CAUSAL PARTIAL SEPARABILITY: LEARNING NONLINEAR DAGS FROM MULTIVARIATE FUNCTIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In functional data analysis, learning directed acyclic graphs (DAGs) from multivariate functional data is crucial for identifying root causes. While recent causal discovery methods for multivariate functional data primarily rely on linear models, there is a lack of nonlinear acyclic models tailored for such data. The fundamental differences between scalars and functions prevent existing scalar nonlinear causal discovery methods from being directly applicable to functional data. To address these limitations, we propose Causal Partial Separability (CPS), a framework where functional variables admit a shared spectral basis and a common causal ordering, thereby mapping the problem to scalar DAG learning across the shared basis. We analyze the equivalence between functional DAGs and the union of blockwise DAGs, deriving the nonlinear identifiability of functional DAGs under the assumption of nonlinear Additive Noise Models (ANMs). Building on CPS, we introduce cps-fDAG, a functional DAG estimation method that combines shared spectral basis learning, searching for a common causal ordering, and hierarchical edge pruning. We establish the theoretical consistency for both the common ordering and edge pruning and provide a sufficient sample size condition. Experiments on synthetic datasets and biophysical fMRI simulation benchmarks demonstrate that cps-fDAG effectively learns causal DAGs for multivariate functional data.

1 INTRODUCTION

Multivariate functional data are becoming increasingly prevalent across various scientific fields, including neuroscience (Zipunnikov et al., 2011; Aljobouri, 2023), genomics (Pickrell, 2014), climate studies (Berrendero et al., 2011b), finance (Cai, 2018; Frois Caldeira et al., 2021), and industrial systems (Jin et al., 2024). The variable is observed as a stochastic process $X_j = \{X_j(t) : t \in \mathcal{T}\}$, rather than as a scalar (Ramsay & Silverman, 2005). Different from estimating the dynamic graphs corresponding to scalars (Runge et al., 2019; Pamfil et al., 2020), inferring the underlying static directed acyclic graph (DAG) among these functions is essential for identifying root causes. In recent years, several methods have been developed for learning DAGs from multivariate functional data. These include functional directed graphical models (DGMs) (Gómez et al., 2020), functional structural equation models (SEMs) (Lee & Li, 2022), Func-LiNGAM (Yang & Suzuki, 2022), functional Bayesian networks (Zhou et al., 2023), directed cyclic graphs for functional data (Roy et al., 2023), functional DAGs (Lee et al., 2024), and MultiFun-DAG (Lan et al., 2024). However, these approaches primarily assume linear relations, which makes them less effective at identifying nonlinear root causes in multivariate functional data.

Significant progress has been made in scalar nonlinear causal discovery (Monti et al., 2020; Zheng et al., 2020), however, directly extending these methods to functional variables poses some limitations. This is due to the fundamental differences between functions and scalars, particularly as the dimension transitions from finite to infinite. Dimension reduction is a common strategy in functional data analysis. For example, one may simply use the mean of each variable sampled across the time domain or apply the functional principal component analysis (FPCA) (Berrendero et al., 2011a; Chiou et al., 2014) for each function separately. However, using the mean to construct a scalar DAG learning model results in a severe information loss from the functional data (Wahl et al., 2024). When applying FPCA to each function, it generates incompatible basis coordinates and re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

quires computationally intensive expenses to accurately characterize the relations between variables across different coordinates (Qiao et al., 2019).

To address these limitations, we propose *Causal Partial Separability* (CPS), a causal framework for nonlinear DAG learning from multivariate functional data. CPS creates a shared spectral basis that structurally decouples functions into equivalent shared scores (Zapata et al., 2022). The DAGs learned by scores possess a common causal ordering. We analyze the functional DAG as equivalent to the union of blockwise DAGs by $\mathcal{G} = \bigcup_{\ell=1}^{\infty} \mathcal{G}_\ell$. CPS assumes that the bivariate relation under each score follows a nonlinear Additive Noise Model (ANM) (Hoyer et al., 2008), and the multivariate relations are additive. Under this assumption, each blockwise graph direction becomes identifiable, consequently making the functional DAG identifiable as well.

We present cps-fDAG, a method for estimating functional DAGs within the Causal Partial Separability framework. First, cps-fDAG learns a shared basis from the joint covariance estimation of the observed data, which allow it to decouple functions into equivalent shared scores. Next, it searches for a common causal ordering using a kernel residual independence criterion formulated in the reproducing kernel Hilbert space (RKHS) (Berlinet & Thomas-Agnan, 2011). cps-fDAG aggregates blockwise evidence to determine whether the candidate residuals are consistent with the nonlinear ANMs. Finally, it performs edge pruning based on the estimated common causal ordering to select parent nodes from their predecessor nodes, thereby constructing an estimated DAG. In addition, we establish the theoretical consistency of both the common ordering and edge pruning, providing the sufficient sample size condition. We evaluate cps-fDAG using both synthetic datasets and biophysical fMRI simulation benchmarks. The experimental results demonstrate that cps-fDAG effectively learns causal DAGs for multivariate functional data. We summarize our contributions as follows.

- We propose the *Causal Partial Separability* (CPS) framework, which facilitates the learning of nonlinear functional DAGs by decomposing it into a series of blockwise scalar DAGs learning problems.
- We present a new estimation method, cps-fDAG, to estimate the functional DAG within the CPS framework. We also establish complete consistency theories and sufficient sample conditions to confirm the reliability of our method.
- We conduct extensive experiments using synthetic datasets and biophysical fMRI simulation benchmarks, comparing functional and scalar methods and demonstrating the effectiveness of our method, yielding more accurate results.

2 Related Work

Functional causal discovery. Func-LiNGAM (Yang & Suzuki, 2022) extends linear non-Gaussian acyclic model to functions. The functional Bayesian network of Zhou et al. (2023) similarly leverages the linear non-Gaussian property for identifiability, while explicitly accounting for noisy functional observations through the Bayesian model. Roy et al. (2023) consider a functional linear structural equation model (SEM) that permits directed cyclic graphs and introduce a causal embedding. MultiFun-DAG (Lan et al., 2024) models relations between multivariate functional variables through hidden bilinear functional regressions and learns an acyclic graph using score-based optimization. Lee & Li (2022) handles functional SEM with the score function at the linear operator level. Zhang & Estrada Gómez (2025) further extend DGMs to heterogeneous multimodal systems with scalar and functional nodes, but their structure learning procedure relies on a prespecified topological ordering. These methods primarily impose linear relations, thus failing to recognize the causal structure under the nonlinear functional relations.

Scalar Nonlinear DAG learning. For scalar variables, nonlinear ANMs are generally identifiable through the independence between regression residuals and candidate causes (Hoyer et al., 2008; Peters et al., 2014; Mooij et al., 2016). CAM combines causal ordering search with sparse additive regression (Bühlmann et al., 2014), whereas nonlinear independent component analysis (ICA) exploits nonstationarity to handle more complex nonlinear mechanisms (Monti et al., 2020). Differentiable estimations use nonlinear structural equations with basis expansions or neural networks (Zheng et al., 2020; Lachapelle et al., 2020; Yu et al., 2019), and score matching methods recover directions under Gaussian or arbitrary additive noise (Rolland et al., 2022; Montagna et al., 2023).

(Göbler et al., 2025) is further extended to grouped data, assuming that the grouped coordinates are observable and are essentially aligned. All of these methods rely on fixed dimensional representations and thus cannot be directly applied to functional variables.

Functional dimension reduction. Multivariate FPCA captures the component variation through vector eigenfunctions (Chiou et al., 2014; Happ & Greven, 2018), but does not produce aligned scalar spectral blocks with an associated graph at each mode. Undirected functional graphical models encode conditional dependence through Gaussian process or truncated score representations (Zhu et al., 2016; Qiao et al., 2019). Partial separability yields a shared basis and a sequence of fixed-dimensional graphical models (Zapata et al., 2022). However, these works merely provide approaches to functional data analysis and do not establish the directed identifiability for multivariate functional data.

Our proposed Causal Partial Separability (CPS) is a unified framework that bridges these three lines of research. Unlike existing functional causal discovery methods, CPS allows nonlinear DAG learning from multivariate functional data. While scalar nonlinear DAG learning methods focus solely on scalar data, CPS extends the nonlinear ANM approaches to functional variables. Instead of relying on existing functional dimension reduction, CPS establishes a shared spectral basis, facilitating the mapping of functional DAG learning to scalar DAGs learning across this shared basis.

3 Causal Partial Separability

We consider a collection of p functional variables $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^\top$, where each $X_j \in L^2(\mathcal{T})$ is a square integrable stochastic process defined on a compact domain $\mathcal{T} \subset \mathbb{R}$. The space $L^2(\mathcal{T})$ is endowed with the inner product $\langle u, v\rangle = \int_{\mathcal{T}} u(t)v(t)\mathrm{d}t$.

Definition 1. *The bivariate functional causal mechanism is a nonlinear Additive Noise Model (ANM) and the multivariate function relations are additive. Let $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ be a Directed Acyclic Graph (DAG) where $\mathcal{V} = \{1, \ldots, p\}$. For each node $j \in \mathcal{V}$, the process $X_j(t)$ is generated by:*

$$X_j(t) = \sum_{k \in \mathrm{PA}_{\mathcal{G}_0}(j)} \mathcal{F}_{jk}(X_k)(t) + E_j(t), \quad \forall t \in \mathcal{T}, \tag{1}$$

where $\mathrm{PA}_{\mathcal{G}_0}(j)$ denotes the parent set of j in $\mathcal{G}_0$, the nonlinear functional operator $\mathcal{F}_{jk} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ is also additive with the expansion of $X_j(t)$, and $E_j(t)$ is the mutually independent Gaussian random error.

A multivariate functional causal system $\mathbf{X}(t)$ can be expanded in terms of an orthonormal basis $\{\phi_\ell(t)\}_{\ell=1}^\infty$ of $L^2(\mathcal{T})$, such that each process $X_j(t)$ and its corresponding error $E_j(t)$ admit the expansion:

$$X_j(t) = \sum_{\ell=1}^{\infty} \theta_{j\ell}\phi_\ell(t), \quad E_j(t) = \sum_{\ell=1}^{\infty} \epsilon_{j\ell}\phi_\ell(t), \tag{2}$$

where $\theta_{j\ell} = \langle X_j, \phi_\ell\rangle$ denotes the corresponding coordinate score, and $\epsilon_{j\ell} = \langle E_j, \phi_\ell\rangle$ denotes the error coefficient, with $\{\epsilon_{j\ell}\}_{j,\ell}$ being mutually independent Gaussian random variables.

Definition 2. *A multivariate functional causal system $\mathbf{X}(t)$ satisfies Causal Partial Separability if:*

1. *The coordinate scores $\{\theta_{\cdot\ell}\}_{\ell\geq 1}$ and exogenous Gaussian noises $\{\epsilon_{\cdot\ell}\}_{\ell\geq 1}$ are mutually independent across spectral blocks, such that*

$$\mathrm{Cov}\big(\langle X_j, \phi_\ell\rangle, \langle X_k, \phi_m\rangle\big) = 0, \qquad j, k \in \mathcal{V},\ \ell \neq m. \tag{3}$$

2. *For each spectral index $\ell \geq 1$, the vector of ℓ-th coordinate scores $\boldsymbol{\theta}_{\cdot\ell} = (\theta_{1\ell}, \ldots, \theta_{p\ell})^\top \in \mathbb{R}^p$ is governed by a functional additive restricted nonlinear ANM:*

$$\theta_{j\ell} = \sum_{k \in \mathrm{PA}_{\mathcal{G}_0}(j)} f_{jk\ell}(\theta_{k\ell}) + \epsilon_{j\ell}, \quad \forall j \in \mathcal{V}, \tag{4}$$

where $f_{jk\ell} : \mathbb{R} \to \mathbb{R}$ is a nonlinear function and the operator $\mathcal{F}_{jk}$ is expanded as

$$\mathcal{F}_{jk}(X_k)(t) = \sum_{\ell=1}^{\infty} f_{jk\ell}(\theta_{k\ell})\,\phi_\ell(t). \tag{5}$$

3. *The blockwise graphs $\{\mathcal{G}_\ell\}_{\ell=1}^{\infty}$ follow a shared topological ordering σ_0; for any $\ell \geq 1$, if $k \in \mathrm{PA}_{\mathcal{G}_\ell}(j)$, then $\sigma_0(k) < \sigma_0(j)$.*

Remark 1. *We assume that the model is not directly formulated as an arbitrary ANM in the time domain as Equation 1, but rather under a partial separability basis as Equations 4 and 5. This ensures both the identifiability of the nonlinear DAG structure and the model's ability to utilize the properties of partial separability.*

Combining Equations 2 and 4, we conclude:

$$\left\langle \sum_{\ell=1}^{\infty} \theta_{j\ell}\phi_\ell, \phi_m \right\rangle = \left\langle \sum_{\ell=1}^{\infty} \sum_{k \in \mathrm{PA}_{\mathcal{G}_0}(j)} f_{jk\ell}(\theta_{k\ell})\phi_\ell, \phi_m \right\rangle + \left\langle \sum_{\ell=1}^{\infty} \epsilon_{j\ell}\phi_\ell, \phi_m \right\rangle. \tag{6}$$

By Definition 1, a node $X_k(t) \in \mathrm{PA}_{\mathcal{G}_0}(j)$ if and only if the structural equation for $X_j(t)$ exhibits a non-trivial functional dependence on $X_k(t)$. In terms of spectral coordinate decompositions, $X_j(t)$ depends non-trivially on $X_k(t)$ if and only if there exists at least one spectral index $m \geq 1$ such that the spectral variable θ_{jm} depends non-trivially on θ_{km}. Consequently, the overall functional parent set for node j satisfies:

$$\mathrm{PA}_{\mathcal{G}_0}(j) = \bigcup_{m=1}^{\infty} \mathrm{PA}_{\mathcal{G}_m}(j). \tag{7}$$

Extending this equivalence across all vertex pairs in $\mathcal{V} \times \mathcal{V}$, we obtain:

$$\mathcal{E}_0 = \Big\{(k,j) \in \mathcal{V} \times \mathcal{V} \mid k \in \mathrm{PA}_{\mathcal{G}_0}(j)\Big\} = \bigcup_{\ell=1}^{\infty} \Big\{(k,j) \mid k \in \mathrm{PA}_{\mathcal{G}_\ell}(j)\Big\} = \bigcup_{\ell=1}^{\infty} \mathcal{E}_\ell. \tag{8}$$

Theorem 1. *Suppose that $\mathbf{X}(t)$ satisfies Causal Partial Separability, the functional level graph edge set is equivalent to the spectral union graph edge sets, that is, $\mathcal{E}_0 = \bigcup_{\ell=1}^{\infty} \mathcal{E}_\ell$.*

In addition, the infinite spectral blockwise graphs $\{\mathcal{G}_\ell\}_{\ell=1}^{\infty}$ share a consistent topological ordering function $\sigma_0 : \mathcal{V} \to \{1, 2, \ldots, p\}$, any directed edge $(k, j) \in \mathcal{E}_\ell$ within any arbitrary spectral block $\ell \in \mathbb{N}^+$ strictly respects the ordering inequality:

$$(k \to j) \in \mathcal{E}_\ell \implies \sigma_0(k) < \sigma_0(j), \quad \forall \ell \in \mathbb{N}^+. \tag{9}$$

Thus, the global spectral union graph is a valid DAG and $\mathcal{G} = \bigcup_{\ell=1}^{\infty} \mathcal{G}_\ell$.

Based on the restricted nonlinear Additive Noise Model (Hoyer et al., 2008) properties and the equivalence Theorem 1, the functional level DAG can be uniquely identified.

Lemma 1. *For any $\ell \geq 1$, $(k, j) \in \mathcal{E}_\ell$, let $\theta_{-k,\ell} := \{\theta_{k'\ell}\}_{k' \in \mathrm{PA}_{\mathcal{G}_\ell}(j) \setminus \{k\}}$, and regard $f_{j\ell}$, with $\theta_{-k,\ell}$ held fixed, as a univariate function $f_{j\ell}(\cdot\,; \theta_{-k,\ell}) \in C^3(\mathbb{R})$ of $\theta_{k\ell}$. Identifiability satisfies:*

(i) $f'_{j\ell}(\cdot\,; \theta_{-k,\ell}) \neq 0$;

(ii) both $\theta_{k\ell} \mid \theta_{-k,\ell}$ and $\epsilon_{j\ell}$ admit strictly positive, C^3 log-densities;

(iii) the triple $\big(f_{j\ell}(\cdot\,; \theta_{-k,\ell}),\, p_{\theta_{k\ell}|\theta_{-k,\ell}},\, p_{\epsilon_{j\ell}}\big)$ fails to satisfy the degenerate third-order ordinary differential equation.

Then, fix $\ell \geq 1$ and $(k, j) \in \mathcal{E}_\ell$, and condition on $\theta_{-k,\ell} = s$. Since $\epsilon_{j\ell} \perp\!\!\!\perp \{\theta_{k'\ell}\}_{k' \in \mathrm{PA}_{\mathcal{G}_\ell}(j)}$, the conditional model

$$\theta_{j\ell} = f_{j\ell}(\theta_{k\ell}; s) + \epsilon_{j\ell}, \qquad \epsilon_{j\ell} \perp\!\!\!\perp \theta_{k\ell} \mid \theta_{-k,\ell} = s, \tag{10}$$

is a bivariate ANM whose ingredients satisfy the hypotheses of Lemma 1. If a backward representation

$$\theta_{k\ell} = g_{k\ell}(\theta_{j\ell}) + \tilde{\epsilon}_{k\ell}, \qquad \tilde{\epsilon}_{k\ell} \perp\!\!\!\perp \theta_{j\ell} \mid \theta_{-k,\ell} = s, \tag{11}$$

existed, $\big(f_{j\ell}(\cdot; s), p_{\theta_{k\ell}|\theta_{-k,\ell}}, p_{\epsilon_{j\ell}}\big)$ has positive Lebesgue measure; this contradicts condition (iii). The details are provided in Appendix A.1.

Hence for $\theta_{-k,\ell}$ almost every s no backward ANM exists, and the direction $k \to j$ is determined by $\boldsymbol{\theta}_{\cdot\ell}$. $\mathcal{E}_\ell$ is uniquely determined by $\boldsymbol{\theta}_{\cdot\ell}$ for every $\ell \geq 1$. By Definition 2 and Theorem 1, $(k, j) \in \mathcal{E}_\ell \Rightarrow \sigma_0(k) < \sigma_0(j)$ for all ℓ, so $\mathcal{E}_0 = \bigcup_{\ell=1}^{\infty} \mathcal{E}_\ell$ admits the topological ordering σ_0 and $\mathcal{G}_0$ is acyclic. Therefore the functional level DAG $\mathcal{G}_0$ is uniquely identified by the law of the functional observations $\mathbf{X}(t)$.

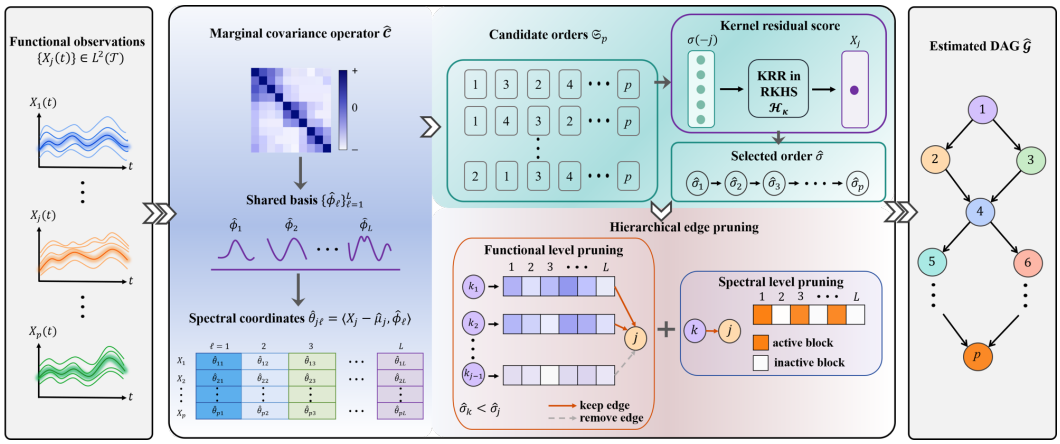


Figure 1: Overview of cps-fDAG.

4 CPS FOR FUNCTIONAL DAG LEARNING

Building upon Causal Partial Separability, we present cps-fDAG, as illustrated in Figure 1, a scalable estimation method for learning functional nonlinear DAGs.

4.1 SHARED SPECTRAL BASIS

Given fully observed i.i.d. samples $\{\mathbf{X}_{(i)}(t)\}_{i=1}^n$, the preliminary mean and empirical marginal covariance operators are computed via finite sample averaging:

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}, \quad \hat{\mathcal{C}}_{jk} = \frac{1}{n}\sum_{i=1}^n (X_{ij} - \hat{\mu}_j) \otimes (X_{ik} - \hat{\mu}_k), \quad \forall j,k = 1,\dots,p, \tag{12}$$

where $\otimes$ denotes the tensor product operator defined by $(f \otimes g)(h) = \langle g, h\rangle f$ for any $f, g, h \in L^2(\mathcal{T})$. This yields the empirical average covariance operator $\hat{\mathcal{H}} = p^{-1}\sum_{j=1}^p \hat{C}_{jj}$.

By solving the spectral decomposition $\hat{\mathcal{H}}\hat{\phi}_\ell = \hat{\lambda}_\ell \hat{\phi}_\ell$, we obtain the empirical shared orthonormal basis $\{\hat{\phi}_\ell\}_{\ell=1}^L$ truncated at level L. Each centered trajectory is then projected onto this basis via the inner product to obtain the empirical spectral coordinates:

$$\hat{\theta}_{ij\ell} = \langle X_{ij} - \hat{\mu}_j, \hat{\phi}_\ell\rangle, \quad \forall i = 1,\dots,n;\ j = 1,\dots,p;\ \ell = 1,\dots,L. \tag{13}$$

Theorem 2. *For any $L \geq 1$ and any orthonormal set $\{\hat{\psi}_\ell\}_{\ell=1}^L \subset L^2(\mathcal{T})$, the total empirical variance satisfies the bound:*

$$\sum_{\ell=1}^L \sum_{j=1}^p \mathrm{Var}\left(\langle X_{ij} - \hat{\mu}_j, \hat{\psi}_\ell\rangle\right) \leq p\sum_{\ell=1}^L \hat{\lambda}_\ell, \tag{14}$$

where equality holds if and only if the subspace spanned by $\{\hat{\psi}_1,\dots,\hat{\psi}_L\}$ is equal to $\{\hat{\phi}_1,\dots,\hat{\phi}_L\}$.

We impose the following condition to quantify the approximation induced by spectral truncation.

Assumption 1. *There exist constants $\omega > 0$, $\eta > 0$, $\varsigma^2 > 0$, and $M < \infty$ independent of n and p, such that the standardized coordinate score is sub-Gaussian with parameter ς^2,*

$$\mathbb{E}\left[\exp\left(u\theta_{ij\ell}\right)\right] \leq \exp\left(\frac{u^2\varsigma^2\,\mathrm{Var}(\theta_{ij\ell})}{2}\right), \quad \forall\, u \in \mathbb{R},\ \ell \in \mathbb{N}^+,\ j \in \mathcal{V}. \tag{15}$$

Furthermore, $\sup_{j\in\mathcal{V}}\sum_{\ell=1}^\infty \mathrm{Var}(\theta_{ij\ell}) \leq M$. The truncation tail error and the adjacent eigengaps satisfy

$$\max_{j\in\mathcal{V}} \sum_{\ell=L+1}^\infty \sigma_{j\ell}^2 \leq L^{-\omega}, \qquad \min_{1\leq\ell\leq L}(\lambda_\ell - \lambda_{\ell+1}) \geq L^{-\eta}. \tag{16}$$

Under Assumption 1, the sub-Gaussian condition on the coordinate scores ensures exponential concentration for the empirical covariance operators and guarantees that the sample covariance operator satisfies bounded deviations (Zapata et al., 2022). The finite-dimensional spectral representation is asymptotically lossless and the eigenvalues $\{\lambda_\ell\}_{\ell=1}^{L}$ exhibit separated spectral gaps (Lee & Li, 2022).

4.2 COMMON CAUSAL ORDERING

Let $\sigma \in \mathfrak{S}_p$ denote a candidate common causal ordering, and define $\sigma(-j) = \{k \in \mathcal{V} : \sigma_k < \sigma_j\}$ as the set of predecessors of node j under σ. Let κ be a positive-definite kernel that is three times Fréchet differentiable, such as the radial basis function (RBF) kernel, and let $\mathcal{H}_\kappa$ denote its associated reproducing kernel Hilbert space (RKHS). For each node j and spectral block ℓ, $\mathbf{h}_{j\ell} = \{h_{jk\ell}\}_{k\in\sigma}(-j) \subset \mathcal{H}_\kappa$, we fit a blockwise Kernel Ridge Regression (KRR):

$$\hat{\mathbf{h}}_{j\ell} = \underset{\{h_{jk\ell}\}\subset\mathcal{H}_\kappa}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left[\hat{\theta}_{ij\ell} - \sum_{k\in\sigma(-j)} h_{jk\ell}\left(\hat{\theta}_{ik\ell}\right) \right]^2 + \beta \sum_{k\in\sigma(-j)} \|h_{jk\ell}\|^2_{\mathcal{H}_\kappa} \right\}. \tag{17}$$

We compute the residual for node j in block ℓ:

$$\hat{\Gamma}_{ij\ell} = \hat{\theta}_{ij\ell} - \sum_{k\in\sigma(-j)} \hat{h}_{jk\ell}\left(\hat{\theta}_{ik\ell}\right). \tag{18}$$

To evaluate the overall causal harmony across nodes, we measure the Hilbert-Schmidt Independence Criterion (HSIC) in spectral block ℓ:

$$\hat{\mathcal{D}}_\ell(\sigma) = \sum_{j=1}^{p} \mathrm{HSIC}\left(\{\hat{\Gamma}_{ij\ell}\}_{i=1}^{n}, \{\left(\hat{\theta}_{ik\ell}\right)_{k\in\sigma(-j)}\}_{i=1}^{n}\right). \tag{19}$$

The additive RKHS regression presents a specific ANM in Lemma 1, and HSIC gives the independence conditions for identifiability. Summing across all spectral truncation levels L, the optimal common topological ordering $\hat{\sigma}$ is identified by solving:

$$\hat{\mathcal{S}}_L(\sigma) = \sum_{\ell=1}^{L} \hat{\mathcal{D}}_\ell(\sigma), \qquad \hat{\sigma} \in \underset{\sigma\in\mathfrak{S}_p}{\arg\min}\, \hat{\mathcal{S}}_L(\sigma). \tag{20}$$

Assumption 2. *Let $\mathcal{S}^0 = \{\sigma \in \mathfrak{S}_p : \sigma_k < \sigma_j, \forall (k,j) \in \mathcal{E}_0\}$ denotes the set of topological orderings of the true functional DAG. There exists a finite $L^\star$, such that for any $L \geq L^\star, \sigma \in \mathcal{S}^0$, we have $\mathcal{S}_L(\sigma) = 0$ and $\mathcal{E}_0 = \bigcup_{\ell=1}^{L^\star} \mathcal{E}_\ell$.*

Theorem 3. *Suppose that $\mathbf{X}(t)$ satisfies Causal Partial Separability and the blockwise ANMs satisfy the identifiability conditions of Lemma 1 and $\beta \to 0$.Then, $\arg\min_{\sigma\in\mathfrak{S}_p} \mathcal{S}_L(\sigma) = \mathcal{S}^0$. For any $\sigma^0 \in \mathcal{S}^0$, define*

$$\Delta_L = \frac{1}{p} \min_{\sigma\notin\mathcal{S}^0} \left\{\mathcal{S}_L(\sigma) - \mathcal{S}_L(\sigma^0)\right\}. \tag{21}$$

If $\mathcal{S}^0 \neq \mathfrak{S}_p$, then $\Delta_L \geq \Delta_{L^\star} > 0$.

Theorem 3 shows that the kernel residual score has a strictly positive population separation between valid and invalid causal orders. The proof is deferred to Appendix A.2.

Theorem 4. *Suppose that Assumptions 1 and 2 hold, the estimated spectral coordinates, the fitted blockwise KRR regressions, and the empirical HSIC statistics satisfy*

$$\frac{1}{p} \sup_{\sigma\in\mathfrak{S}_p} \left|\hat{\mathcal{S}}_L(\sigma) - \mathcal{S}_L(\sigma)\right| = O_p(r_{n,p,L}), \tag{22}$$

where

$$r_{n,p,L} = L\left(\frac{L^\eta}{\beta}\sqrt{\frac{\log(pL)}{n}} + \sqrt{\frac{\log(pL)}{n\beta}} + \sqrt{\beta} + \sqrt{\frac{\log(pL)}{n}}\right). \tag{23}$$

If $r_{n,p,L} = o(\Delta_L)$, then $\mathbb{P}\left\{\hat{\sigma} \in \mathcal{S}^0\right\} \to 1$ as $n \to \infty$.

Theorem 4 states that order recovery is consistent whenever the uniform estimation error of the empirical score is asymptotically dominated by the population separation. A detailed derivation is provided in Appendix A.3.

4.3 HIERARCHICAL EDGE PRUNING

Once the common topological ordering $\hat{\sigma}$ is identified, we constrain the candidate parents of each node $j \in \mathcal{V}$ strictly to its predecessors $\hat{\sigma}(-j)$. For each target node $j \in \mathcal{V}$, we jointly optimize the functional mappings $\mathbf{f}_j = \{f_{jk\ell}\}_{k\in\hat{\sigma}(-j),\ell\le L} \subset \mathcal{H}_\kappa$ by minimizing:

$$\hat{\mathbf{f}}_j = \underset{\{f_{jk\ell}\}\subset\mathcal{H}_\kappa}{\arg\min} \left\{ \frac{1}{2n}\sum_{\ell=1}^{L}\sum_{i=1}^{n}\left[\hat{\theta}_{ij\ell} - \sum_{k\in\hat{\sigma}(-j)} f_{jk\ell}\left(\hat{\theta}_{ik\ell}\right)\right]^2 + \gamma \sum_{k\in\hat{\sigma}(-j)} \mathcal{P}(\mathbf{f}_{jk}) \right\}, \tag{24}$$

where the hierarchical penalty $\mathcal{P}(\mathbf{f}_{jk})$ enforces a nested sparse selection:

$$\mathcal{P}(\mathbf{f}_{jk}) = \alpha\sqrt{\sum_{\ell=1}^{L}\|f_{jk\ell}\|^2_{\mathcal{H}_\kappa}} + (1-\alpha)\sum_{\ell=1}^{L}\|f_{jk\ell}\|_{\mathcal{H}_\kappa}, \tag{25}$$

where $\alpha \in (0,1]$. The first penalty controls the existence of the edge $k \to j$ at the global functional level across all frequencies. The second penalty prunes inactive spectral blocks, isolating active frequency bands for each connection. Finally, the estimated functional level edge set is

$$\hat{\mathcal{E}} = \bigcup_{\ell=1}^{L}\hat{\mathcal{E}}_\ell = \left\{ (k,j) : k \in \hat{\sigma}(-j),\ \sqrt{\sum_{\ell=1}^{L}\|f_{jk\ell}\|^2_{\mathcal{H}_\kappa}} > 0 \right\}. \tag{26}$$

Assumption 3. *There exists a constant $\xi \in (0,1]$ such that, uniformly over $\sigma \in S^0$, $j \in V$, $L \ge L^\star$, and $k \in \sigma(-j) \setminus \mathrm{PA}_{\mathcal{G}_0}(j)$,*

$$\sup_{\mathcal{P}(\mathbf{g}_{jk})\le 1} \frac{1}{\gamma}\sum_{\ell=1}^{L}\left\langle \nabla_{f_{jk\ell}} \frac{1}{2n}\sum_{i=1}^{n}\left[\theta_{ij\ell} - \sum_{k\in\sigma(-j)} f_{jk\ell}(\theta_{ik\ell})\right]^2, g_{jk\ell}\right\rangle_{\mathcal{H}_\kappa} \le 1-\xi, \tag{27}$$

where the gradient is evaluated at the penalized oracle solution restricted to $k \notin \mathrm{PA}_{\mathcal{G}_0}(j)$.

Assumption 4. *Uniformly over $\sigma \in S^0$, $j \in V$, and $L \ge L^\star$, the Hessian of the squared loss in Equation 27 restricted to $k \in \mathrm{PA}_{\mathcal{G}_0}(j)$ has minimum eigenvalue bounded below by a constant $\mu > 0$, and $\rho_L = \min_{j\in\mathcal{V},\, k\in\mathrm{PA}_{\mathcal{G}_0}(j)} \mathcal{P}(\mathbf{f}_{jk}) > 0$.*

Assumption 3 controls non-parent effects, while Assumption 4 ensures parent-set identifiability. We next establish the consistency of edge pruning in Theorem 5, and give the proof in Appendix A.4.

Theorem 5. *Suppose that $X(t)$ satisfies Causal Partial Separability, Assumptions 1–4 hold, and the conditions of Theorem 4 are satisfied. Let $\alpha \in (0,1]$ be fixed and $L \ge L^\star$. If $L^{\eta+\frac{1}{2}}\sqrt{\frac{\log(pL)}{n}} = o(\gamma\xi)$ and $L^{\eta+\frac{1}{2}}\sqrt{\frac{\log(pL)}{n}} + \gamma = o(\rho_L)$, then $\mathbb{P}\left\{\hat{\mathcal{E}} = \mathcal{E}_0\right\} \to 1$ as $n \to \infty$.*

Corollary 1. *Under the consistency conditions of Theorems 4 and 5, the optimal regularization parameters are chosen as $\beta = \left(L^\eta\sqrt{\frac{\log(pL)}{n}}\right)^{2/3}$ and $\gamma = \left(\frac{\rho_L}{\xi}L^{\eta+\frac{1}{2}}\sqrt{\frac{\log(pL)}{n}}\right)^{1/2}$. Then, a sufficient sample size scaling for both common causal ordering recovery and hierarchical edge pruning consistency is*

$$n \gg \max\left\{\frac{L^{2\eta+6}}{\Delta_L^6},\ \frac{L^{2\eta+1}}{\xi^2\rho_L^2}\right\}\log(pL). \tag{28}$$

Remark 2. *If Δ_L, ξ, and ρ_L are uniformly bounded away from zero, the sufficient sample size condition equation 28 reduces to*

$$n \gg L^{2\eta+6}\log(pL). \tag{29}$$

Appendix presents the proof of Corollary 1.

In practice, there are $p!$ possible permutations for optimization. When dealing with high-dimensional scenarios, exhaustive search becomes computationally intractable. We give two strategies. First, we identify candidate neighborhoods $\hat{\mathcal{N}}_j$ via RKHS-based group sparse regression across all L spectral components. Confining the conditioning set to $\sigma(-j) \cap \hat{\mathcal{N}}_j$ shrinks the regression dimension from $\mathcal{O}(p)$ to $|\hat{\mathcal{N}}_j| \ll p$. Next, we replace global enumeration with a greedy local swap search accelerated by hash caching, reducing the search complexity from factorial $\mathcal{O}(p!)$ to polynomial time.

5 EXPERIMENTS

In this section, we conduct comprehensive evaluations to validate the performance and theoretical properties of cps-fDAG. We evaluate cps-fDAG against two baseline categories: (1) Functional methods: MultiFun-DAG (Lan et al., 2024) and fSEM (Lee & Li, 2022); (2) Scalar methods via FPCA coordinates: NOTEARS-MLP (Zheng et al., 2020) and SCORE (Rolland et al., 2022). Graph recovery is measured by SHD, TPR, FDR, and F1-score on directed edges. Additional baseline implementations and extended metric definitions are provided in Appendix B.1.

Synthetic Datasets Results. Ground truth functional DAGs $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ are synthesized across Erdős–Rényi (ER-m) and Scale-Free (SF-m) graphs ($m \in \{1, 2, 4\}$, node count $p \in \{10, 20, 50\}$). Functional trajectories $X_j(t)$ are generated using $L_0 = 10$ shared Fourier bases on $\mathcal{T} = [0, 1]$. The blockwise causal interactions incorporate nonlinear operator mechanisms including trigonometric, sigmoidal, and zero-mean RBF kernel transformations. Exogenous errors $E_j(t)$ are modeled via independent spectral noise. Continuous trajectories are discretely sampled over time grid $T = 100$ across sample sizes $n \in \{500, 1000\}$. All results are averaged over 10 independent trials. Detailed mathematical formulations are provided in Appendix B.2.1.

Table 1: Synthetic graph recovery results across different ER and SF graph settings.

	ER-2 ($p = 20, n = 500$)				**SF-2** ($p = 20, n = 500$)			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	17 ± 2	0.717 ± 0.023	0.189 ± 0.017	0.761 ± 0.020	18 ± 1	0.651 ± 0.027	0.207 ± 0.026	0.713 ± 0.023
fSEM	26 ± 4	0.744 ± 0.037	0.407 ± 0.040	0.653 ± 0.035	25 ± 2	0.751 ± 0.023	0.415 ± 0.030	0.654 ± 0.025
NOTEARS-MLP (FPCA)	36 ± 3	0.375 ± 0.040	0.445 ± 0.039	0.435 ± 0.033	39 ± 2	0.297 ± 0.026	0.571 ± 0.026	0.345 ± 0.021
SCORE (FPCA)	33 ± 3	0.355 ± 0.038	0.360 ± 0.057	0.454 ± 0.044	36 ± 1	0.286 ± 0.026	0.483 ± 0.033	0.363 ± 0.026
cps-fDAG (Ours)	$\mathbf{10 \pm 2}$	$\mathbf{0.863 \pm 0.019}$	$\mathbf{0.103 \pm 0.022}$	$\mathbf{0.878 \pm 0.015}$	$\mathbf{13 \pm 1}$	$\mathbf{0.773 \pm 0.030}$	$\mathbf{0.144 \pm 0.016}$	$\mathbf{0.810 \pm 0.022}$
	ER-4 ($p = 20, n = 500$)				**ER-2** ($p = 50, n = 1000$)			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	56 ± 3	0.608 ± 0.018	0.324 ± 0.018	0.639 ± 0.016	49 ± 1	0.646 ± 0.015	0.161 ± 0.009	0.729 ± 0.011
fSEM	58 ± 4	0.444 ± 0.033	0.395 ± 0.030	0.510 ± 0.031	50 ± 4	0.796 ± 0.018	0.319 ± 0.025	0.733 ± 0.019
NOTEARS-MLP (FPCA)	79 ± 3	0.376 ± 0.009	0.490 ± 0.011	0.432 ± 0.007	103 ± 6	0.295 ± 0.031	0.517 ± 0.030	0.360 ± 0.032
SCORE (FPCA)	74 ± 3	0.286 ± 0.022	0.394 ± 0.029	0.386 ± 0.026	92 ± 4	0.295 ± 0.021	0.399 ± 0.026	0.394 ± 0.023
cps-fDAG (Ours)	$\mathbf{42 \pm 4}$	$\mathbf{0.793 \pm 0.020}$	$\mathbf{0.296 \pm 0.023}$	$\mathbf{0.744 \pm 0.020}$	$\mathbf{26 \pm 2}$	$\mathbf{0.801 \pm 0.012}$	$\mathbf{0.068 \pm 0.013}$	$\mathbf{0.861 \pm 0.009}$

As shown in Table 1, our method consistently delivers superior structural recovery capability. It achieves a stronger trade-off between true recovery and false discovery, yielding lower SHD and higher F1 than competing methods. More results of synthetic datasets see Appendix B.2.2.

Table 2: Causal structure recovery performance on the fMRI NetSim benchmark.

	NetSim-1				**NetSim-2**				**NetSim-3**			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	3	0.800	0.333	0.727	9	0.364	0.429	0.444	16	0.500	0.500	0.500
fSEM	1	0.800	0.000	0.889	13	0.636	0.650	0.452	14	0.611	0.522	0.537
NOTEARS-MLP (FPCA)	4	0.600	0.400	0.600	10	0.273	0.500	0.353	27	0.222	0.765	0.229
SCORE (FPCA)	2	1.000	0.286	0.833	12	0.455	0.545	0.455	20	0.278	0.615	0.323
cps-fDAG (Ours)	**0**	**1.000**	**0.000**	**1.000**	**3**	**0.727**	**0.273**	**0.727**	**6**	**0.722**	**0.188**	**0.765**

fMRI Benchmark. Estimating brain networks from functional magnetic resonance imaging (fMRI) data is crucial for understanding neural mechanisms (Logothetis, 2008). Moreover, fMRI data have been shown to exhibit an inherent separable structure (Smith et al., 2011; Zapata et al., 2022). We evaluate our method on the simulated fMRI benchmark **NetSim** (Smith et al., 2011), using the standard configurations Sim1–Sim3. The graph dimensions scale as $p \in \{5, 10, 15\}$ regions of interest (ROI), each containing $N = 50$ subjects with continuous blood-oxygenation-level-dependent (BOLD) trajectories sampled at $T = 200$ time grid. As shown in Table 2, cps-fDAG demonstrates robust and biologically plausible causal discovery in complex neural systems. More analysis is available in Appendix B.3.

Ablation Study. We investigate the performance of cps-fDAG arises from its components: w/o Shared Basis, which replaces the shared eigensystem with node-wise marginal FPCA bases; w/o HSIC Score, which replaces $\mathcal{S}_L$ with a Gaussian likelihood score; and w/o Hierarchical Pruning, which removes the cross-spectral hierarchical coupling by setting $\alpha = 0$ in Equation 25. To evaluate causal order quality, we compute normalized Kendall tau distance τ. Details on the ablation variants and τ definition are provided in Appendix B.4.1.

Table 3: Mechanistic ablation study on synthetic ER-2 graph ($p = 20$, $n = 500$).

Method	SHD ↓	TPR ↑	FDR ↓	F1 ↑	Kendall τ ↓
w/o Shared Basis	33 ± 3	0.473 ± 0.040	0.400 ± 0.038	0.526 ± 0.037	0.275 ± 0.020
w/o HSIC Score	11 ± 1	0.826 ± 0.016	0.135 ± 0.021	0.843 ± 0.014	0.226 ± 0.024
w/o Hierarchical Pruning	13 ± 2	0.847 ± 0.020	0.165 ± 0.026	0.837 ± 0.015	$\mathbf{0.178 \pm 0.018}$
cps-fDAG (Full)	$\mathbf{10 \pm 2}$	$\mathbf{0.863 \pm 0.019}$	$\mathbf{0.103 \pm 0.022}$	$\mathbf{0.878 \pm 0.015}$	$\mathbf{0.178 \pm 0.018}$

As shown in Table 3, ablating the shared basis causes a performance collapse (F_1 drops to 0.526, τ rises to 0.275), confirming that unified spectral coordinate alignment is indispensable for capturing functional dependencies. Replacing $\mathcal{S}_L$ with a Gaussian likelihood score degrades topological ordering (τ worsens to 0.226), validating its robustness in orienting functional edges. Disabling hierarchical coupling elevates FDR from 0.103 to 0.165, substantiating its isolated mechanism in suppressing spurious edges downstream. Results for more datasets are attached in Appendix B.4.2.

Robustness Analysis. We introduce a cross-basis contamination parameter $\chi \in [0, 1)$ in the synthetic data generation to simulate cases where the data do not fully satisfy causal partial separability. Here, $\chi = 0$ corresponds to the standard CPS setting, while increasing to 1 systematically amplifies the strength of cross-basis causal interventions. Appendix B.5.1 provides more details.

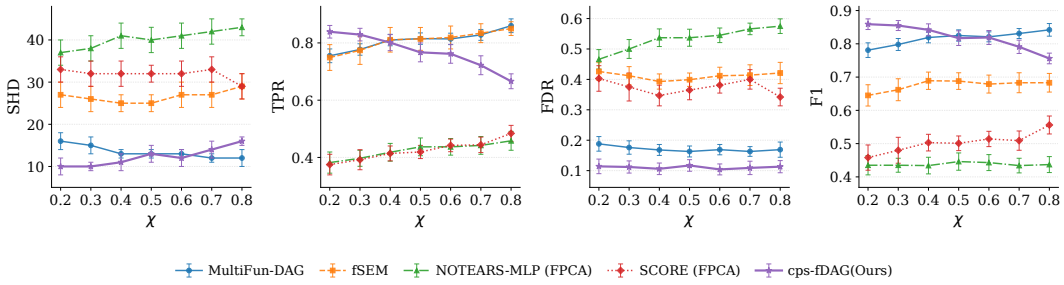


Figure 2: Robustness of all metrics with respect to χ on synthetic ER-2 graph ($p = 20$, $n = 500$).

As shown in Figure 2, cps-fDAG retains a clear performance advantage for contamination levels up to $\chi = 0.5$, while its performance gradually degrades under stronger cross-basis contamination. Additional results on different graphs are provided in Appendix B.5.2.

6 Conclusion

In this work, we propose Causal Partial Separability (CPS), a causal framework for nonlinear DAG learning from multivariate functional data. CPS decouples the functions with a shared spectral basis and a common causal ordering. We analyze that the functional DAG is equivalent to the union of blockwise DAGs, and derive the identifiability of the functional DAG under the nonlinear ANMs assumption. Building on CPS, we introduce cps-fDAG, a functional DAG estimation method that combines shared spectral basis learning, common causal ordering search, and hierarchical edge pruning. We establish the theoretical consistency of both common ordering and edge pruning, and provide a sufficient sample size condition. Experiments on synthetic datasets and biophysical fMRI simulation benchmark demonstrate that cps-fDAG outperforms baselines in recovery accuracy and provides an effective causal DAG. Future work will focus on estimating causal effects for multivariate functional data under generalized nonlinear models that do not impose the additivity of functions.

7 LLM USAGE

In this work, we used generative AI tools solely to assist with manuscript writing and editing to improve clarity and readability. Generative AI was not used to develop the research concepts or ideas or to conduct the analyses. We carefully reviewed all AI-assisted content and take full responsibility for the final manuscript, including any text produced with the assistance of generative AI.

REFERENCES

Hadeel K Aljobouri. Independent component analysis with functional neuroscience data analysis. *Journal of Biomedical Physics & Engineering*, 13(2):169, 2023.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

José R Berrendero, Ana Justel, and Marcela Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634, 2011a.

José R Berrendero, Ana Justel, and Marta Svarc. Modelling daily minimum temperatures by functional data analysis. *Computational Statistics & Data Analysis*, 55(11):2964–2980, 2011b.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-AOS1260.

Tianming Cai. Financial risk management based on functional data analysis. *Journal of Discrete Mathematical Sciences and Cryptography*, 21(6):1397–1400, 2018.

Jeng-Min Chiou, Yu-Ting Chen, and Ya-Fang Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pp. 1571–1596, 2014.

Joao Frois Caldeira, Rangan Gupta, Muhammad Tahir Suleman, and Hudson S Torrent. Forecasting the term structure of interest rates of the brics: Evidence from a nonparametric functional data analysis. *Emerging Markets Finance and Trade*, 57(15):4312–4329, 2021.

Konstantin Göbler, Tobias Windisch, and Mathias Drton. Nonlinear causal discovery for grouped data. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, pp. 1453–1475, 2025.

Ana María Estrada Gómez, Kamran Paynabar, and Massimo Pacella. Functional directed graphical models and applications in root-cause analysis and diagnosis. *Journal of Quality Technology*, 53 (4):421–437, 2020.

Clara Happ and Sonja Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113 (522):649–659, 2018. doi: 10.1080/01621459.2016.1273115.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Zhongnan Jin, Jie Min, Yili Hong, Pang Du, and Qingyu Yang. Multivariate functional clustering with variable selection and application to sensor data from engineering systems. *INFORMS Journal on Data Science*, 3(2):203–218, 2024.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

Tian Lan, Ziyue Li, Junpeng Lin, Zhishuai Li, Lei Bai, Man Li, Fugee Tsung, Rui Zhao, and Chen Zhang. Multifun-dag: Multivariate functional directed acyclic graph. *arXiv preprint arXiv:2404.13836*, 2024.

Kuang-Yao Lee and Lexin Li. Functional structural equation model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):600–629, 2022.

Kuang-Yao Lee, Lexin Li, and Bing Li. Functional directed acyclic graphs. *Journal of Machine Learning Research*, 25(78):1–48, 2024.

Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751. PMLR, 2023.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 186–195. PMLR, 2020.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The american journal of human genetics*, 94(4):559–573, 2014.

Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

James Ramsay and BW Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753. PMLR, 2022.

Saptarshi Roy, Raymond KW Wong, and Yang Ni. Directed cyclic graph for causal discovery from multivariate functional data. *Advances in neural information processing systems*, 36:42762–42774, 2023.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1):20230041, 2024.

Tianle Yang and Joe Suzuki. The functional lingam. In *International Conference on Probabilistic Graphical Models*, pp. 25–36. PMLR, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163. PMLR, 2019.

Javier Zapata, Sang-Yun Oh, and Alexander Petersen. Partial separability and functional graphical models for multivariate gaussian processes. *Biometrika*, 109(3):665–681, 2022.

Manni Zhang and Ana María Estrada Gómez. Directed graphical models for root cause analysis in multimodal systems collecting heterogeneous data. *IISE Transactions*, pp. 1–19, 2025.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3414–3425. PMLR, 2020.

Fangting Zhou, Kejun He, Kunbo Wang, Yanxun Xu, and Yang Ni. Functional bayesian networks for discovering causality from multivariate functional data. *Biometrics*, 79(4):3279–3293, 2023.

Hongxiao Zhu, Nate Strawn, and David B. Dunson. Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(204):1–27, 2016.

Vadim Zipunnikov et al. Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4):852–873, 2011.

A Proofs

A.1 Details of Identifiability

Condition (i) states $f'_{j\ell}(\cdot;\theta_{-k,\ell}) \neq 0$ for every $k \in \mathrm{PA}_{\mathcal{G}_\ell}(j)$, so $f_{j\ell}$ depends non-trivially on each of its arguments and $\theta_{j\ell} \not\perp\!\!\!\perp \theta_{k\ell} \mid \{\theta_{k'\ell}\}_{k' \in \mathrm{PA}_{\mathcal{G}_\ell}(j)\setminus\{k\}}$; thus $P_{\boldsymbol{\theta}_{\cdot\ell}}$ satisfies causal minimality with respect to $\mathcal{G}_\ell$.

We present the third-order differential constraint from Hoyer et al. (2008) as follows.

Lemma 2. *Let (x, y) have a strictly positive density $p \in C^3(\mathbb{R}^2)$ admitting the forward representation $y = f(x) + \epsilon$ with $\epsilon \perp\!\!\!\perp x$, $f \in C^3$, and write $\xi := \log p_x$, $\nu := \log p_\epsilon$, $q(x, y) := \log p(x, y) = \xi(x) + \nu\big(y - f(x)\big)$. If in addition a backward representation $x = g(y) + \tilde{\epsilon}$ with $\tilde{\epsilon} \perp\!\!\!\perp y$, $g \in C^3$ exists, then on the open set $\mathcal{O} := \{(x, y) : f'(x)\,\nu''(y - f(x)) \neq 0\}$ the triple (f, ξ, ν) obeys*

$$\xi''' = \xi''\Big(\frac{f''}{f'} - \frac{f'\nu'''}{\nu''}\Big) - 2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}, \tag{30}$$

where f, ξ and their derivatives are evaluated at x, and ν and its derivatives at $y - f(x)$.

If the backward representation Equation 11 existed, Lemma 2 would force $\big(f_{j\ell}(\cdot; s), p_{\theta_{k\ell}|\theta_{-k,\ell}}, p_{\epsilon_{j\ell}}\big)$ to satisfy Equation 30 on the set $\mathcal{O}$, which has positive Lebesgue measure because $f'_{j\ell} \neq 0$ and $\nu'' \not\equiv 0$; this contradicts condition (iii).

A.2 Proof of Theorem 3

Proof. Fix $\sigma^0 \in \mathcal{S}^0$. For every spectral block ℓ and node j, all parents of j precede j under σ^0. By the blockwise ANM,

$$\theta_{j\ell} = \sum_{k \in \mathrm{PA}_{\mathcal{G}_\ell}(j)} f_{jk\ell}(\theta_{k\ell}) + \epsilon_{j\ell}.$$

Under correct specification of the additive RKHS regression,

$$\Gamma^{\sigma^0}_{j\ell} = \epsilon_{j\ell} \perp\!\!\!\perp \theta_{\sigma^0(-j)\ell}.$$

Therefore every HSIC term vanishes and

$$S_L(\sigma^0) = 0.$$

Conversely, suppose that $S_L(\sigma) = 0$. Since HSIC is nonnegative and the kernels are characteristic,

$$\Gamma^{\sigma}_{j\ell} \perp\!\!\!\perp \theta_{\sigma(-j)\ell}, \qquad j \in V, \quad \ell \leq L.$$

Hence, for every $\ell \le L$, the joint distribution of $\theta_{\cdot\ell}$ admits an additive noise representation compatible with σ. By the blockwise ANM identifiability in Lemma 1, σ must be a topological ordering of $\mathcal{G}_\ell$. By Theorem 1,

$$\mathcal{E}_0 = \bigcup_{\ell \ge 1} \mathcal{E}_\ell.$$

Since $\mathcal{E}_0$ is finite, there exists $L_\star < \infty$ such that

$$\mathcal{E}_0 = \bigcup_{\ell=1}^{L_\star} \mathcal{E}_\ell.$$

Thus, for every $L \ge L_\star$,

$$S_L(\sigma) = 0 \quad \Longleftrightarrow \quad \sigma \in \mathcal{S}^0.$$

This proves

$$\arg\min_{\sigma \in \mathfrak{S}_p} S_L(\sigma) = \mathcal{S}^0.$$

Finally, $S_L(\sigma) > 0$ for every $\sigma \notin \mathcal{S}^0$. Since $\mathfrak{S}_p$ is finite,

$$\Delta_{L_\star} > 0.$$

Moreover, $S_L(\sigma)$ is nondecreasing in L because each additional HSIC term is nonnegative. Hence

$$\Delta_L \ge \Delta_{L_\star} > 0, \qquad L \ge L_\star.$$

□

A.3 Proof of Theorem 4

Proof. Under Assumption 1, operator-norm concentration for the average covariance operator yields

$$\|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{op}} \le \|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

By the Davis–Kahan $\sin\Theta$ theorem and the spectral gap bound $\min_{1\le\ell\le L}(\lambda_\ell - \lambda_{\ell+1}) \ge L^{-\eta}$, with sign alignment $s_\ell = \mathrm{sign}(\langle\hat{\phi}_\ell, \phi_\ell\rangle) \in \{-1, +1\}$,

$$\max_{1\le\ell\le L} \|\hat{\phi}_\ell - s_\ell \phi_\ell\|_{L^2(\mathcal{T})} \le \frac{2\sqrt{2}\|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{op}}}{\min_{1\le\ell\le L}(\lambda_\ell - \lambda_{\ell+1})} = O_p\left(\frac{L^\eta}{\sqrt{n}}\right).$$

Combining this with the identity $\hat{\theta}_{ij\ell} - s_\ell\theta_{ij\ell} = \langle X_{ij} - \mu_j, \hat{\phi}_\ell - s_\ell\phi_\ell\rangle - \langle\hat{\mu}_j - \mu_j, \hat{\phi}_\ell\rangle$ and sub-Gaussian concentration over $j \in \mathcal{V}, \ell \le L$,

$$\max_{j\in\mathcal{V},\ell\le L}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\theta}_{ij\ell} - s_\ell\theta_{ij\ell}|^2\right)^{1/2} = O_p\left(L^\eta\sqrt{\frac{\log(pL)}{n}}\right).$$

For any candidate ordering $\sigma \in \mathfrak{S}_p$, the β-regularized KRR error and corresponding residual perturbation satisfy

$$\sup_{\sigma\in\mathfrak{S}_p}\max_{j\in\mathcal{V},\ell\le L}\|\hat{h}_{j\ell} - h_{j\ell}\|_{\mathcal{H}_\kappa} = O_p\left(\frac{L^\eta}{\beta}\sqrt{\frac{\log(pL)}{n}} + \sqrt{\frac{\log(pL)}{n\beta}} + \sqrt{\beta}\right),$$

$$\sup_{\sigma\in\mathfrak{S}_p}\max_{j\in\mathcal{V},\ell\le L}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\Gamma}_{ij\ell}(\sigma) - s_\ell\Gamma_{ij\ell}(\sigma)|^2\right)^{1/2} = O_p\left(\frac{L^\eta}{\beta}\sqrt{\frac{\log(pL)}{n}} + \sqrt{\frac{\log(pL)}{n\beta}} + \sqrt{\beta}\right).$$

Since the HSIC kernel is bounded, Lipschitz continuous, and invariant under the reflection $s_\ell \in \{-1, +1\}$, empirical process deviation implies

$$\sup_{\sigma\in\mathfrak{S}_p}\max_{j\in\mathcal{V},\ell\le L}\left|\widehat{\mathrm{HSIC}}_{j\ell}(\sigma) - \mathrm{HSIC}_{j\ell}(\sigma)\right| = O_p\left(\frac{L^\eta}{\beta}\sqrt{\frac{\log(pL)}{n}} + \sqrt{\frac{\log(pL)}{n\beta}} + \sqrt{\beta} + \sqrt{\frac{\log(pL)}{n}}\right).$$

Summing over $j \in \mathcal{V}$ and $\ell \le L$ gives

$$\frac{1}{p} \sup_{\sigma \in \mathfrak{S}_p} \left| \hat{\mathcal{S}}_L(\sigma) - \mathcal{S}_L(\sigma) \right| \le \sum_{\ell=1}^{L} \sup_{\sigma \in \mathfrak{S}_p} \max_{j \in \mathcal{V}} \left| \widehat{\mathrm{HSIC}}_{j\ell}(\sigma) - \mathrm{HSIC}_{j\ell}(\sigma) \right| = O_p(r_{n,p,L}).$$

Let $R_{n,p,L} = \frac{1}{p} \sup_{\sigma \in \mathfrak{S}_p} |\hat{\mathcal{S}}_L(\sigma) - \mathcal{S}_L(\sigma)|$. Under $r_{n,p,L} = o(\Delta_L)$,

$$R_{n,p,L} = o_p(\Delta_L) \implies \mathbb{P}\left(R_{n,p,L} < \frac{\Delta_L}{2} \right) \to 1.$$

By Theorem 3, for any $\sigma^0 \in \mathcal{S}^0$ and $\sigma \notin \mathcal{S}^0$,

$$\mathcal{S}_L(\sigma) - \mathcal{S}_L(\sigma^0) \ge p\Delta_L.$$

On the event $\{R_{n,p,L} < \Delta_L/2\}$,

$$\hat{\mathcal{S}}_L(\sigma) - \hat{\mathcal{S}}_L(\sigma^0) \ge \mathcal{S}_L(\sigma) - \mathcal{S}_L(\sigma^0) - 2pR_{n,p,L} \ge p\Delta_L - 2p\left(\frac{\Delta_L}{2}\right) > 0.$$

Consequently,

$$\mathbb{P}\left\{ \hat{\sigma} \in \mathcal{S}^0 \right\} \ge \mathbb{P}\left(R_{n,p,L} < \frac{\Delta_L}{2} \right) \to 1 \quad \text{as } n \to \infty.$$

□

A.4 Proof of Theorem 5

Proof. Let $\Phi_{n,p,L} = L^{\eta+\frac{1}{2}} \sqrt{\frac{\log(pL)}{n}}$. All bounds below hold uniformly over $\sigma \in \mathcal{S}^0$, $j \in \mathcal{V}$, and $L \ge L^\star$. For any fixed $\sigma \in \mathcal{S}^0$ and $j \in \mathcal{V}$, let $A_j = \mathrm{PA}_{\mathcal{G}_0}(j) \subseteq \sigma(-j)$. Define the empirical squared loss and its population counterpart:

$$\hat{\mathcal{L}}^{\sigma}_{j,L}(f_j) = \frac{1}{2n} \sum_{\ell=1}^{L} \sum_{i=1}^{n} \left[\hat{\theta}_{ij\ell} - \sum_{k \in \sigma(-j)} f_{jk\ell}(\hat{\theta}_{ik\ell}) \right]^2,$$

$$\mathcal{L}^{\sigma}_{j,L}(f_j) = \frac{1}{2} \sum_{\ell=1}^{L} \mathbb{E}\left[\theta_{ij\ell} - \sum_{k \in \sigma(-j)} f_{jk\ell}(\theta_{ik\ell}) \right]^2.$$

By the coordinate error bound $\max_{j \in \mathcal{V}, \ell \le L} (\frac{1}{n} \sum_{i=1}^{n} |\hat{\theta}_{ij\ell} - s_\ell \theta_{ij\ell}|^2)^{1/2} = O_p(L^\eta \sqrt{\frac{\log(pL)}{n}})$ and RKHS gradient concentration aggregated across L spectral blocks,

$$\max_{\sigma \in \mathcal{S}^0, j \in \mathcal{V}} \max_{k \in \sigma(-j)} \sup_{\mathcal{P}(g_{jk}) \le 1} \left| \sum_{\ell=1}^{L} \left\langle \nabla_{f_{jk\ell}} \hat{\mathcal{L}}^{\sigma}_{j,L}(f_j) - \nabla_{f_{jk\ell}} \mathcal{L}^{\sigma}_{j,L}(f_j),\, g_{jk\ell} \right\rangle_{\mathcal{H}_\kappa} \right| = O_p(\Phi_{n,p,L}).$$

Define the restricted oracle and empirical estimators on the true parent set A_j:

$$\bar{f}^{\sigma}_j = \operatorname*{arg\,min}_{\{f_{jk\ell}\}_{k \in A_j, \ell \le L}} \left\{ \mathcal{L}^{\sigma}_{j,L}(f_j) + \gamma \sum_{k \in A_j} \mathcal{P}(f_{jk}) \right\},$$

$$\widetilde{f}^{\sigma}_j = \operatorname*{arg\,min}_{\{f_{jk\ell}\}_{k \in A_j, \ell \le L}} \left\{ \hat{\mathcal{L}}^{\sigma}_{j,L}(f_j) + \gamma \sum_{k \in A_j} \mathcal{P}(f_{jk}) \right\}.$$

By the restricted strong convexity in Assumption 4, the first-order optimality condition implies

$$\max_{\sigma \in \mathcal{S}^0} \max_{j \in \mathcal{V}} \sum_{k \in A_j} \mathcal{P}\left(\widetilde{f}^{\sigma}_{jk} - \bar{f}^{\sigma}_{jk} \right) = O_p(\Phi_{n,p,L}),$$

$$\max_{\sigma \in \mathcal{S}^0} \max_{j \in \mathcal{V}} \sum_{k \in A_j} \mathcal{P}\left(\bar{f}^{\sigma}_{jk} - f_{jk} \right) = O(\gamma),$$

$$\max_{\sigma\in\mathcal{S}^0}\max_{j\in\mathcal{V}}\sum_{k\in A_j}\mathcal{P}\left(\widetilde{f}_{jk}^{\sigma}-f_{jk}\right)=O_p(\Phi_{n,p,L}+\gamma).$$

For each true parent $k\in A_j$, the triangle inequality gives

$$\mathcal{P}(\widetilde{f}_{jk}^{\sigma})\geq\mathcal{P}(f_{jk})-\mathcal{P}(\widetilde{f}_{jk}^{\sigma}-f_{jk})\geq\rho_L-O_p(\Phi_{n,p,L}+\gamma).$$

Under $\Phi_{n,p,L}+\gamma=o(\rho_L)$, no true edge is pruned:

$$\mathbb{P}\left\{\min_{\sigma\in\mathcal{S}^0}\min_{j\in\mathcal{V}}\min_{k\in A_j}\mathcal{P}(\widetilde{f}_{jk}^{\sigma})>0\right\}\to 1.$$

For any non-parent $k\in\sigma(-j)\setminus A_j$, strict dual feasibility from Assumption 4 gives

$$\sup_{\mathcal{P}(g_{jk})\leq 1}\frac{1}{\gamma}\sum_{\ell=1}^{L}\left\langle\nabla_{f_{jk\ell}}\mathcal{L}_{j,L}^{\sigma}(\bar{f}_j^{\sigma}),\,g_{jk\ell}\right\rangle_{\mathcal{H}_\kappa}\leq 1-\xi.$$

Perturbing to the empirical restricted minimizer $\widetilde{f}_j^{\sigma}$ yields

$$\sup_{\mathcal{P}(g_{jk})\leq 1}\frac{1}{\gamma}\left|\sum_{\ell=1}^{L}\left\langle\nabla_{f_{jk\ell}}\hat{\mathcal{L}}_{j,L}^{\sigma}(\widetilde{f}_j^{\sigma})-\nabla_{f_{jk\ell}}\mathcal{L}_{j,L}^{\sigma}(\bar{f}_j^{\sigma}),\,g_{jk\ell}\right\rangle_{\mathcal{H}_\kappa}\right|=O_p\left(\frac{\Phi_{n,p,L}}{\gamma}\right).$$

Under $\Phi_{n,p,L}=o(\gamma\xi)$, the triangle inequality ensures

$$\sup_{\mathcal{P}(g_{jk})\leq 1}\frac{1}{\gamma}\sum_{\ell=1}^{L}\left\langle\nabla_{f_{jk\ell}}\hat{\mathcal{L}}_{j,L}^{\sigma}(\widetilde{f}_j^{\sigma}),\,g_{jk\ell}\right\rangle_{\mathcal{H}_\kappa}\leq 1-\xi+O_p\left(\frac{\Phi_{n,p,L}}{\gamma}\right)<1,$$

which holds with probability tending to 1 uniformly over all $k\in\sigma(-j)\setminus A_j$:

$$\mathbb{P}\left\{\max_{\sigma\in\mathcal{S}^0}\max_{j\in\mathcal{V}}\max_{k\in\sigma(-j)\setminus A_j}\sup_{\mathcal{P}(g_{jk})\leq 1}\frac{1}{\gamma}\sum_{\ell=1}^{L}\left\langle\nabla_{f_{jk\ell}}\hat{\mathcal{L}}_{j,L}^{\sigma}(\widetilde{f}_j^{\sigma}),\,g_{jk\ell}\right\rangle_{\mathcal{H}_\kappa}<1\right\}\to 1.$$

By Karush–Kuhn–Tucker (KKT) optimality, the zero-padded estimator $(\widetilde{f}_j^{\sigma},\mathbf{0})$ is the unique global solution. Combined with Theorem 4,

$$\mathbb{P}\left\{\hat{\mathcal{E}}=\mathcal{E}_0\right\}\to 1\quad\text{as } n\to\infty.$$

□

A.5 Proof of Corollary 1

Proof. First, consider the uniform common ordering error in Theorem 4:

$$r_{n,p,L}=L\left(\frac{L^{\eta}}{\beta}\sqrt{\frac{\log(pL)}{n}}+\sqrt{\frac{\log(pL)}{n\beta}}+\sqrt{\beta}+\sqrt{\frac{\log(pL)}{n}}\right).$$

Balancing the two leading β-dependent terms,

$$\frac{L^{1+\eta}}{\beta}\sqrt{\frac{\log(pL)}{n}}=L\sqrt{\beta},$$

yields the optimal regularization choice

$$\beta=\left(L^{\eta}\sqrt{\frac{\log(pL)}{n}}\right)^{2/3}.$$

Under this choice of β,

$$r_{n,p,L}=O_p\left(L^{1+\frac{\eta}{3}}\left[\frac{\log(pL)}{n}\right]^{1/6}\right).$$

Consequently, the ordering consistency condition $r_{n,p,L} = o(\Delta_L)$ requires

$$n \gg \frac{L^{2\eta+6}}{\Delta_L^6} \log(pL).$$

Next, consider the hierarchical edge pruning conditions in Theorem 5:

$$L^{\eta+\frac{1}{2}} \sqrt{\frac{\log(pL)}{n}} = o(\gamma\xi), \qquad L^{\eta+\frac{1}{2}} \sqrt{\frac{\log(pL)}{n}} + \gamma = o(\rho_L).$$

Balancing the lower and upper constraints on γ,

$$\frac{L^{\eta+\frac{1}{2}}}{\xi} \sqrt{\frac{\log(pL)}{n}} \ll \gamma \ll \rho_L,$$

yields the optimal choice

$$\gamma = \left(\frac{\rho_L}{\xi} L^{\eta+\frac{1}{2}} \sqrt{\frac{\log(pL)}{n}} \right)^{1/2}.$$

Under this choice of γ,

$$\frac{L^{\eta+\frac{1}{2}} \sqrt{\log(pL)/n}}{\gamma\xi} = \frac{\gamma}{\rho_L} = \left(\frac{L^{\eta+\frac{1}{2}} \sqrt{\log(pL)/n}}{\xi\rho_L} \right)^{1/2}.$$

Hence, both conditions hold simultaneously if and only if

$$L^{\eta+\frac{1}{2}} \sqrt{\frac{\log(pL)}{n}} = o(\xi\rho_L),$$

which is equivalent to

$$n \gg \frac{L^{2\eta+1}}{\xi^2 \rho_L^2} \log(pL).$$

Combining the two sufficient sample size conditions yields

$$n \gg \max\left\{ \frac{L^{2\eta+6}}{\Delta_L^6}, \frac{L^{2\eta+1}}{\xi^2\rho_L^2} \right\} \log(pL).$$

□

B Supplementary Experimental Details and Results

B.1 Implementation Details

Baseline Implementation and Experimental Setup. To ensure fair comparison and reproducibility, all methods process identical densely observed functional trajectories generated under a unified protocol. Public codebases are linked via footnotes.

- **cps-fDAG (Ours)** uses $L = 3$ components and kernel ridge regularization $\beta = 0.01$. Pruned via pruning penalty $\gamma = 0.08$ and hierarchical ordering $\alpha = 0.65$.
- **MultiFun-DAG** (Lan et al., 2024)[1] retains 3 Functional Principal Component Analysis (FPCA) components per node. Optimized using Adam with learning rate 0.03 , 300 inner iterations, 8 outer updates. Final edges are thresholded at 0.3.
- **fSEM** (Lee & Li, 2022)[2] uses 30 random Fourier features. First layer basis dimension is set to 10, and maximum allowed parents per node is 10. Ridge regularization parameter is 0.1, and structural edge penalty is 0.3.

[1] https://github.com/Lantian72/MultiFun-DAG

[2] https://sites.google.com/site/kuangyaolee1217/software

- **NOTEARS-MLP (FPCA)** (Zheng et al., 2020)[3] reduces functional curves to scores using FPCA. The underlying MLP architecture employs a $(10, 1)$ hidden layer layout with bias enabled. Optimization uses continuous acyclicity constraints with 0.01, stopping tolerance 10^{-8}, and edge threshold 0.1.
- **SCORE (FPCA)** (Rolland et al., 2022)[4] reduces functional curves to scores using FPCA. Score estimation steps use step-sizes 0.001 with variance dispersion. Graph pruning is performed via Stein pruning at threshold 0.25 and CAM cutoff 0.001.

Metric Definitions.

- **Structural Hamming Distance (SHD, lower is better):** Total count of edge additions $\mathcal{E}_A$, deletions $\mathcal{E}_D$, and reversals $\mathcal{E}_R$:

$$\mathrm{SHD} = \mathcal{E}_A + \mathcal{E}_D + \mathcal{E}_R.$$

- **True Positive Rate (TPR, higher is better):**

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

- **False Discovery Rate (FDR, lower is better):**

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}.$$

- **F1-Score (higher is better):**

$$\mathrm{Precision} = 1 - \mathrm{FDR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

B.2 Experiments on Synthetic Data

B.2.1 Synthetic Functional Data Generation

Functional observations $X_j(t) = \mu_j(t) + \sum_{\ell=1}^{L} \theta_{j\ell}\phi_\ell(t)$ on $\mathcal{T} = [0, 1]$ are generated using $L = 10$ orthonormal Fourier bases:

$$\phi_1(t) = 1, \quad \phi_{2k}(t) = \sqrt{2}\cos(2\pi kt), \quad \phi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt), \quad k \geq 1.$$

For a given DAG $\mathcal{G}_0$, let $\mathrm{PA}(j)$ denote the set of parents of node j. The spectral coordinates $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jL})^\top$ follow a nonlinear additive noise structural equation model (ANM):

$$\theta_{j\ell} = \sum_{k \in \mathrm{PA}(j)} f_{kj}^{(\ell)}(\theta_{k\ell}) + \varepsilon_{j\ell},$$

where $\varepsilon_{j\ell} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ is the mutually independent Gaussian random error. For each edge $k \to j$ in the specific block ℓ, the univariate nonlinear transformation $f_{kj}^{(\ell)}$ is uniformly sampled from three different nonlinear operators:

Trigonometric Operator: $f_{kj}^{(\ell)}(x) = \sin(1.2\,x)$;

Sigmoidal Operator: $f_{kj}^{(\ell)}(x) = (1 + \exp(-1.5\,x))^{-1} - 0.5$;

Zero-Mean RBF Operator: $f_{kj}^{(\ell)}(x) = \exp\left(-0.5\,x^2\right) - \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\exp(-0.5X^2)]$.

[3] https://github.com/xunzheng/notears

[4] https://github.com/paulrolland1307/SCORE

B.2.2 Experiment Results on More Synthetic Datasets

In this section, we provide comprehensive experimental evaluations across the remaining graph topologies, graph densities, and sample size configurations mentioned in Section 5.

We evaluate the baseline recovery performance on low-dimensional graphs with varying graph densities ($m \in \{1, 2\}$) at sample size $n = 500$ in Table 4.

To investigate whether our framework remains effective under complex network topologies, we examine dense and large graphs, shown in Table 5.

To empirically validate the asymptotic consistency, we report the performance trajectory as the sample size scales increasing to $n = 1000$ on the benchmark ER-4 ($p = 20$) and SF-4 ($p = 20$) system in Table 6.

Table 4: Synthetic causal graph discovery on low-dimensional graphs.

	ER-1 ($p = 20, n = 500$)				**ER-2** ($p = 10, n = 500$)			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	8 ± 1	0.683 ± 0.039	0.085 ± 0.024	0.777 ± 0.031	7 ± 1	0.739 ± 0.023	0.175 ± 0.028	0.778 ± 0.023
fSEM	14 ± 1	0.534 ± 0.036	0.424 ± 0.025	0.553 ± 0.030	9 ± 1	0.736 ± 0.046	0.316 ± 0.050	0.704 ± 0.046
NOTEARS-MLP (FPCA)	19 ± 2	0.253 ± 0.038	0.334 ± 0.077	0.352 ± 0.046	16 ± 1	0.227 ± 0.046	0.494 ± 0.077	0.303 ± 0.055
SCORE (FPCA)	16 ± 1	0.355 ± 0.024	0.269 ± 0.047	0.471 ± 0.023	16 ± 1	0.265 ± 0.040	0.429 ± 0.077	0.353 ± 0.050
CPSFDAG (Ours)	$\mathbf{7 \pm 1}$	$\mathbf{0.709 \pm 0.034}$	$\mathbf{0.052 \pm 0.020}$	$\mathbf{0.807 \pm 0.024}$	$\mathbf{5 \pm 1}$	$\mathbf{0.815 \pm 0.023}$	$\mathbf{0.094 \pm 0.015}$	$\mathbf{0.856 \pm 0.014}$

Table 5: Synthetic graph recovery on dense (SF-4) and high-dimensional (SF-2, $p = 50$) graphs.

	SF-4 ($p = 20, n = 500$)				**SF-2** ($p = 50, n = 500$)			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	44 ± 2	0.594 ± 0.022	0.295 ± 0.022	0.643 ± 0.018	50 ± 2	0.607 ± 0.016	0.202 ± 0.016	0.689 ± 0.014
fSEM	42 ± 2	0.551 ± 0.036	0.324 ± 0.029	0.602 ± 0.029	85 ± 5	0.738 ± 0.021	0.505 ± 0.019	0.591 ± 0.018
NOTEARS-MLP (FPCA)	72 ± 3	0.313 ± 0.019	0.541 ± 0.028	0.368 ± 0.018	109 ± 4	0.325 ± 0.016	0.598 ± 0.026	0.357 ± 0.017
SCORE (FPCA)	64 ± 1	0.249 ± 0.011	0.415 ± 0.031	0.348 ± 0.014	88 ± 2	0.318 ± 0.014	0.441 ± 0.022	0.404 ± 0.016
cps-fDAG (Ours)	$\mathbf{28 \pm 3}$	$\mathbf{0.780 \pm 0.016}$	$\mathbf{0.214 \pm 0.030}$	$\mathbf{0.781 \pm 0.020}$	$\mathbf{30 \pm 5}$	$\mathbf{0.787 \pm 0.024}$	$\mathbf{0.119 \pm 0.030}$	$\mathbf{0.830 \pm 0.026}$

Table 6: Sample size scaling increasing to $n = 1000$ on ER-4 and SF-4 graphs ($p = 20$).

	ER-4 ($p = 20, n = 1000$)				**SF-4** ($p = 20, n = 1000$)			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	52 ± 3	0.602 ± 0.012	0.278 ± 0.017	0.655 ± 0.009	43 ± 2	0.561 ± 0.022	0.256 ± 0.024	0.638 ± 0.019
fSEM	54 ± 3	0.488 ± 0.034	0.355 ± 0.028	0.554 ± 0.031	43 ± 3	0.497 ± 0.032	0.374 ± 0.041	0.554 ± 0.035
NOTEARS-MLP (FPCA)	79 ± 4	0.316 ± 0.018	0.471 ± 0.029	0.392 ± 0.017	71 ± 2	0.254 ± 0.025	0.539 ± 0.031	0.325 ± 0.026
SCORE (FPCA)	74 ± 3	0.251 ± 0.017	0.365 ± 0.040	0.357 ± 0.021	64 ± 2	0.210 ± 0.023	0.395 ± 0.040	0.307 ± 0.026
cps-fDAG (Ours)	$\mathbf{36 \pm 4}$	$\mathbf{0.834 \pm 0.016}$	$\mathbf{0.266 \pm 0.029}$	$\mathbf{0.779 \pm 0.021}$	$\mathbf{26 \pm 2}$	$\mathbf{0.807 \pm 0.016}$	$\mathbf{0.207 \pm 0.024}$	$\mathbf{0.798 \pm 0.015}$

In the default synthetic protocol, structural equations across spectral blocks are sampled uniformly from the three candidate nonlinear families (Trigonometric, Sigmoidal, and Zero-Mean RBF) with an equal $1 : 1 : 1$ ratio. In this section, we generate four distinct mixture weight configurations $\mathbf{w}^{(k)} = (w_{\text{Trig}}, w_{\text{Sigm}}, w_{\text{RBF}})$ via random sampling: Mixture 1: $\mathbf{w}^{(1)} = (0.50, 0.35, 0.15)$; Mixture 2: $\mathbf{w}^{(2)} = (0.15, 0.65, 0.20)$; Mixture 3: $\mathbf{w}^{(3)} = (0.20, 0.25, 0.55)$; Mixture 4: $\mathbf{w}^{(4)} = (0.10, 0.45, 0.45)$. As reported in Table 7, cps-fDAG consistently maintains stable and high recovery fidelity across all four randomized mixture settings.

Table 7: Synthetic graph recovery under randomized nonlinear mechanism mixtures (Settings 1–4) on ER-2 graphs ($p = 20, n = 500$). Results are averaged over 10 independent trials.

	Mixture 1: $(0.50, 0.35, 0.15)$				**Mixture 2:** $(0.15, 0.65, 0.20)$			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	20 ± 3	0.844 ± 0.026	0.287 ± 0.022	0.772 ± 0.022	21 ± 2	0.528 ± 0.029	0.125 ± 0.024	0.657 ± 0.028
fSEM	26 ± 3	0.664 ± 0.044	0.441 ± 0.034	0.603 ± 0.035	33 ± 4	0.765 ± 0.053	0.491 ± 0.033	0.605 ± 0.036
NOTEARS-MLP (FPCA)	46 ± 4	0.449 ± 0.019	0.580 ± 0.016	0.432 ± 0.014	36 ± 2	0.205 ± 0.032	0.414 ± 0.071	0.289 ± 0.039
SCORE (FPCA)	36 ± 3	0.436 ± 0.038	0.460 ± 0.023	0.479 ± 0.031	33 ± 2	0.286 ± 0.026	0.391 ± 0.054	0.389 ± 0.034
cps-fDAG (Ours)	$\mathbf{13 \pm 2}$	$\mathbf{0.936 \pm 0.015}$	$\mathbf{0.210 \pm 0.021}$	$\mathbf{0.856 \pm 0.017}$	$\mathbf{13 \pm 2}$	$\mathbf{0.735 \pm 0.030}$	$\mathbf{0.086 \pm 0.025}$	$\mathbf{0.811 \pm 0.022}$
	Mixture 3: $(0.20, 0.25, 0.55)$				**Mixture 4:** $(0.10, 0.45, 0.45)$			
Method	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑	**SHD** ↓	**TPR** ↑	**FDR** ↓	**F1** ↑
MultiFun-DAG	20 ± 2	0.546 ± 0.029	0.093 ± 0.017	0.678 ± 0.026	24 ± 2	0.404 ± 0.029	0.103 ± 0.025	0.554 ± 0.030
fSEM	36 ± 3	0.800 ± 0.016	0.488 ± 0.037	0.617 ± 0.025	50 ± 5	0.768 ± 0.025	0.590 ± 0.032	0.527 ± 0.028
NOTEARS-MLP (FPCA)	35 ± 2	0.218 ± 0.032	0.424 ± 0.037	0.308 ± 0.035	36 ± 2	0.150 ± 0.025	0.272 ± 0.065	0.235 ± 0.033
SCORE (FPCA)	31 ± 3	0.289 ± 0.032	0.230 ± 0.046	0.416 ± 0.039	31 ± 2	0.245 ± 0.027	0.197 ± 0.031	0.370 ± 0.032
cps-fDAG (Ours)	$\mathbf{15 \pm 1}$	$\mathbf{0.662 \pm 0.020}$	$\mathbf{0.056 \pm 0.017}$	$\mathbf{0.776 \pm 0.016}$	$\mathbf{18 \pm 2}$	$\mathbf{0.551 \pm 0.029}$	$\mathbf{0.023 \pm 0.008}$	$\mathbf{0.701 \pm 0.026}$

B.3 Additional Analysis on fMRI Benchmark

The fMRI NetSim benchmark dataset (Smith et al., 2011) [5] is a widely used benchmark for evaluating causal discovery methods on functional magnetic resonance imaging (fMRI) data. NetSim is constructed upon realistic biophysical forward models, specifically integrating Dynamic Causal Modeling (DCM) with nonlinear balloon hemodynamics, to closely emulate continuous BOLD dynamics under physiological confounds such as spatial hemodynamic response variability and observation noise. Figure 3 presents the estimated graphs.

B.4 Additional Details on Ablation Study

B.4.1 Ablation Variants and Kendall tau

w/o Shared Basis. In contrast to the design in Section 4.1, for each functional node j, we independently compute its marginal covariance operator $\hat{C}_{jj}$ and extract the leading K eigenfunctions $\{\hat{\psi}_{j\ell}\}_{\ell=1}^{L}$ via functional principal component analysis (FPCA):

$$\hat{C}_{jj}\hat{\psi}_{j\ell} = \hat{\lambda}_{j\ell}\hat{\psi}_{j\ell}.$$

The corresponding scores are computed as

$$\widetilde{\theta}_{ij\ell} = \left\langle X_{ij} - \hat{\mu}_j, \hat{\psi}_{j\ell} \right\rangle.$$

Components with the same rank are directly treated as the same spectral block across nodes, without cross-node alignment or component matching.

w/o HSIC Score. We replace the proposed ordering score $\mathcal{S}_L$ with a Gaussian profile likelihood. Different from Equation 19 and Equation 20, the alternative criterion is

$$\hat{\mathcal{S}}_{L_G}(\sigma) = \frac{n}{2} \sum_{\ell=1}^{L} \sum_{j=1}^{p} \log\left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Gamma}_{ij\ell}^2 \right),$$

and the estimated ordering is

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_p} \hat{\mathcal{S}}_{L_G}(\sigma).$$

w/o Hierarchical Pruning. We remove the cross-spectral hierarchical coupling by setting $\alpha = 0$ in Equation 25, leaving only independent blockwise sparsity regularization. The resulting flat

[5] https://www.fMRIb.ox.ac.uk/datasets/netsim

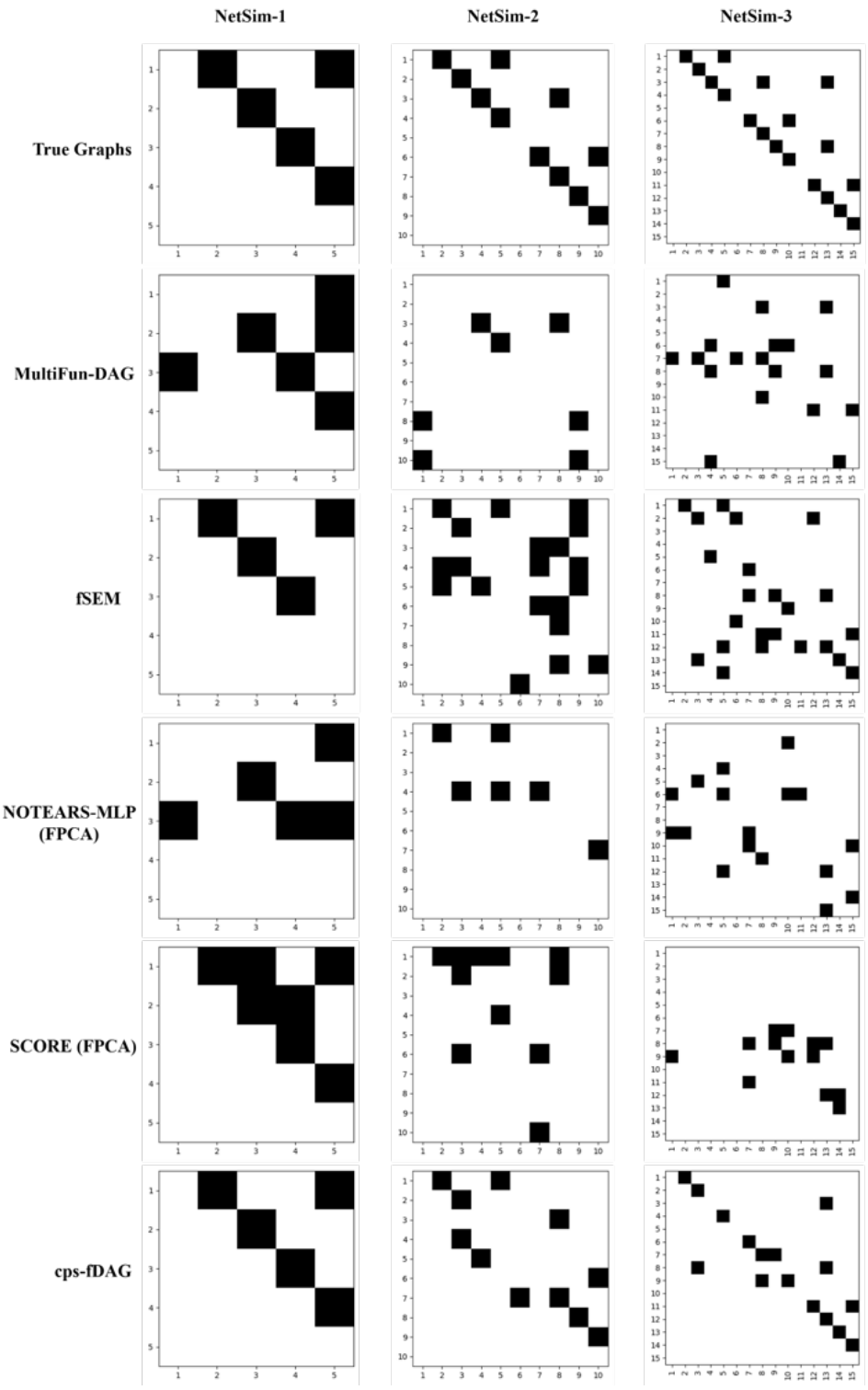


Figure 3: The estimated graphs obtained from different methods on the fMRI NetSim benchmark.

blockwise RKHS sparsity penalty as

$$\mathcal{P}_{\text{flat}}(\mathbf{f}_{jk}) = \sum_{\ell=1}^{L} \|f_{jk\ell}\|_{\mathcal{H}_\kappa}.$$

Kendall tau distance. The Kendall tau distance between two rankings is the number of pairs that are in different order in the two lists τ_1 and τ_2 is

$K_d(\tau_1, \tau_2) = |\{(i,j) : i < j, [\tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j)] \vee [\tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j)]\}|$,

where $\tau_1(i)$ and $\tau_2(i)$ are the rankings of the element i in τ_1 and τ_2 respectively. The normalized Kendall tau distance τ is

$$\tau = \frac{2K_d}{p(p-1)},$$

which lies in the interval $[0, 1]$ and lower is better.

B.4.2 Additional Ablation Study Results

We extend ablation studies to Scale-Free (SF-2, SF-4) topologies, dense graphs (ER-4),and high-dimensional regimes ($p = 50$) to verify the necessity of the cps-fDAG components. Table 8 reports the additional ablation results.

Table 8: Extended ablation study across diverse graph topologies, densities, and dimensions.

Setting & Variant	SHD ↓	TPR ↑	FDR ↓	F1 ↑	Kendall τ ↓
SF-2 ($p = 20, n = 500$)					
w/o Shared Basis	35 ± 2	0.359 ± 0.020	0.464 ± 0.029	0.424 ± 0.015	0.315 ± 0.013
w/o HSIC Score	14 ± 1	0.738 ± 0.026	0.162 ± 0.029	0.782 ± 0.022	0.245 ± 0.011
w/o Hierarchical Pruning	15 ± 1	0.757 ± 0.035	0.192 ± 0.026	0.777 ± 0.026	$\mathbf{0.233 \pm 0.020}$
cps-fDAG (Full)	$\mathbf{13 \pm 1}$	$\mathbf{0.773 \pm 0.030}$	$\mathbf{0.144 \pm 0.016}$	$\mathbf{0.810 \pm 0.022}$	$\mathbf{0.233 \pm 0.020}$
SF-4 ($p = 20, n = 500$)					
w/o Shared Basis	66 ± 3	0.399 ± 0.019	0.515 ± 0.027	0.436 ± 0.021	0.306 ± 0.017
w/o HSIC Score	34 ± 3	0.699 ± 0.025	0.287 ± 0.033	0.705 ± 0.028	0.272 ± 0.018
w/o Hierarchical Pruning	35 ± 2	$\mathbf{0.789 \pm 0.015}$	0.285 ± 0.015	0.749 ± 0.013	$\mathbf{0.191 \pm 0.015}$
cps-fDAG (Full)	$\mathbf{28 \pm 3}$	0.780 ± 0.016	$\mathbf{0.214 \pm 0.030}$	$\mathbf{0.781 \pm 0.020}$	$\mathbf{0.191 \pm 0.015}$
ER-2 ($p = 50, n = 1000$)					
w/o Shared Basis	86 ± 4	0.397 ± 0.021	0.372 ± 0.028	0.483 ± 0.020	0.356 ± 0.014
w/o HSIC Score	27 ± 2	0.793 ± 0.011	0.078 ± 0.013	0.852 ± 0.009	0.271 ± 0.019
w/o Hierarchical Pruning	32 ± 2	0.791 ± 0.013	0.113 ± 0.015	0.835 ± 0.009	$\mathbf{0.249 \pm 0.011}$
cps-fDAG (Full)	$\mathbf{26 \pm 2}$	$\mathbf{0.801 \pm 0.012}$	$\mathbf{0.068 \pm 0.013}$	$\mathbf{0.861 \pm 0.009}$	$\mathbf{0.249 \pm 0.011}$
ER-4 ($p = 20, n = 500$)					
w/o Shared Basis	79 ± 3	0.457 ± 0.007	0.513 ± 0.015	0.470 ± 0.008	0.297 ± 0.022
w/o HSIC Score	$\mathbf{42 \pm 4}$	0.762 ± 0.027	0.312 ± 0.025	0.722 ± 0.024	0.195 ± 0.022
w/o Hierarchical Pruning	48 ± 3	$\mathbf{0.802 \pm 0.021}$	0.350 ± 0.017	0.716 ± 0.015	$\mathbf{0.184 \pm 0.017}$
cps-fDAG (Full)	$\mathbf{42 \pm 4}$	0.793 ± 0.020	$\mathbf{0.296 \pm 0.023}$	$\mathbf{0.744 \pm 0.020}$	$\mathbf{0.184 \pm 0.017}$

B.5 Additional Details on Robustness Analysis

B.5.1 Cross-basis contamination mechanism

When the data do not fully satisfy causal partial separability, $\bigcup_{\ell=1}^{L} \hat{\mathcal{E}}_\ell$ is no longer guaranteed to consistently recover the full functional graph $\mathcal{E}_0$. Nevertheless, as established in Section 4.1, the eigenbasis of $\hat{\mathcal{H}}$ optimally preserves the total variability of $(\langle X_1, \hat{\phi}_\ell \rangle, \ldots, \langle X_p, \hat{\phi}_\ell \rangle)^\top$ under orthogonality constraints, irrespective of CPS. The learned structure is therefore inherently valuable as a principal component approximation of $\mathcal{E}_0$.

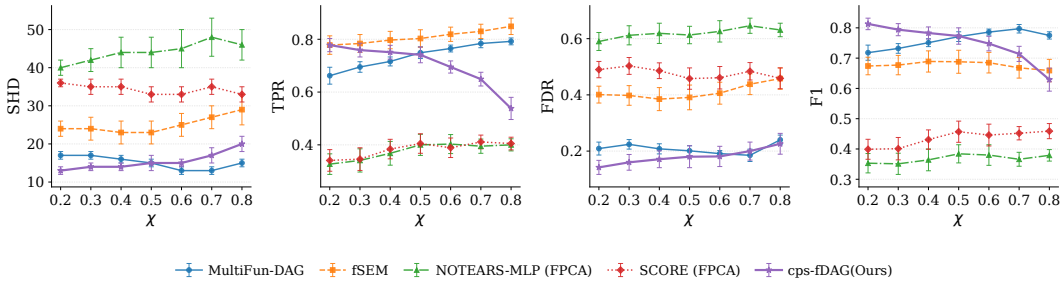
We introduce $\chi \in [0, 1)$ as a cross-basis contamination parameter. Here, $\chi = 0$ corresponds to the standard CPS setting, while increasing to 1 systematically amplifies the strength of cross-basis causal interventions. Specifically, the structural updating rule for the parameter vector $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jL})^\top$ associated with node j is formulated as follows. For each dimension $\ell \in \{1, \ldots, L\}$, the latent state $\theta_{j\ell}$ is modeled as

$$\theta_{j\ell} = \sum_{k \in \mathrm{pa}(j)} \left[\sqrt{1 - \chi^2}\, f(\theta_{k\ell}) + \chi \sum_{r \neq \ell} f(\theta_{kr}) \right] + \epsilon_{j\ell},$$

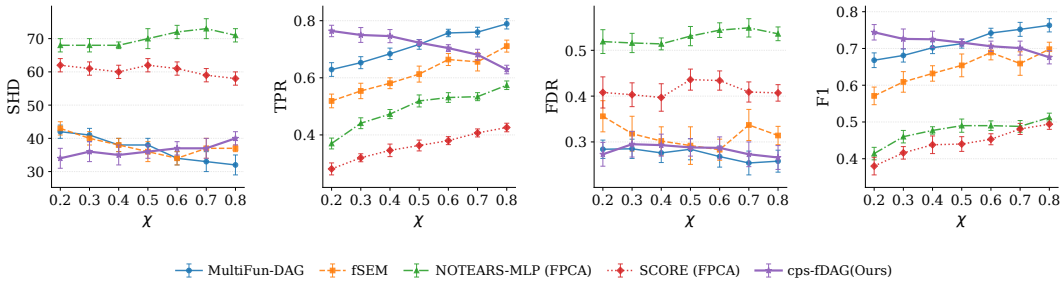
where $\eta \in [0, 1)$ serves as a trade-off coefficient that balances direct component-wise inheritance against cross-component interactions.

B.5.2 Additional Results of Robustness Analysis

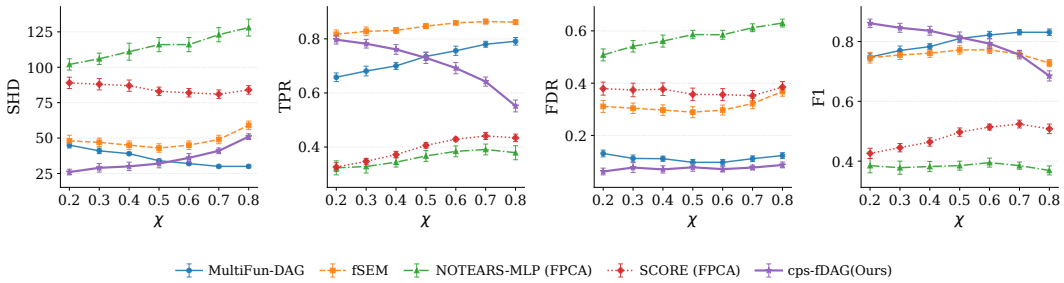
We extend the cross-basis contamination experiments described in Section 5 across diverse graph topologies, higher node dimensions, and variable spectral decay rates. Figure 4 presents the detailed structural recovery metrics.



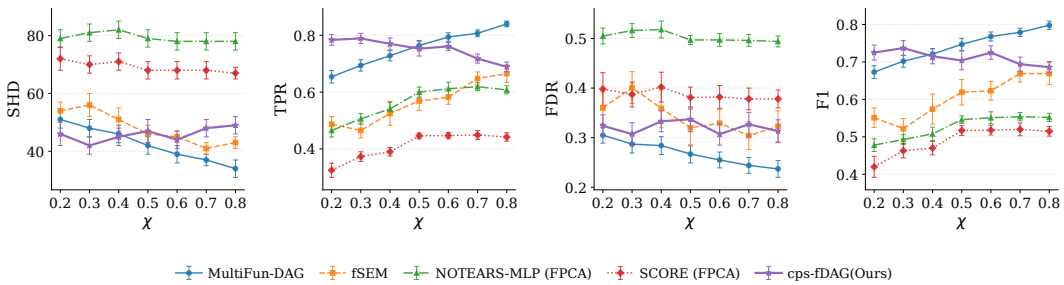
(a) SF-2 ($p = 20$, $n = 500$).



(b) SF-4 ($p = 20$, $n = 500$).



(c) ER-2 ($p = 50$, $n = 1000$).



(d) ER-4 ($p = 20$, $n = 500$).

Figure 4: Robustness of all metrics with respect to χ across different synthetic graph settings.