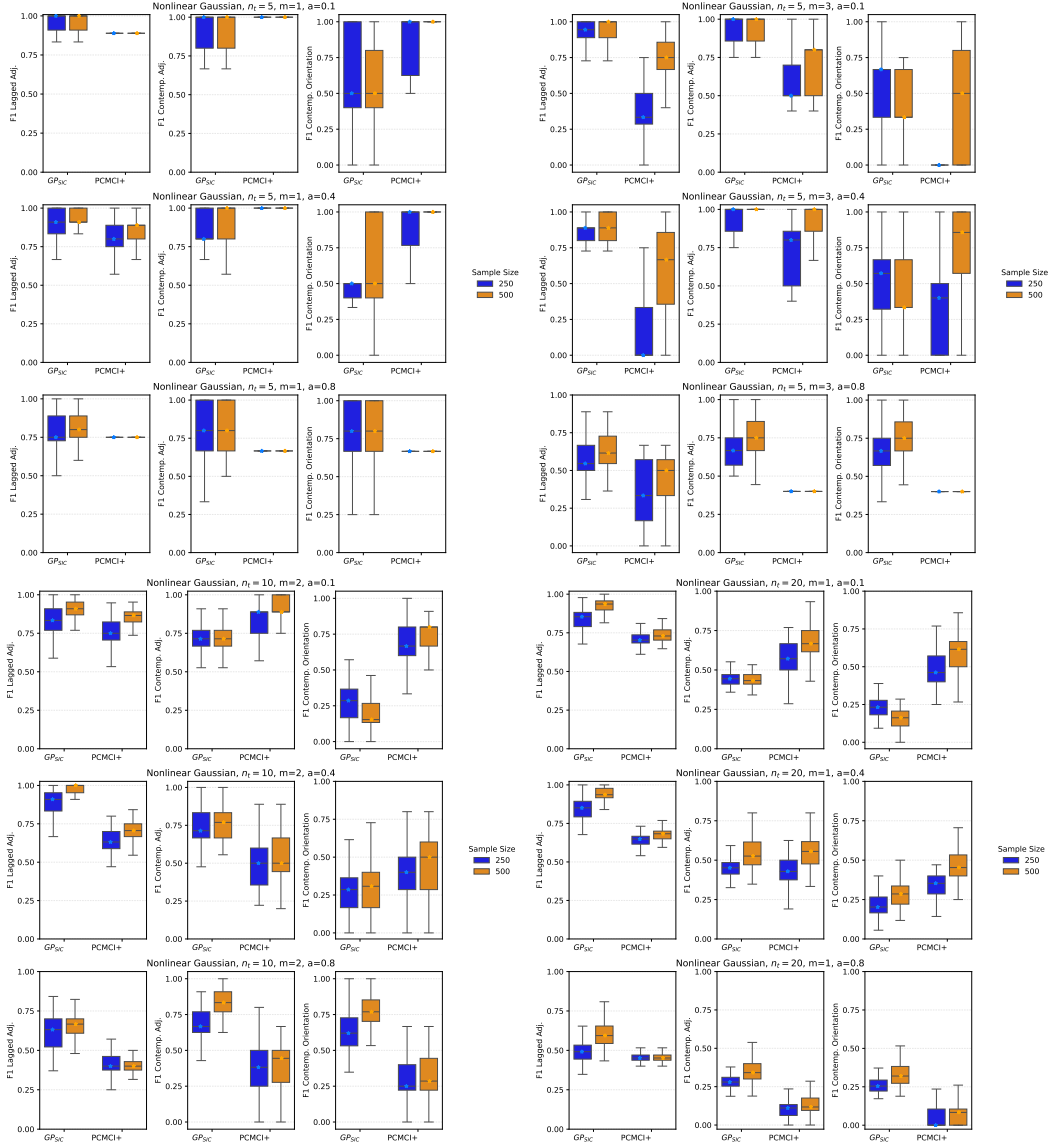


Fig. 10 F1 scores for the contemporaneous nonlinear, Gaussian noise simulated experiments.

pH Plant Application

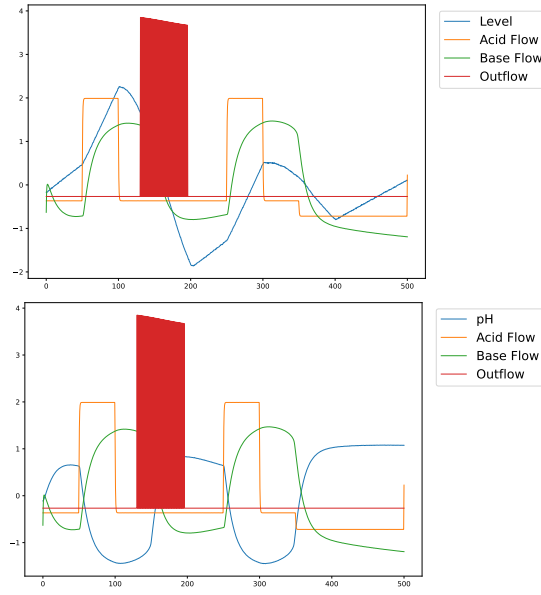


Fig. 11 The normalised time series data for the level interactions (top) and pH interactions (bottom) for one MC run in the regulatory mode.

A.5 FMRI Application

We additionally test the GP_{SIC} and PCMCi+ contemporaneous algorithms on a FMRI dataset from the work [Smith et al. \(2011\)](#), which simulates realistic FMRI data, measuring neuronal activity. The simulation model followed a mean neural lag of 50 ms, while the sampling rate of the data was 3 s. As a result, we expect there to be contemporaneous connections, and it is reported in [Smith et al. \(2011\)](#) that causal discovery methods that were strictly lag-based had poor performance. We tested GP_{SIC} and PCMCi+ on the Sim 7 dataset, using the time series $\{X_1, X_2, X_3, X_4\}$ and $n = 500$ data points, and having ground-truth causal edges $X_1 \rightarrow X_2, X_2 \rightarrow X_3, X_3 \rightarrow X_4$. There are 50 realisations of the dataset.

Figure 14 reports the overall F1 score and the F1 score for adjacencies. The results are consistent with the contemporaneous simulated experiments in that the methods have high F1 for detecting the causal adjacencies, i.e. the presence of a relationship between $X_1 - X_2, X_2 - X_3, X_3 - X_4$, where the GP_{SIC} outperforms the PCMCi+ method with lower variation. The overall F1 score is lower than the F1 for adjacencies due to decreased performance in contemporaneous orientation, which is consistent with the contemporaneous simulated experiments.

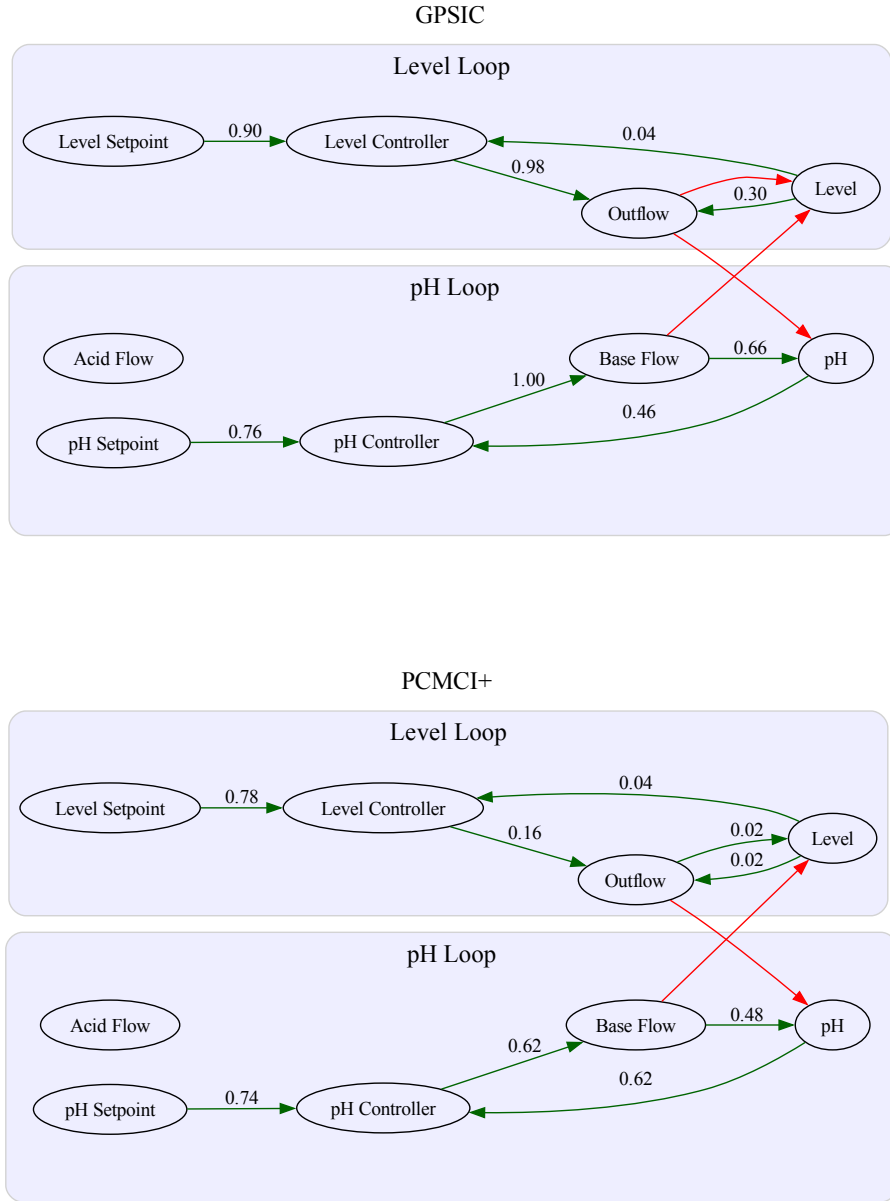


Fig. 12 The causal diagram recovered from the *GPSIC*-based algorithm and *PCMCI+* for the pH neutralisation plant data in *servo* mode. Green arrows correspond to true edges and are labeled with the rate of identified edges for 50 MC simulations. Red arrows correspond to false negatives. False positives were excluded from the diagrams for clarity.

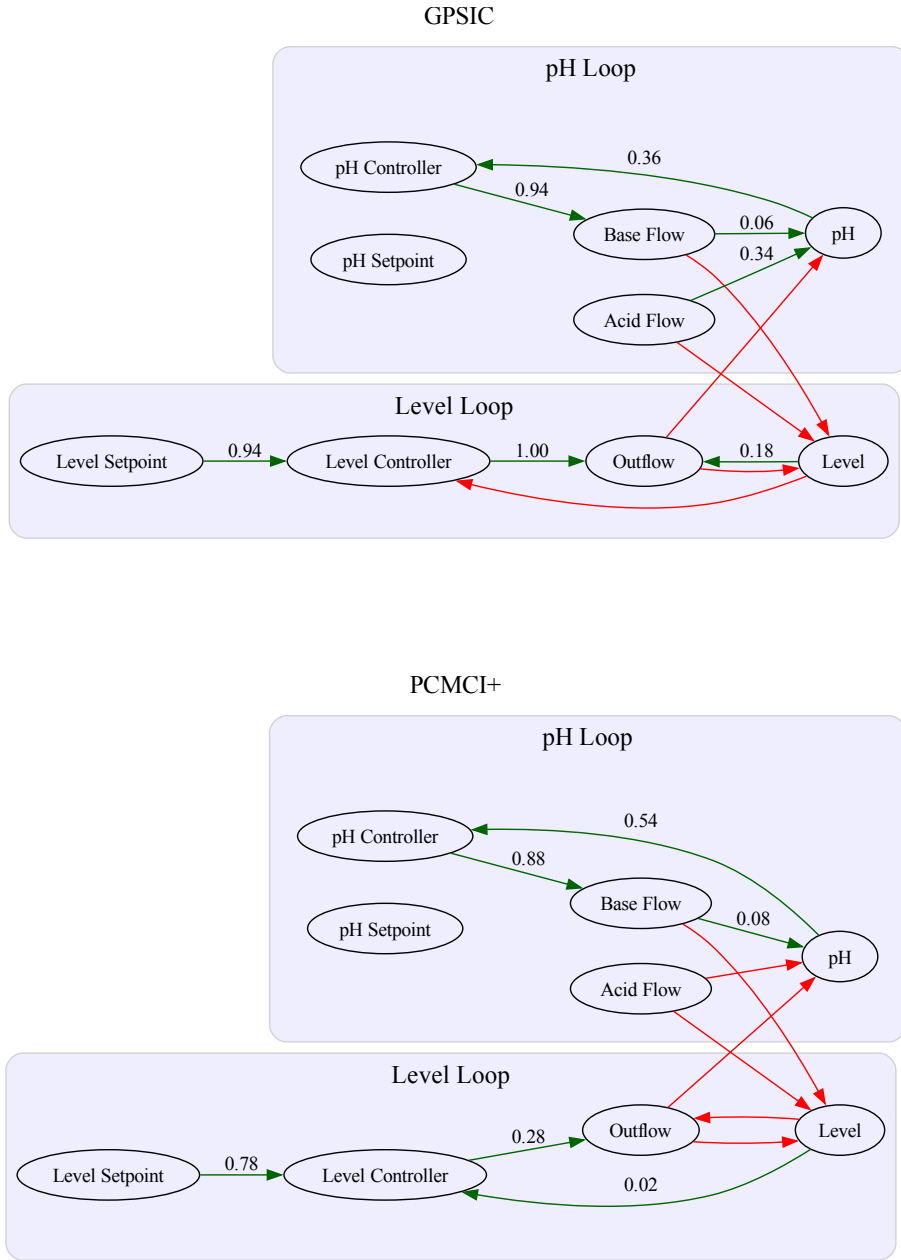


Fig. 13 The causal diagram recovered from the $GPSIC$ -based algorithm and PCMCi+ for the pH neutralisation plant data in *regulatory* mode. Green arrows correspond to true edges and are labeled with the rate of identified edges for 50 MC simulations. Red arrows correspond to false negatives. False positives were excluded from the diagrams for clarity.

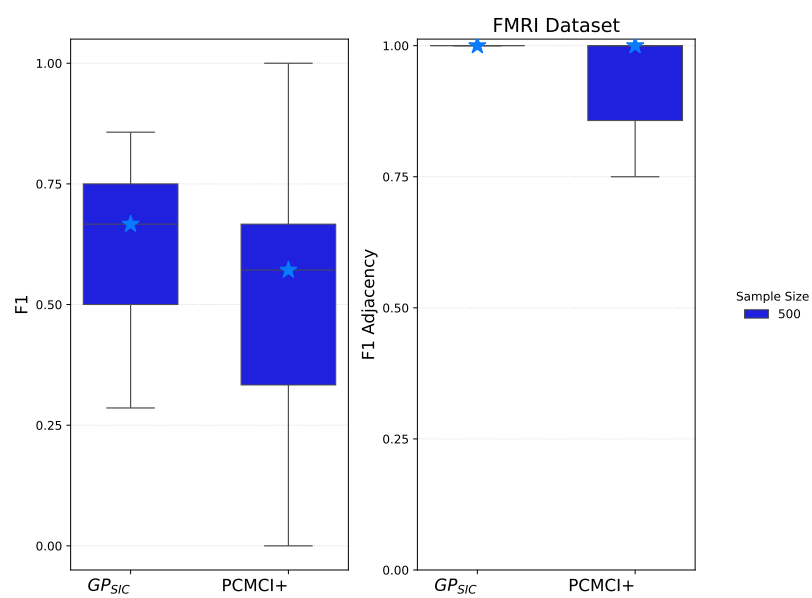


Fig. 14 Overall F1 score and F1 score for adjacencies for the GP_{SIC} and PCMCI+ contemporaneous causal methods on the FMRI dataset from [Smith et al. \(2011\)](#). The median is marked with a $\star$.

A.6 Method Implementations

LsNGC

The Large-Scale Nonlinear Granger Causality method was evaluated following the implementation described in (Wismüller et al. 2021). That is, we used False Discovery Rate correction for multiple tests, and the number of cluster centers was set to $c_f = 25$ and $c_g = 5$ for each simulated dataset.

KGC

The Kernel Granger Causality method was evaluated according to the implementation described in (Marinazzo et al. 2008). The method was evaluated with a Square Exponential kernel. The lengthscale for the SE kernel was set according to the heuristic $l = 2n_t m$.

PCMCI

The PCMCI and PCMCI+ methods, proposed in (Runge et al. 2019b; Runge 2020), were evaluated using the Tigramite library. There are several different Conditional Independence Tests that can be selected as the first phase of the PCMCI algorithm for causal identification. We used the Gaussian Process Distance Correlation (GPDC) conditional independence test, which is suitable for identifying a variety of nonlinear (and linear) causal relationships (Runge et al. 2019b). We used the default setting of no multiple testing correction for both PCMCI and PCMCI+.

KPCR and GP_{SIC}

We have made the code available for the KPCR and GP_{SIC} methods. We have also included our implementations for the GP-based GC methods (Amblard et al. 2012; Cui et al. 2022; Zaremba and Peters 2022). An analogous method to the GP method from Cui et al. (2022) was implemented by removing the SIC penalisation from the GP_{SIC} method. The GP GLRT Zaremba and Peters (2022) was implemented with a critical value $\alpha = 0.05$.

A.7 Simulation Setup

Two Variable Systems

The Logistic One-way and Logistic Two-way systems (Martínez-Sánchez et al. 2024; Sugihara et al. 2012) are defined by

$$x_1(t+1) = x_1(t)(r_1 - r_1 x_1(t) - \gamma_{1,2} x_2(t)) + \varepsilon(t), \quad (46)$$

$$x_2(t+1) = x_2(t)(r_2 - r_2 x_2(t) - \gamma_{2,1} x_1(t)) + \varepsilon(t), \quad (47)$$

where $\varepsilon(t)$ is additive Gaussian noise with variance 0.01, and the initial values were set to $x_1(0) = 0.2$ and $x_2(0) = 0.4$. The first 50 time points were excluded. The Logistic One-way system has the parameter values: $r_1 = 3.7$, $r_2 = 3.7$, $\gamma_{2,1} = 0.2$, $\gamma_{1,2} = 0$. The Logistic Two-way system has the parameter values: $r_1 = 3.5$, $r_2 = 3.9$, $\gamma_{2,1} = 0.2$, $\gamma_{1,2} = 0.01$.

The Stochastic Linear system (Martínez-Sánchez et al. 2024; Richardson 1922) is defined as

$$x_1(t+1) = 0.95\sqrt{2}x_1(t) - 0.9025x_1(t-1) + \varepsilon_1(t), \quad (48)$$

$$x_2(t+1) = 0.5x_1(t-1) + \varepsilon_2(t), \quad (49)$$

where $\varepsilon_i \sim N(0, 1)$. The Stochastic Nonlinear system (Martínez-Sánchez et al. 2024; Bueso et al. 2020) is defined as

$$x_1(t+1) = 3.4x_1(t)(1 - x_1(t)^2) \exp(-x_1(t-1)^2) + \varepsilon_1(t), \quad (50)$$

$$x_2(t+1) = 3.4x_2(t)(1 - x_2(t)^2) \exp(-x_2(t)^2) + \frac{x_1(t-1)x_2(t)}{2} + \varepsilon_2(t) \quad (51)$$

where $\varepsilon_i \sim N(0, 1)$.

Three Variable Systems

The Three Fan-in and Three Fan-out systems (Wismüller et al. 2021; Ma et al. 2014) are described by

$$x_j(t+1) = x_j(t)(\gamma_{jj} - \sum_{i=1,2,3} \gamma_{ji}x_i(t)), j = 1, 2, 3, \quad (52)$$

where the parameters for the fan-in system are $\gamma_{1,1} = 4, \gamma_{2,2} = 3.6, \gamma_{3,3} = 2.12, \gamma_{3,1} = 0.636, \gamma_{3,2} = -0.636$, and all others zero. The fan-out case uses the parameters $\gamma_{1,1} = 4, \gamma_{2,2} = 3.1, \gamma_{3,3} = 2.12, \gamma_{2,1} = 0.21, \gamma_{3,1} = -0.636$, and all others zero. We used initial conditions drawn from a uniform distribution on the interval $[0, 1)$ for the Fan-in and Fan-out examples. We discard the first 50 time points generated for each time series.

The Mediator, Confounder, Synergistic Collider, and Redundant Collider systems are described in (Martínez-Sánchez et al. 2024) as standard examples of different types of interactions that can occur in systems with more than two variables, where confounding effects may be present. The Mediator system is defined by

$$x_1(t+1) = \sin(x_2(t)) + 0.001\epsilon_1(t), \quad (53)$$

$$x_2(t+1) = \cos(x_3(t)) + 0.01\epsilon_2(t), \quad (54)$$

$$x_3(t+1) = 0.5x_3(t) + 0.1\epsilon_3(t). \quad (55)$$

The Confounder system is defined by

$$x_1(t+1) = \sin(x_1(t) + x_3(t)) + 0.01\epsilon_1(t), \quad (56)$$

$$x_2(t+1) = \cos(x_2(t) - x_3(t)) + 0.01\epsilon_2(t), \quad (57)$$

$$x_3(t+1) = 0.5x_3(t) + 0.1\epsilon_3(t). \quad (58)$$

The Synergistic Collider system is defined by

$$x_1(t+1) = \sin(x_2(t)x_3(t)) + 0.001\epsilon_1(t), \quad (59)$$

$$x_2(t+1) = 0.5x_2(t) + 0.1\epsilon_2(t), \quad (60)$$

$$x_3(t+1) = 0.5x_3(t) + 0.1\epsilon_3(t). \quad (61)$$

The Redundant Collider system is defined by

$$x_1(t+1) = 0.3x_1(t) + \sin(x_2(t)x_3(t)) + 0.001\epsilon_1(t), \quad (62)$$

$$x_2(t+1) = 0.5x_2(t) + 0.1\epsilon_2(t), \quad (63)$$

$$x_3(t+1) \equiv x_2(t+1). \quad (64)$$

In each example, $\epsilon_i(t) \sim N(0, 1)$, and the first 50 time points were excluded.

The Synchronization system (Martínez-Sánchez et al. 2024) with various coupling scenarios is defined by

$$x_1(t+1) = r_1x_1(t)(1 - x_1(t)) + \epsilon_1(t), \quad (65)$$

$$x_2(t+1) = r_2\left(\frac{x_2(t) + c_{1,2}x_1(t)}{1 + c_{1,2}}\right)\left(1 - \frac{x_2(t) + c_{1,2}x_1(t)}{1 + c_{1,2}}\right) + \epsilon_2(t), \quad (66)$$

$$x_3(t+1) = r_3\left(\frac{x_3(t) + c_{1,2,3}x_1(t) + c_{1,2,3}x_2(t)}{1 + 2c_{1,2,3}}\right)\left(1 - \frac{x_3(t) + c_{1,2,3}x_1(t) + c_{1,2,3}x_2(t)}{1 + 2c_{1,2,3}}\right) + \epsilon_3(t). \quad (67)$$

The parameters $r_1 = 3.68, r_2 = 3.67, r_3 = 3.78$ were used for every coupling scenario. The coupling parameters were varied by scenario: One-way intermediate uses $c_{1,2} = 0.1, c_{1,2,3} = 0$, One-way strong uses $c_{1,2} = 1, c_{1,2,3} = 0$ and Two-way strong uses $c_{1,2} = 0, c_{1,2,3} = 1$. The noise terms are independently sampled from $\epsilon_i(t) \sim N(0, 1e-5)$.

Moran-Effect Model

The Moran Effect (Moran 1953; Martínez-Sánchez et al. 2024) model is defined by the equations

$$R_i(t+1) = r_i N_i(t)(1 - N_i(t))e^{-\psi_i V(t)}, \quad (68)$$

$$N_i(t+1) = s_i N_i(t) + \max(R_i(t - D_i), 0), \quad (69)$$

for $i = 1, 2$. N_i are the variables of the system, R_i are mediators of a shared external influence $V \sim N(0, 1)$. Thus, while correlated, the variables N_1 and N_2 maintain causal independence, representing the ecological phenomenon of simultaneous changes in the sizes of two non-interacting populations due to a shared environmental factor. The parameters we used to simulate the data are as follows: $r_1 = 3.4, r_2 = 2.9, s_1 = 0.4, s_2 = 0.35, D_i = 4, \psi_1 = 0.5, \psi_2 = 0.6$. The initial value for both R_i was 1, and for both N_i was 0.5. We drop out the first 50 generated time points.

Five Variable Systems

The Five Linear system (Wismüller et al. 2021; Baccalá and Sameshima 2001) is defined as

$$x_1(t) = x_1(t) + 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \sigma\epsilon(t), \quad (70)$$

$$x_2(t) = x_2(t) + 0.5x_1(t-2) + \sigma\epsilon(t), \quad (71)$$

$$x_3(t) = -0.4x_1(t-3) + \sigma\epsilon(t), \quad (72)$$

$$x_4(t) = x_4(t) - 0.5x_1(t-2) + 0.5\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \sigma\epsilon(t), \quad (73)$$

$$x_5(t) = x_5(t) - 0.5\sqrt{2}x_4(t-1) + 0.5\sqrt{2}x_5(t-1) + \sigma\epsilon(t), \quad (74)$$

where $\epsilon(t) \sim N(0, 1)$ and $\sigma = 0.01$. The Five Nonlinear system (Wismüller et al. 2021; Baccalá and Sameshima 2001) incorporates nonlinear dependencies via

$$x_1(t) = x_1(t) + 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \sigma\epsilon(t), \quad (75)$$

$$x_2(t) = x_2(t) + 0.5x_1^2(t-2) + \sigma\epsilon(t), \quad (76)$$

$$x_3(t) = -0.4x_1(t-3) + \sigma\epsilon(t), \quad (77)$$

$$x_4(t) = x_4(t) - 0.5x_1^2(t-2) + 0.5\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \sigma\epsilon(t), \quad (78)$$

$$x_5(t) = x_5(t) - 0.5\sqrt{2}x_4(t-1) + 0.5\sqrt{2}x_5(t-1) + \sigma\epsilon(t), \quad (79)$$

where $\epsilon(t) \sim N(0, 1)$ and $\sigma = 0.01$. We dropped out the first 50 time points that were generated, and used initial conditions drawn from the standard normal distribution $N(0, 1)$.

Eight Variable System

We adapt the Eight Species system from (Martínez-Sánchez et al. 2024; Leng et al. 2020) via the model

$$x_1(t+1) = x_1(t)(3.9 - 3.9x_1(t)) + \epsilon_1(t), \quad (80)$$

$$x_2(t+1) = x_2(t)(3.5 - 3.5x_2(t)) + \epsilon_2(t), \quad (81)$$

$$x_3(t+1) = x_3(t)(3.62 - 3.62x_3(t) - 0.35x_1(t) - 0.35x_2(t)) + \epsilon_3(t), \quad (82)$$

$$x_4(t+1) = x_4(t)(3.75 - 3.75x_4(t) - 0.35x_2(t)) + \epsilon_4(t), \quad (83)$$

$$x_5(t+1) = x_5(t)(3.65 - 3.65x_5(t) - 0.35x_3(t)) + \epsilon_5(t), \quad (84)$$

$$x_6(t+1) = x_6(t)(3.72 - 3.72x_6(t) - 0.35x_3(t)) + \epsilon_6(t), \quad (85)$$

$$x_7(t+1) = x_7(t)(3.57 - 3.57x_7(t) - 0.35x_6(t)) + \epsilon_7(t), \quad (86)$$

$$x_8(t+1) = x_8(t)(3.68 - 3.68x_8(t) - 0.35x_6(t)) + \epsilon_8(t), \quad (87)$$

where $\epsilon_i(t) \sim N(0, 0.005)$. The first 50 time points are dropped out.

20 and 30 Nonlinear Systems

The 20 and 30 node nonlinear systems were simulated according to the model

$$x_j(t+1) = ax_j(t) + \sum_i cf(x_i(t - \tau_i)) + \varepsilon_j(t), \varepsilon_j(t) \sim N(0, 1), \quad (88)$$

where $a = 0.4, c = 0.4$ and $f(x) = (1 + 5xe^{-\frac{x^2}{20}})x$. The number of edges in the true causal graph was selected according to $L = \lfloor 1.5n_t \rfloor$, and edges were randomly assigned.

Contemporaneous Systems

The Contemporaneous systems were simulated in a similar manner to the numerical experiments in [Runge \(2020\)](#). We simulated the data according to the model

$$x_j(t+1) = ax_j(t) + \sum_i cf(x_i(t - \tau_i)) + \varepsilon_j(t), \varepsilon_j(t) \sim N(0, 1), \quad (89)$$

where the autocorrelation and causal magnitude parameters were selected as either $(a, c) = (0.4, 0.4)$ or $(a, c) = (0.8, 0.4)$ for the nonlinear examples to assess the effect of varying degrees of autocorrelation. For the linear case with $n_t = 5, m = 3$, we tested $a = \{0.4, 0.5\}$. The nonlinear function used for the contemporaneous systems was $f(x) = (1 + 5xe^{-\frac{x^2}{20}})$. The number of edges in the true causal graph was selected according to $L = \lfloor 1.5n_t \rfloor$, with 30% contemporaneous edges, and edges were randomly assigned.