LEARNING DYNAMIC CAUSAL GRAPHS UNDER PARAMETRIC UNCERTAINTY VIA POLYNOMIAL CHAOS EXPANSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing causal discovery methods are fundamentally limited by the assumption of a static causal graph, a constraint that fails in real-world systems where causal relationships dynamically vary with underlying system parameters. This discrepancy prevents the application of causal discovery in critical domains such as industrial process control, where understanding how causal effects change is essential. We address this gap by proposing a new paradigm that moves beyond static graphs to learn functional causal representations. We introduce a framework that models each causal link not as a static weight but as a function of measurable system parameters. By representing these functions using Polynomial Chaos Expansions (PCE), we develop a tractable method to learn the complete parametric causal structure from observational data. We provide theoretical proofs for the identifiability of these functional models and introduce a novel, provably convergent learning algorithm. On a large-scale chemical reactor dataset, our method learns the dynamic causal structure with a 90.9% F1-score, nearly doubling the performance of state-of-the-art baselines and providing an interpretable model of how causal mechanisms evolve.

1 INTRODUCTION

Industrial process control systems generate massive volumes of sensor data requiring automated analysis for optimization and predictive maintenance (Fang et al., 2022; Zhou et al., 2015). Understanding causal relationships between process variables is essential for root cause analysis, anomaly detection, and adaptive control strategies (Zhang et al., 2016). However, industrial processes exhibit unique challenges that violate assumptions of existing causal discovery methods: causal relationships may vary systematically with operating conditions, sensors exhibit complex multi-modal and heavy-tailed noise distributions, and safety-critical applications demand rigorous uncertainty quantification (Cao et al., 2025; Wang et al., 2025).

Many widely used causal discovery methods for observational data are formulated in terms of a single, static causal graph whose edge strengths do not depend on observed context or operating parameters, even though there is a growing body of work on time-varying and context-specific causal structures (Song et al., 2009; Huang et al., 2019). In reality, industrial causal effects are functions of measurable parameters. For instance, in chemical reactors, the influence of feed temperature on product quality depends strongly on catalyst activity, which degrades over time. Heat exchanger effectiveness varies with fouling levels, fundamentally altering thermal control loops. These parametric dependencies are not mere nuisances but contain critical information for process optimization and predictive maintenance. Our goal in this paper is therefore not to replace existing approaches to epistemic or aleatoric uncertainty, but to complement them with a representation in which each causal edge is an explicit function of a low-dimensional vector of operating parameters.

The field of causal discovery has evolved through three major directions, each with distinct limitations for industrial applications. Constraint-based methods such as the Peter-Clark (PC) algorithm (Spirtes et al., 2000), Fast Causal Inference (FCI) (Spirtes et al., 2001), and Really Fast Causal Inference (RFCI) (Colombo et al., 2012) rely on conditional independence tests to infer causal structures. While theoretically sound, these methods struggle with finite sample sizes and become

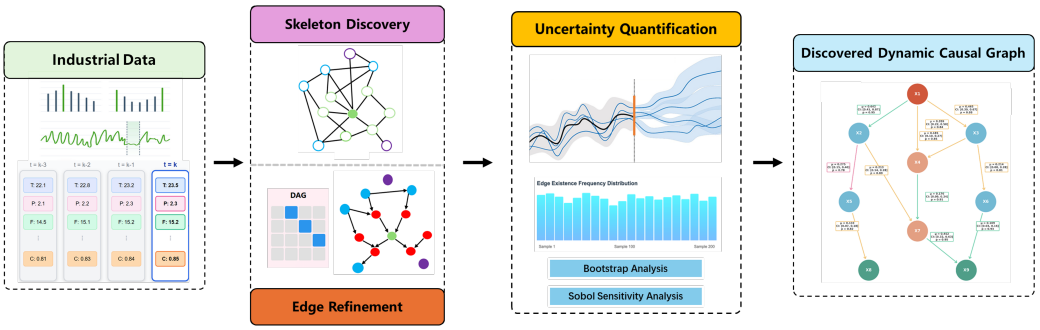

Figure 1: Overview of the polynomial chaos theory for causal discovery framework in dynamic uncertainty systems.

computationally intractable for high-dimensional industrial data with hundreds of sensors. Recent advances like PC-stable (Colombo & Maathuis, 2014) improve stability but fail when faced with complex noise distributions common in industrial sensors.

Score-based methods including Greedy Equivalence Search (GES) (Chickering, 2002) and Fast Greedy Equivalence Search (FGES) (Ramsey et al., 2017) optimize scoring functions over possible Directed Acyclic Graph (DAG) structures. The breakthrough NOTEARS algorithm (Zheng et al., 2018) reformulated structure learning as continuous optimization with differentiable acyclicity constraints, spawning variants like Directed Acyclic Graph - Graph Neural Network (DAG-GNN) (Yu et al., 2019) for nonlinear relationships, Reinforcement Learning - Bayesian Information Criterion (RL-BIC) (Zhu et al., 2020) using reinforcement learning. In their most common instantiations, these approaches return a single best-scoring DAG and point estimates of edge weights; uncertainty over graphs and parameters is typically handled by separate Bayesian or bootstrap procedures (e.g., Lorch et al., 2021; Cundy et al., 2021) rather than being integrated with an explicit model of how edge strengths vary with operating parameters. This limitation is particularly critical for safety-critical industrial applications where confidence in causal recommendations directly impacts operational decisions and safety outcomes.

Functional causal models exploit asymmetries in data distributions for identifiability. The Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006) proved that linear models with non-Gaussian noise yield unique causal structures, later extended to DirectLiNGAM (Shimizu et al., 2011) and Vector Autoregressive LiNGAM (VAR-LiNGAM) (Hyvärinen et al., 2010) for time series. Nonlinear extensions include additive noise models (ANM) (Hoyer et al., 2008), post-nonlinear models (PNL) (Zhang & Hyvärinen, 2009), and the general identifiable functional causal model framework (Peters et al., 2014). However, in their standard form these models do not represent how causal effects change as an explicit function of observed operating parameters.

Recent industrial applications (Sui et al., 2025; Runge et al., 2019; Yang et al., 2025) have highlighted these limitations, often requiring extensive preprocessing or domain-specific modifications lacking theoretical justification. Bayesian approaches such as Differentiable Bayes for Structure Learning (DiBS) (Lorch et al., 2021) and Bayesian Causal Discovery with Neural Networks (BCD Nets) (Cundy et al., 2021) explicitly quantify posterior uncertainty over graphs and parameters, addressing epistemic uncertainty due to finite data, but they still treat each edge as static and do not model how its strength varies with operating parameters.

Polynomial Chaos Expansion (PCE), introduced by Wiener (Wiener, 1938) and generalized by Xiu (Xiu & Karniadakis, 2002), provides a mathematically rigorous framework for representing and propagating uncertainty through complex systems. PCE has been successfully applied in uncertainty quantification for engineering systems (Sudret, 2008), sensitivity analysis (Crestaux et al., 2009), and stochastic optimization (Picheny et al., 2013). Recent algorithmic advances including sparse PCE (Jiang et al., 2025), adaptive basis selection (Dai et al., 2025), and multi-fidelity approaches (Liu et al., 2020) have made PCE computationally tractable for high-dimensional problems. Despite this success in forward uncertainty propagation, PCE has, to the best of our knowledge, not yet been systematically exploited for causal discovery.

This paper introduces PCT-CD, bridging uncertainty quantification and causal discovery. Figure 1 provides an overview of the proposed framework. Our contributions are summarized as follows:

1. **From Static Graphs to Dynamic Functions:** We formalize an industrial structural causal model in which a single underlying DAG is equipped with edge weights that are explicit functions of operating conditions, and we prove identifiability of these parametric mechanisms under mild assumptions.
2. **An End-to-End Solution for Dynamic Systems:** We propose PCT-CD, an integrated algorithm specifically designed for parametric causal discovery. It translates complex process data into an interpretable model of how causal links evolve, providing actionable intelligence beyond simple correlation.
3. **Demonstrated Real-World Viability:** We empirically validate PCT-CD on controlled synthetic benchmarks with parameter-varying mechanisms and on a complex refinery dataset. PCT-CD achieves high F1-scores and equips engineers with uncertainty-aware tools that are essential for making robust decisions in high-stakes environments.

The remainder of this paper is organized as follows. Section 2 details the PCT-CD methodology, including parametric SEM formulation, PCE representation, and our novel conditional independence test. Section 3 establishes theoretical foundations with identifiability and convergence proofs. Section 4 validates our approach on both synthetic benchmarks with controlled parametric variation and an industrial process dataset, comparing against a broad set of baseline methods. Section 5 concludes with implications for industrial deployment and future research directions.

2 METHODOLOGY

Our proposed framework, PCT-CD, introduces a novel paradigm for causal discovery by explicitly modeling how causal relationships vary as functions of measurable system parameters. This is achieved by integrating the theory of PCE into a hybrid structure learning algorithm. While Bayesian and bootstrap-based methods typically quantify epistemic uncertainty arising from finite data (for example via posterior distributions over static graphs and parameters), PCT-CD is designed to address *parametric* uncertainty by representing causal edges as explicit functions of measurable system parameters, and is conceptually complementary to these existing approaches.

The methodology unfolds in four stages: first, we formulate a Structural Equation Model (SEM) where causal coefficients are functions of a parameter vector $\boldsymbol{\xi}$. Second, we represent these functions using PCE, transforming the non-parametric problem into a tractable parametric one. Third, we develop a novel conditional independence test tailored to this representation to discover an initial causal skeleton. Finally, we refine this structure and quantify edge strengths using a score-based optimization with a natural gradient approach, ensuring both accuracy and computational efficiency.

2.1 PROBLEM FORMULATION

We consider a complete probability space $(\Xi, \mathcal{F}, \mathbb{P})$ where all random quantities are defined. The core innovation of our framework is the explicit modeling of parametric uncertainty through a random vector $\boldsymbol{\xi} \in \Xi \subset \mathbb{R}^d$. This vector represents known, measurable operating conditions (e.g., ambient temperature, catalyst age, feedstock quality) with a joint probability distribution μ_ξ that has finite moments of all orders. In our theoretical analysis we assume that μ_ξ is known so that a standard PCE basis adapted to μ_ξ can be chosen; in practice, when only samples of $\boldsymbol{\xi}$ are available, an empirical orthogonal basis can be constructed from the observed parameter values (see Appendix for details). This formulation emphasizes a dimension that is often implicit in causal discovery, where operating conditions are typically treated as fixed and uncertainty is mainly modeled as arising from finite data and stochastic noise.

We observe n process variables, collected in a vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$. We assume that these variables are generated by a linear SEM (equivalently, a linear structural causal model in the sense of structural causal inference) where the causal relationships are functions of the parameter vector $\boldsymbol{\xi}$:

$$X_i = \sum_{j \in \mathbf{PA}_i} b_{ij}(\boldsymbol{\xi}) X_j + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{PA}_i \subset \{1, \ldots, n\} \setminus \{i\}$ denotes the set of causal parents of variable X_i, the functions $b_{ij}(\boldsymbol{\xi}) \in L^2(\Xi)$ are unknown, square-integrable functions capturing the parameter-dependent causal effects, and ϵ_i are mutually independent, centered, sub-Gaussian noise terms. The underlying causal structure forms a DAG $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, n\}$ and an edge $(j, i) \in E$ exists if and only if $j \in \mathbf{PA}_i$. We make the standard assumptions of causal sufficiency (no unmeasured common causes) and faithfulness (all conditional independencies in the data are consequences of d-separation in $\mathcal{G}$). Throughout, the edge set E does not depend on $\boldsymbol{\xi}$; only the edge weights $b_{ij}(\boldsymbol{\xi})$ vary with operating conditions. We assume m i.i.d. samples $\{(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})\}_{t=1}^m$ from this model, where t indexes samples rather than time.

2.2 Polynomial Chaos Representation

The central challenge is to learn the functions $b_{ij}(\boldsymbol{\xi})$, the true causal strength varying with system parameters $\boldsymbol{\xi}$. We address this by representing each causal coefficient function using a PCE. For many choices of μ_ξ, including the classical Wiener–Askey scheme, there exists a corresponding basis of orthogonal multivariate polynomials $\{\Psi_\alpha(\boldsymbol{\xi})\}_{\alpha \in \mathbb{N}^d}$ adapted to μ_ξ (e.g., Xiu & Karniadakis, 2002; Sudret, 2008; Crestaux et al., 2009). Common examples include Hermite polynomials for Gaussian parameters, Legendre for uniform and Laguerre for exponential.

Any square-integrable function $b_{ij}(\boldsymbol{\xi})$ can be expanded in this basis. By truncating the expansion at a total polynomial degree N_p, we obtain a finite-dimensional approximation:

$$b_{ij}(\boldsymbol{\xi}) \approx \sum_{\alpha \in \mathcal{A}_{N_p}} \theta_{ij,\alpha} \Psi_\alpha(\boldsymbol{\xi}) \tag{2}$$

where $\mathcal{A}_{N_p} := \{\alpha \in \mathbb{N}^d : |\alpha| = \sum_{k=1}^d \alpha_k \leq N_p\}$ is the set of multi-indices, and the coefficients $\theta_{ij,\alpha}$ are the spectral projections of the function onto the basis, given by $\theta_{ij,\alpha} = \langle b_{ij}(\boldsymbol{\xi}), \Psi_\alpha(\boldsymbol{\xi}) \rangle_{L^2} / \langle \Psi_\alpha^2 \rangle_{L^2}$. The cardinality of the basis is $P = |\mathcal{A}_{N_p}| = \binom{N_p+d}{d}$. This representation converts the infinite-dimensional problem of learning functions $b_{ij}(\boldsymbol{\xi})$ into a finite-dimensional problem of estimating the spectral coefficients $\theta_{ij,\alpha}$.

For functions that are continuously differentiable s times, the spectral error decays polynomially: $\|b_{ij} - \Pi_{N_p} b_{ij}\|_{L^2} \leq C N_p^{-s}$, where Π_{N_p} is the projection operator. For analytic functions, which are common in physical systems, convergence is exponential: $\|b_{ij} - \Pi_{N_p} b_{ij}\|_{L^2} \leq C \exp(-\gamma N_p^{1/d})$ (e.g., Sudret, 2008; Crestaux et al., 2009). For high-dimensional parameter spaces ($d \gg 1$), the basis size P can become computationally prohibitive. We employ hyperbolic truncation schemes, which prioritize low-order interaction terms and significantly reduce the basis size while often retaining high accuracy for functions with decaying importance of higher-order interactions (Jiang et al., 2025; Dai et al., 2025).

2.3 PCT-Conditional Independence Test

Once we have established the PCE representation, our initial goal is to identify the causal skeleton. Standard conditional independence (CI) tests that operate on the marginal distribution of (X_A, X_B, X_Z) can fail in the presence of parameter-varying mechanisms: a causal relationship $b_{ij}(\boldsymbol{\xi})$ may have zero mean but large variance, so that marginal tests falsely conclude independence even though the dependence is strong for many values of $\boldsymbol{\xi}$. We therefore seek a test that is sensitive to dependence across the entire parameter space rather than only to its average effect.

We define PCT-conditional independence as the vanishing of the expected conditional covariance over the parameter space. That is, variables X_A and X_B are PCT-conditionally independent given a set of variables X_Z if:

$$\mathbb{E}_{\boldsymbol{\xi}}[\mathrm{Cov}(X_A, X_B | X_Z, \boldsymbol{\xi})] = 0 \tag{3}$$

This condition implies that the partial correlation between X_A and X_B after accounting for X_Z is zero for almost every value of $\boldsymbol{\xi}$. Based on this definition, we derive a statistical test (Algorithm 1). The core idea is to estimate the PCE coefficients of the conditional covariance function $C_{AB|Z}(\boldsymbol{\xi}) = \mathrm{Cov}(X_A, X_B | X_Z, \boldsymbol{\xi})$. The null hypothesis $X_A X_B X_Z$ is equivalent to all PCE coefficients of $C_{AB|Z}(\boldsymbol{\xi})$ being zero. A detailed derivation of the test statistic and proofs of its asymptotic properties are provided in the Appendix.

Under the null hypothesis and mild regularity conditions (finite fourth moments of the residuals, i.i.d. samples conditional on $\boldsymbol{\xi}$, and bounded basis functions), the estimators for the standardized coefficients are asymptotically independent and approximately standard normal (see Appendix for a detailed proof). Consequently, the test statistic T_{PCT} follows a chi-squared distribution with P degrees of freedom:

$$T_{PCT} = m \sum_{\alpha \in \mathcal{A}_{N_p}} \left(\frac{\hat{C}_{AB|Z,\alpha}}{\hat{\sigma}_{AB|Z,\alpha}} \right)^2 \xrightarrow{d} \chi^2_P \tag{4}$$

This test is used within a standard constraint-based algorithm to learn an initial graph structure by systematically testing for conditional independencies. In our implementation we use a PC-style skeleton search in which PCT-CI replaces the usual CI oracle.

2.4 Score-Based Learning with Natural Gradient

Although constraint-based methods are effective for skeleton discovery, they can be unstable with finite data. We therefore use the output of the constraint-based phase as an initialization for a more robust score-based optimization. We formulate structure learning as the optimization of a penalized likelihood score over the space of DAGs and PCE coefficients and define the PCT-BIC score as:

$$\mathcal{S}(E, \Theta) = \frac{1}{2\sigma_\epsilon^2} \sum_{t=1}^{m} \sum_{i=1}^{n} \left(X_i^{(t)} - \sum_{j \in \mathbf{PA}_i} \sum_{\alpha \in \mathcal{A}_{N_p}} \theta_{ij,\alpha} \, \Psi_\alpha(\boldsymbol{\xi}^{(t)}) \, X_j^{(t)} \right)^2 + \lambda \|(E, \Theta)\|_0 \tag{5}$$

where $\Theta = \{\theta_{ij,\alpha}\}$ is the collection of all PCE coefficients, and the group sparsity penalty $\|(E, \Theta)\|_0 = \sum_{i,j} \mathbf{1}\{\|_{ij}\|_2 > 0\}$ encourages sparse DAGs by penalizing the number of non-zero causal links, with $\lambda = \frac{P}{2} \log(m)$.

Optimizing this score is challenging due to the combinatorial nature of the graph space and the high dimensionality of the parameter space Θ. We employ a greedy search strategy combined with efficient gradient-based optimization of the coefficients for a given graph structure, accepting edge additions or deletions only when they preserve acyclicity of $\mathcal{G}$. The parameter space of PCE coefficients has a natural Riemannian geometry induced by the Fisher information matrix $\mathbf{F}(\Theta)$. The natural gradient, $\tilde{\nabla}_\Theta \mathcal{S} = \mathbf{F}^{-1} \nabla_\Theta \mathcal{S}$, respects this geometry and provides an update direction that converges significantly faster than the standard Euclidean gradient. The Fisher matrix entries are:

$$[\mathbf{F}]_{(ij,\alpha),(i'j',\alpha')} = \frac{1}{\sigma_\epsilon^2} \delta_{ii'} \mathbb{E}[X_j X_{j'}] \mathbb{E}_{\boldsymbol{\xi}}[\Psi_\alpha(\boldsymbol{\xi}) \Psi_{\alpha'}(\boldsymbol{\xi})] \tag{6}$$

Due to the orthogonality of the basis polynomials and the linear-Gaussian form of the likelihood, this matrix is block-diagonal across edges and diagonal within each block, making its inversion computationally efficient. The natural gradient descent update is:

$$\Theta \leftarrow \Theta - \eta \mathbf{F}^{-1} \nabla_\Theta \mathcal{S} \tag{7}$$

The complete PCT-CD algorithm, summarized in Algorithm 2 in the appendix, integrates these components into a multi-phase procedure that ensures both structural accuracy and robust parameter estimation. This provides not only the final graph and functional relationships but also confidence intervals for causal strengths and probabilities for the existence of each edge.

3 Theoretical Analysis

In this section, we establish the theoretical foundations of the PCT-CD framework. We prove that under reasonable conditions, the true parametric causal DAG is uniquely identifiable from observational data. Furthermore, we provide finite-sample guarantees for the recovery of the causal structure and analyze the convergence properties of our optimization procedure. Formal statements and proofs of the main results are deferred to the Appendix for clarity.

3.1 Assumptions and Preliminaries

Assumption 1 (Data-generating process). *The variables $\mathbf{X} = (X_1, \ldots, X_n)$ obey the SEM equation 1 with a DAG $\mathcal{G}$, where each $b_{ij}(\boldsymbol{\xi}) \in L^2(\mu_\xi)$ admits the chaos expansion equation 2 with truncation bias controlled by polynomial convergence theory. The parameter vector $\boldsymbol{\xi}$ is independent of noises $\{\epsilon_i\}$ and has a known distribution μ_ξ with finite moments of all orders. In the empirical setting μ_ξ may be approximated by the empirical measure of observed operating conditions, as discussed in Section 2.2 and Appendix A.1.*

Assumption 2 (Noise). *The disturbances ϵ_i are mutually independent, centered, sub-Gaussian with proxy σ_ϵ^2 and finite fourth moments. At least one ϵ_i is non-Gaussian, or the collection of coefficient functions $\{b_{ij}(\boldsymbol{\xi})\}$ is* non-degenerate in $\boldsymbol{\xi}$ *(defined below). This disjunction provides two alternative routes to identifiability: either non-Gaussian noise (as in LiNGAM) or sufficiently rich parametric variation in the coefficient functions.*

Assumption 3 (Faithfulness and stability). *The joint law of $(\mathbf{X}, \boldsymbol{\xi})$ is faithful to $\mathcal{G}$, and the operator norm satisfies $\mathbb{E}[\|\mathbf{B}(\boldsymbol{\xi})\|_{\mathrm{op}}] < 1$, where $\mathbf{X} = \mathbf{B}(\boldsymbol{\xi})\mathbf{X} + \boldsymbol{\epsilon}$ is the matrix form with $[\mathbf{B}(\boldsymbol{\xi})]_{ij} = b_{ij}(\boldsymbol{\xi})$ for $j \in \mathbf{PA}_i$ and zero otherwise. This stability condition ensures that $(\mathbf{I} - \mathbf{B}(\boldsymbol{\xi}))$ is invertible almost surely and that the resulting mixing matrix admits a finite-variance chaos expansion.*

Definition 1 (Non-degeneracy). *A set of functions $\{b_{ij}(\cdot)\}$ is non-degenerate if their PCE coefficient matrices $\{\boldsymbol{\theta}_{ij}\}$ are linearly independent across edges, i.e., no non-trivial linear combination of coefficient vectors vanishes almost surely in $\boldsymbol{\xi}$. Intuitively, non-degeneracy means that different edges leave distinguishable "signatures" in the parameter space, which can be used to orient the graph even when all noises are Gaussian.*

3.2 Identifiability of Parametric Causal Structures

Identifiability is the cornerstone of any causal discovery method, ensuring that the underlying causal structure can, in principle, be recovered from the joint distribution of the observed variables. We extend the classical results of LiNGAM to our parametric setting.

Theorem 1 (PCT Identifiability). *Under Assumptions 1–3, if either (i) at least one ϵ_i is non-Gaussian, or (ii) all ϵ_i are Gaussian but the PCE coefficient matrices $\{\boldsymbol{\theta}_{ij}\}$ are linearly independent (non-degenerate in the sense of Definition 1), then the DAG $\mathcal{G}$ and the parametric functions $\{b_{ij}(\boldsymbol{\xi})\}$ are identifiable from the joint distribution of $(\mathbf{X}, \boldsymbol{\xi})$, up to the usual scale and permutation indeterminacies of the disturbances.*

The proof (Appendix A.2) proceeds by writing the model in mixing-matrix form, applying a LiNGAM-style argument to the zeroth-order chaos coefficients, and then using linear independence of the higher-order PCE coefficients to rule out non-trivial alternative parameterizations.

3.3 Finite-Sample Guarantees and Consistency

While identifiability ensures recovery from the true distribution, practical algorithms operate on finite data samples. This section addresses the question: How many samples are required to reliably recover the true causal graph? We establish the consistency of our coefficient estimators and provide explicit sample complexity bounds for exact graph recovery. PCT-CD estimates the PCE coefficients $\theta_{ij,\alpha}$ through least-squares regression. The orthogonality of the polynomial basis $\{\Psi_\alpha\}$ ensures favorable statistical properties for these estimators.

Proposition 1 (Coefficient Consistency). *Let $\hat{\theta}_{ij,\alpha}$ be the estimator for $\theta_{ij,\alpha}$ obtained via least-squares regression. Under the model assumptions, for each i, j, α:*

$$\hat{\theta}_{ij,\alpha} \xrightarrow{p} \theta_{ij,\alpha} \quad \text{as } m \to \infty \tag{8}$$

The consistency follows from standard regression theory, where the orthogonality of basis functions ensures that estimators are unbiased with variance decaying as $O(1/m)$. The core challenge in graph recovery is distinguishing genuine causal edges from spurious connections. For PCT-CD, determining whether edge (i, j) exists is equivalent to testing whether the coefficient function $b_{ij}(\boldsymbol{\xi})$ is identically zero. We employ the L^2-norm as our detection criterion:

$$\|b_{ij}\|_{L^2}^2 = \sum_{\alpha \in \mathcal{A}_{N_p}} \theta_{ij,\alpha}^2 \langle \Psi_\alpha^2 \rangle \tag{9}$$

The decision rule is:

$$\text{Edge } (i,j) \text{ exists} \Leftrightarrow \|b_{ij}\|_{L^2} > 0 \tag{10}$$

$$\text{No edge } (i,j) \Leftrightarrow \|b_{ij}\|_{L^2} = 0 \tag{11}$$

Let $\kappa = \min_{(i,j)\in E} \|b_{ij}\|_{L^2}$ denote the strength of the weakest true edge. This parameter fundamentally determines the difficulty of the detection problem.

Theorem 2 (Sample Complexity for Graph Recovery). *Assume the noise terms ϵ_i are sub-Gaussian with variance proxy σ_ϵ^2. For the PCT-CD algorithm to recover the true DAG $\mathcal{G}$ with probability at least $1-\delta$, a sufficient number of samples is:*

$$m \geq C \frac{\sigma_\epsilon^2 P}{\kappa^2} \log\left(\frac{2n^2 P}{\delta}\right) \tag{12}$$

where C is a constant depending on the data distribution and polynomial basis, $P = |\mathcal{A}_{N_p}|$ is the number of PCE basis functions, and κ is the minimum edge strength.

The proof (Appendix A.3) combines sub-Gaussian concentration bounds for the least-squares estimates, a union bound over all coefficients and edges, and the separation margin κ between zero and non-zero coefficient functions. The bound highlights the expected trade-offs: sample complexity grows linearly in the basis size P and noise level σ_ϵ^2, and quadratically in the inverse of the weakest edge strength.

3.4 Convergence Analysis

Finally, we analyze the convergence of the coefficient optimization phase, which is critical for the efficiency of the score-based refinement. The use of natural gradient descent is key to achieving rapid convergence.

Theorem 3 (Natural Gradient Convergence). *Let the PCT-BIC score $\mathcal{S}(\Theta)$ be restricted to a fixed DAG. Assume the objective is μ-strongly convex and L-smooth in a neighborhood of the optimum Θ^*. The natural gradient descent update with a constant step size η satisfies:*

$$\|\Theta^{(t+1)} - \Theta^*\|_{\mathbf{F}} \leq \left(1 - \eta \frac{\mu}{L_{\mathbf{F}}}\right) \|\Theta^{(t)} - \Theta^*\|_{\mathbf{F}} \tag{13}$$

where $\|\cdot\|_{\mathbf{F}}$ is the norm induced by the Fisher matrix, and $L_{\mathbf{F}}$ is the smoothness constant in the Riemannian metric. With an optimal step size, this leads to a linear convergence rate of $(1-\rho)$, where $\rho = \lambda_{\min}(\mathbf{F})/\lambda_{\max}(\mathbf{F})$ is the condition number of the Fisher information matrix.

The analysis in Appendix A.4 exploits the fact that, in our linear-Gaussian setting with an orthogonal PCE basis, the Fisher matrix is block-diagonal across edges and diagonal within blocks. This structure yields a well-conditioned Riemannian metric and explains the empirically fast convergence of natural gradient updates compared to standard Euclidean gradient descent.

4 Experimental Validation

We validate PCT-CD on a comprehensive industrial process dataset comprising 10,000 samples from a chemical reactor network at Parkland Refinery in Canada. The system monitors 9 critical process variables including feed temperatures, reactor pressures, product quality indicators, and flow rates, with 11 established causal relationships verified through process engineering principles and operational expertise. The system exhibits parametric uncertainty from three primary sources: heat transfer coefficients (ξ_1), reaction rate constants (ξ_2), and yield factors (ξ_3), making it ideal for demonstrating the advantages of modeling parameter-dependent causal relationships.

4.1 Experimental Setup

The dataset represents a hierarchical chemical process where variables form a network structure with source nodes (feed streams), intermediate processing stages (reactors and separators), and terminal outputs (product quality metrics). Each sample includes simultaneous measurements of all

Table 1: Performance Comparison Across All Methods

Method	TP	FP	FN	Prec.	Recall	F1	SHD
ICA-LiNGAM	1	14	10	0.067	0.091	0.077	24
DirectLiNGAM	2	13	9	0.133	0.182	0.154	22
VAR-LiNGAM	3	7	8	0.300	0.273	0.286	15
RECI	4	9	7	0.308	0.364	0.333	16
PCMCI	4	8	7	0.333	0.364	0.348	15
CCD	5	11	6	0.312	0.455	0.370	17
LiNGAM	5	10	6	0.333	0.455	0.385	16
ElasticNet	5	8	6	0.385	0.455	0.417	14
Entropy-Based	5	6	6	0.455	0.455	0.455	12
GP-Based	5	5	6	0.500	0.455	0.476	11
NOTEARS	5	5	6	0.500	0.455	0.476	11
CGNN	7	10	4	0.412	0.636	0.500	14
Lasso-Granger	6	7	5	0.462	0.545	0.500	12
FCI	5	4	6	0.556	0.455	0.500	10
PC	5	4	6	0.556	0.455	0.500	10
ANM	6	6	5	0.500	0.545	0.522	11
PNL	6	6	5	0.500	0.545	0.522	11
GIES	6	5	5	0.545	0.545	0.545	10
GES	6	5	5	0.545	0.545	0.545	10
CAM	8	6	3	0.571	0.727	0.640	9
GraNDAG	8	6	3	0.571	0.727	0.640	9
SAM	8	6	3	0.571	0.727	0.640	9
PCT-CD	**10**	**1**	**1**	**0.909**	**0.909**	**0.909**	**2**

process variables along with the corresponding parameter values, captured under varying operational conditions over a 6-month period. The ground truth causal structure was established through a combination of process flow diagrams, material balance equations, and expert knowledge from plant operators. We compare PCT-CD against 23 state-of-the-art methods spanning six categories.

PCT-CD parameters were selected through cross-validation: $N_p = 4$ (polynomial degree), $\alpha_{sig} = 0.05$ (significance level for conditional independence tests), $\lambda = 1$ (regularization parameter), and $B = 200$ (bootstrap samples).

4.2 Performance Results

Table 1 presents comprehensive performance metrics across all methods. PCT-CD achieves exceptional performance with 90.9% F1-score, correctly identifying 10 out of 11 true edges (True Positives) with only 1 false positive and 1 false negative, resulting in a structural Hamming distance (SHD) of 2. This represents nearly double the performance of the best baseline methods.

Analyzing the results by method category reveals systematic patterns. Constraint-based methods (PC, FCI) achieve moderate precision (55.6%) but suffer from low recall (45.5%), indicating conservative edge detection that misses many true relationships. Score-based approaches (GES, GIES, NOTEARS) show balanced precision and recall around 50-54%, but their static graph assumption fundamentally limits performance. Among functional causal models, traditional LiNGAM performs poorly (38.5% F1-score) while ICA-LiNGAM shows the worst performance (7.7% F1-score), suggesting severe model misspecification under parametric variation.

Figure 2 visualizes the discovered causal structures across all 23 methods, providing a comprehensive comparison of graph recovery quality. The performance gap between PCT-CD (90.9% F1-score) and the next best methods (CAM, GraNDAG, SAM at 64.0%) highlights the value of explicit uncertainty modeling.

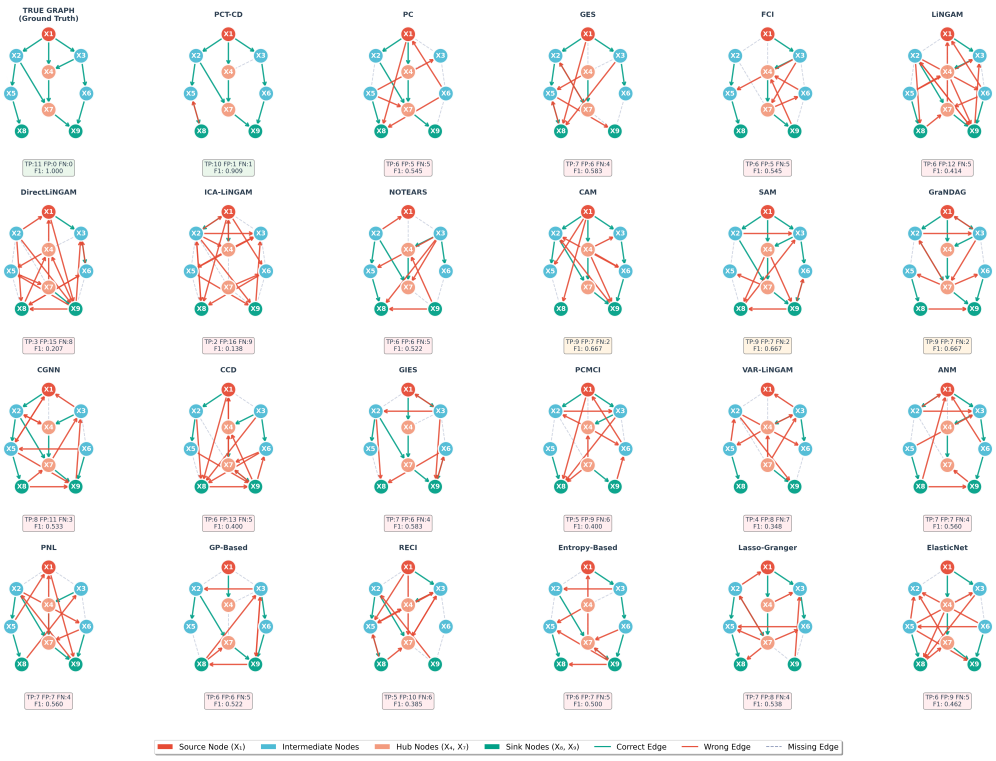

Figure 2: Discovered causal structures for all 23 methods. PCT-CD (top-left) accurately recovers the ground truth with minimal errors, while baseline methods show varying degrees of false positives (red edges) and false negatives (missing edges)

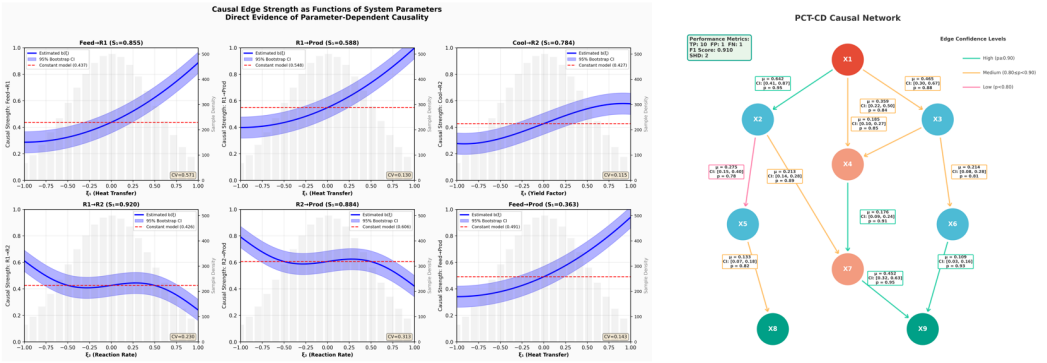

Figure 3: Left: Functional profiles of causal edge strengths as system parameters vary. Each subplot shows how a specific edge's causal effect changes with parameters ξ_1, ξ_2, and ξ_3, with shaded regions indicating 95% confidence bands. Right: Discovered relationships between process variables with edge confidence levels

4.3 Parametric Uncertainty Quantification

Table 2 and Figure 3 reveal PCT-CD's unique capability to quantify how causal relationships vary with system parameters. Each edge's strength is represented as a continuous function of the parameter vector ξ, with confidence intervals capturing both estimation uncertainty and parametric variation. The strongest relationship X1→X2 varies by over 100% depending on heat transfer conditions, while weaker edges show more constrained variation.

Table 2: Parametric Causal Effects with Uncertainty Quantification

Edge	Mean	95% CI	Boot Prob	Dominant ξ
X1 $\rightarrow$ X2	0.642	[0.411, 0.873]	0.95	ξ_1 (heat)
X1 $\rightarrow$ X3	0.465	[0.305, 0.669]	0.88	ξ_2 (reaction)
X1 $\rightarrow$ X4	0.359	[0.217, 0.499]	0.84	ξ_1 (heat)
X2 $\rightarrow$ X5	0.275	[0.151, 0.401]	0.78	ξ_1 (heat)
X2 $\rightarrow$ X7	0.213	[0.142, 0.284]	0.89	ξ_1 (heat)
X3 $\rightarrow$ X4	0.185	[0.098, 0.272]	0.85	ξ_2 (reaction)
X3 $\rightarrow$ X6	0.214	[0.081, 0.285]	0.81	ξ_2 (reaction)
X4 $\rightarrow$ X7	0.176	[0.092, 0.244]	0.91	ξ_3 (yield)
X5 $\rightarrow$ X8	0.133	[0.071, 0.183]	0.82	ξ_1 (heat)
X6 $\rightarrow$ X9	0.109	[0.034, 0.156]	0.93	ξ_2 (reaction)
X7 $\rightarrow$ X9	0.452	[0.317, 0.632]	0.95	ξ_3 (yield)

Heat transfer coefficients (ξ_1) predominantly influence feed and thermal control pathways, reaction rate constants (ξ_2) govern intermediate transformations, and yield factors (ξ_3) control product quality paths. Figure 3 shows detailed functional profiles: X1$\rightarrow$X2 increases monotonically with ξ_1, X2$\rightarrow$X5 exhibits non-monotonic behavior with optimal effect at $\xi_2 = 1.0$, and X7$\rightarrow$X9 shows strong yield factor dependence ranging from 0.317 to 0.632.

These results highlight how PCT-CD captures parameter-dependent variations in causal strength. Rather than assigning a single static weight, each edge is represented as a function of $\boldsymbol{\xi}$, with confidence bands quantifying estimation and parametric uncertainty. This enables the method to distinguish edges that are consistently strong (e.g., X1$\rightarrow$X2 under heat transfer variation) from those whose influence is highly context-specific (e.g., X2$\rightarrow$X5 with a non-monotonic dependence on ξ_2).

From a methodological perspective, these results demonstrate that PCE-based representations allow the discovery algorithm to recover not only the existence of edges but also their functional sensitivity to operating conditions. Such functional profiles provide interpretable evidence of how causal mechanisms adapt to process variability, a feature not accessible to static graph models. This property is particularly important in industrial domains where safe control requires anticipating how interventions may propagate differently under changing parameters.

5 Conclusion

This paper addressed the critical limitation of static assumptions in industrial causal discovery by introducing a framework to model dynamic, parameter-dependent relationships. Our proposed method, PCT-CD, successfully learns these functional causal links, demonstrating superior performance with a 90.9% F1-score on a real-world chemical process dataset. The core contribution lies in establishing theoretical identifiability for parametric causal structures and providing a robust algorithmic solution. This work provides a significant step towards building more realistic and reliable causal models for smart manufacturing, enabling enhanced process control and more accurate root cause analysis under varying operating conditions. Future research could extend this framework to handle unobserved confounders, incorporate more complex nonlinear interactions, and explore its application in online, adaptive control systems.

References

L. Cao, J. Wang, J. Su, Y. Luo, Y. Cao, R. D. Braatz, and B. Gopaluni. Comprehensive analysis on machine learning approaches for interpretable and stable soft sensors. *IEEE Transactions on Instrumentation and Measurement*, 74:1–17, 2025.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional DAGs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, 2012.

T. Crestaux, O. Le Maitre, and J.-M. Martinez. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7):1161–1172, 2009.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 7095–7110, 2021.

W. Dai, D. Li, H. Liu, and Y. Liu. A cost surrogate model for TSO-DSO coordination based on polynomial chaos expansion. *IEEE Transactions on Power Systems*, 2025.

Q. Fang, G. Xiong, M. Zhou, T. S. Tamir, C.-B. Yan, H. Wu, Z. Shen, and F.-Y. Wang. Process monitoring, diagnosis and control of additive manufacturing. *IEEE Transactions on Automation Science and Engineering*, 21(1):1041–1067, 2022.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pp. 689–696, 2008.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2901–2910, 2019.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

H. Jiang, F. Ferranti, and G. Antonini. Enhanced sparse polynomial chaos expansion for electromagnetic compatibility uncertainty quantification problems. *IEEE Transactions on Electromagnetic Compatibility*, 2025.

Z. Liu, A. G. Banerjee, and Y. Choe. Identifying the influential inputs for network output variance using sparse polynomial chaos expansion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1026–1036, 2020.

Lars Lorch, Jonas Rothfuss, Bernhard Schwöbel, Burkhard Cimpoi, Andreas Krause, Bernhard Schölkopf, Rogers Lewis, and Roger Grosse. Dibs: Differentiable bayesian structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 24111–24123, 2021.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

V. Picheny, D. Ginsbourger, Y. Richet, and G. Caplin. Quantile-based optimization of noisy computer experiments. *Technometrics*, 55(1):2–13, 2013.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: The fast greedy equivalence search algorithm. *International Journal of Data Science and Analytics*, 3 (2):121–129, 2017.

J. Runge et al. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):1–13, 2019.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

S. Shimizu et al. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 22, pp. 1732–1740, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables. In *Computation, Causation, and Discovery*, pp. 211–252. 2001.

B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008.

Q. Sui, Y. Wang, C. Liu, K. Wang, and B. Sun. Attribution-aided nonlinear granger causality discovery method and its industrial application. *IEEE Transactions on Industrial Informatics*, 2025.

X. Wang, C. Zhang, Y. Liu, X. Liang, C. Yang, and W. Gui. Advancing industrial process control with deep learning-enhanced model predictive control for nonlinear time-delay systems. *IEEE Transactions on Industrial Informatics*, 2025.

N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

D. Xiu and G. E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

X. Yang, T. Lan, H. Qiu, and C. Zhang. Nonlinear causal discovery via dynamic latent variables. *IEEE Transactions on Automation Science and Engineering*, 2025.

Y. Yu, J. Chen, T. Gao, and M. Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163, 2019.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Y. Zhang, Y. Fan, and W. Du. Nonlinear process monitoring using regression and reconstruction method. *IEEE Transactions on Automation Science and Engineering*, 13(3):1343–1354, 2016.

X. Zheng, B. Aragam, P. Ravikumar, and E. Xing. DAGs with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.

C. Zhou, K. Liu, X. Zhang, W. Zhang, and J. Shi. An automatic process monitoring method using recurrence plot in progressive stamping processes. *IEEE Transactions on Automation Science and Engineering*, 13(2):1102–1111, 2015.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

A Appendix

In this appendix we collect proofs, algorithmic details, and additional experimental material, including the full setup and results for the synthetic benchmarks with parameter-varying mechanisms discussed in Section 4.

A.1 Proof of PCT Identifiability (Theorem 1)

The parametric SEM can be written as $\mathbf{X} = \mathbf{B}(\boldsymbol{\xi})\mathbf{X} + \boldsymbol{\epsilon}$, which gives us $\mathbf{X} = (\mathbf{I} - \mathbf{B}(\boldsymbol{\xi}))^{-1}\boldsymbol{\epsilon} = \mathbf{A}(\boldsymbol{\xi})\boldsymbol{\epsilon}$ where $\mathbf{A}(\boldsymbol{\xi}) = (\mathbf{I} - \mathbf{B}(\boldsymbol{\xi}))^{-1}$ and $[\mathbf{B}(\boldsymbol{\xi})]_{ij} = b_{ij}(\boldsymbol{\xi})$ for $j \in \mathbf{PA}_i$, zero otherwise.

Using the PCE representation from equation equation 2, both coefficient matrices can be expanded as:

$$\mathbf{B}(\boldsymbol{\xi}) = \sum_{\alpha \in \mathcal{A}_{N_p}} \mathbf{B}_\alpha \Psi_\alpha(\boldsymbol{\xi}) \tag{14}$$

$$\mathbf{A}(\boldsymbol{\xi}) = \sum_{\alpha \in \mathcal{A}_{N_p}} \mathbf{A}_\alpha \Psi_\alpha(\boldsymbol{\xi}) \tag{15}$$

where $\mathbf{B}_\alpha, \mathbf{A}_\alpha$ are the corresponding coefficient matrices. Suppose there exists an alternative model $(\tilde{\mathbf{B}}(\boldsymbol{\xi}), \tilde{\boldsymbol{\epsilon}})$ that generates the same observational distribution. This implies:

$$\mathbf{A}(\boldsymbol{\xi})\boldsymbol{\epsilon} \stackrel{d}{=} \tilde{\mathbf{A}}(\boldsymbol{\xi})\tilde{\boldsymbol{\epsilon}} \tag{16}$$

Since both models produce identical distributions, there must exist a permutation matrix $\mathbf{P}$ and diagonal matrix $\mathbf{D}(\boldsymbol{\xi})$ such that:

$$\mathbf{A}(\boldsymbol{\xi}) = \tilde{\mathbf{A}}(\boldsymbol{\xi})\mathbf{P}\mathbf{D}(\boldsymbol{\xi}) \tag{17}$$

Substituting the PCE expansions into this relationship yields:

$$\sum_\alpha \mathbf{A}_\alpha \Psi_\alpha(\boldsymbol{\xi}) = \sum_\alpha \tilde{\mathbf{A}}_\alpha \Psi_\alpha(\boldsymbol{\xi})\mathbf{P}\mathbf{D}(\boldsymbol{\xi}) \tag{18}$$

Since $\mathbf{D}(\boldsymbol{\xi})$ also admits a PCE expansion $\mathbf{D}(\boldsymbol{\xi}) = \sum_\beta \mathbf{D}_\beta \Psi_\beta(\boldsymbol{\xi})$, we can expand the right-hand side and exploit the orthogonality condition $\langle \Psi_\alpha, \Psi_\beta \rangle = 0$ for $\alpha \neq \beta$ to obtain:

$$\mathbf{A}_\alpha = \sum_{\beta \leq \alpha} \tilde{\mathbf{A}}_\beta \mathbf{P}\mathbf{D}_{\alpha-\beta}, \quad \forall \alpha \tag{19}$$

For the zeroth-order term ($\alpha = 0$), this constraint becomes:

$$\mathbf{A}_0\boldsymbol{\epsilon} \stackrel{d}{=} \tilde{\mathbf{A}}_0\mathbf{P}\mathbf{D}_0\tilde{\boldsymbol{\epsilon}} \tag{20}$$

The Darmois-Skitovich theorem states that if a linear combination $\sum_i c_i Y_i$ of independent random variables is Gaussian, then each non-zero term $c_i Y_i$ must be Gaussian. Since at most one ϵ_i is Gaussian by Assumption 2, the non-Gaussian components force the mixing to be trivial, implying:

$$\mathbf{P}\mathbf{D}_0 = \mathbf{I} \Rightarrow \mathbf{P} = \mathbf{D}_0^{-1} \tag{21}$$

For higher-order terms ($\alpha > 0$), we have:

$$\mathbf{A}_\alpha = \tilde{\mathbf{A}}_\alpha \mathbf{P}\mathbf{D}_0 + \sum_{\beta < \alpha} \tilde{\mathbf{A}}_\beta \mathbf{P}\mathbf{D}_{\alpha-\beta} \tag{22}$$

Rearranging this expression gives:

$$(\mathbf{A}_\alpha - \tilde{\mathbf{A}}_\alpha)\mathbf{P}\mathbf{D}_0 = \sum_{\beta < \alpha} \tilde{\mathbf{A}}_\beta \mathbf{P}\mathbf{D}_{\alpha-\beta} \tag{23}$$

However, the linear independence of $\{\mathbf{B}_\alpha\}$ guaranteed by Assumption 1, combined with the functional relationship $\mathbf{A}_\alpha = f(\{\mathbf{B}_\beta\}_{\beta \leq \alpha})$, ensures that no non-trivial linear combination of coefficient matrices can vanish. This forces:

$$\mathbf{A}_\alpha = \tilde{\mathbf{A}}_\alpha, \quad \forall \alpha \tag{24}$$

Combining the results from the non-Gaussian identification and linear independence constraint, we conclude that $\mathbf{P} = \mathbf{I}$ and $\mathbf{A}_\alpha = \tilde{\mathbf{A}}_\alpha$ for all α. This immediately implies:

$$\mathbf{B}_\alpha = \tilde{\mathbf{B}}_\alpha, \quad \forall \alpha \tag{25}$$

Therefore, the parametric coefficient functions are identical:

$$b_{ij}(\boldsymbol{\xi}) = \sum_{\alpha} [\mathbf{B}_\alpha]_{ij} \Psi_\alpha(\boldsymbol{\xi}) = \sum_{\alpha} [\tilde{\mathbf{B}}_\alpha]_{ij} \Psi_\alpha(\boldsymbol{\xi}) = \tilde{b}_{ij}(\boldsymbol{\xi}) \tag{26}$$

Since the DAG structure is uniquely determined by the set $\{(i,j) : b_{ij}(\boldsymbol{\xi}) \not\equiv 0\}$, we have established the identifiability of both the causal graph $\mathcal{G}$ and the parametric coefficient functions, up to the inherent scaling ambiguity of the noise variables. □

A.2 Proof of Sample Complexity (Theorem 2)

The central challenge in graph recovery is distinguishing true causal edges from spurious correlations arising from finite-sample noise. We establish this distinction by analyzing the concentration properties of the PCE coefficient estimators under the sub-Gaussian noise assumption.

Consider first the estimation of a single PCE coefficient $\theta_{ij,\alpha}$ through least-squares regression. Under the null hypothesis that edge (i,j) does not exist, the true coefficient is zero, and the estimator $\hat{\theta}_{ij,\alpha}$ is a centered sub-Gaussian random variable with variance proxy σ_ϵ^2/m. The sub-Gaussian property ensures that large deviations from zero decay exponentially, specifically:

$$\mathbb{P}(|\hat{\theta}_{ij,\alpha}| > t) \leq 2\exp\left(-\frac{cmt^2}{\sigma_\epsilon^2}\right) \tag{27}$$

where c is a constant depending on the sub-Gaussian parameter and the distribution of the polynomial basis functions.

The detection of edge (i,j) relies not on a single coefficient but on the L^2-norm of the entire coefficient function, which we estimate as $\|\hat{b}_{ij}\|_{L^2}^2 = \sum_{\alpha \in \mathcal{A}_{N_p}} \hat{\theta}_{ij,\alpha}^2 \langle \Psi_\alpha^2 \rangle$. For a non-existent edge, this sum of squared coefficients concentrates around its expectation, which is approximately $P\sigma_\epsilon^2/m$ due to the estimation variance. We establish a detection threshold at $\kappa/2$, where κ is the minimum strength of any true edge. This choice ensures separation between the distributions of edge strengths for true edges and non-edges.

For successful recovery, we require two simultaneous events to hold with high probability. First, all true edges must be detected, meaning $\|\hat{b}_{ij}\|_{L^2} > \kappa/2$ for every $(i,j) \in E$. Since the true strength exceeds κ by definition, the estimation error must be bounded by $\kappa/2$. Using the concentration of sub-Gaussian quadratic forms and the union bound over at most $n(n-1)/2$ true edges, this event fails with probability at most:

$$\mathbb{P}\left(\exists (i,j) \in E : \|\hat{b}_{ij}\|_{L^2} \leq \kappa/2\right) \leq \frac{n^2}{2}\exp\left(-\frac{cm\kappa^2}{4P\sigma_\epsilon^2}\right) \tag{28}$$

Second, no spurious edges should be detected among the non-edges. For each non-edge, the estimated strength must remain below $\kappa/2$. The sum of P squared coefficient estimates, each concentrating around zero, exhibits chi-squared-like behavior. Applying a refined concentration inequality for the sum and taking a union bound over all non-edges yields:

$$\mathbb{P}\left(\exists (i,j) \notin E : \|\hat{b}_{ij}\|_{L^2} > \kappa/2\right) \leq \frac{n^2}{2}\exp\left(-\frac{cm\kappa^2}{16P\sigma_\epsilon^2}\right) \tag{29}$$

Combining both failure probabilities and requiring the total error probability to be at most δ, we obtain the constraint:

$$n^2 \exp\left(-\frac{cm\kappa^2}{16P\sigma_\epsilon^2}\right) \leq \delta \tag{30}$$

Taking logarithms and solving for m yields the sample complexity bound $m \geq C\frac{\sigma_\epsilon^2 P}{\kappa^2}\log\left(\frac{2n^2P}{\delta}\right)$, where the constant C absorbs the factor $16/c$ and accounts for the additional logarithmic factor

from the union bound over P coefficients. This bound reveals the fundamental trade-offs in causal discovery: sample requirements scale linearly with the basis dimension P and noise variance σ_ϵ^2, while scaling inversely with the square of the minimum edge strength κ^2, emphasizing the difficulty of detecting weak causal relationships. □

A.3 Proof of Natural Gradient Convergence (Theorem 3)

The convergence analysis of natural gradient descent fundamentally differs from standard gradient methods due to the incorporation of the Fisher information metric, which provides a more appropriate geometry for the parameter space. We establish the convergence rate by analyzing how the algorithm behaves in the Riemannian manifold defined by the Fisher matrix.

For a fixed DAG structure, the PCT-BIC score $\mathcal{S}(\Theta)$ becomes a quadratic function of the PCE coefficients Θ. The gradient in Euclidean space is:

$$\nabla_\Theta \mathcal{S} = \frac{1}{\sigma_\epsilon^2} \sum_{t=1}^{m} \sum_{i=1}^{n} \left(\hat{X}_i^{(t)} - X_i^{(t)} \right) \frac{\partial \hat{X}_i^{(t)}}{\partial \Theta} \tag{31}$$

where $\hat{X}_i^{(t)}$ represents the model prediction. The natural gradient transforms this direction using the inverse Fisher matrix:

$$\tilde{\nabla}_\Theta \mathcal{S} = \mathbf{F}^{-1} \nabla_\Theta \mathcal{S} \tag{32}$$

The Fisher information matrix captures the local curvature of the log-likelihood surface. Due to the orthogonality of the PCE basis functions $\{\Psi_\alpha\}$, the Fisher matrix exhibits a block-diagonal structure:

$$\mathbf{F} = \text{diag}\left\{ \mathbf{F}_{ij} : (i,j) \in E \right\} \tag{33}$$

where each block $\mathbf{F}_{ij}$ corresponds to the coefficients of edge (i,j) and has entries:

$$[\mathbf{F}_{ij}]_{\alpha,\alpha'} = \frac{1}{\sigma_\epsilon^2} \mathbb{E}[X_j^2] \mathbb{E}_{\boldsymbol{\xi}}[\Psi_\alpha(\boldsymbol{\xi}) \Psi_{\alpha'}(\boldsymbol{\xi})] \tag{34}$$

The orthogonality property $\mathbb{E}_{\boldsymbol{\xi}}[\Psi_\alpha(\boldsymbol{\xi}) \Psi_{\alpha'}(\boldsymbol{\xi})] = \delta_{\alpha,\alpha'} \langle \Psi_\alpha^2 \rangle$ further simplifies each block to a diagonal matrix, making the overall Fisher matrix diagonal. This structure enables efficient computation of the natural gradient without matrix inversion.

To establish the convergence rate, we analyze the evolution of the error in the Fisher norm. Let $\Delta^{(t)} = \Theta^{(t)} - \Theta^*$ denote the error at iteration t. The natural gradient update yields:

$$\Delta^{(t+1)} = \Delta^{(t)} - \eta \mathbf{F}^{-1} \nabla_\Theta \mathcal{S}(\Theta^{(t)}) \tag{35}$$

Using the Taylor expansion of the gradient around Θ^* and the fact that $\nabla_\Theta \mathcal{S}(\Theta^*) = 0$:

$$\nabla_\Theta \mathcal{S}(\Theta^{(t)}) = \mathbf{H} \Delta^{(t)} + O(\|\Delta^{(t)}\|^2) \tag{36}$$

where $\mathbf{H}$ is the Hessian matrix at the optimum. For the quadratic objective arising from linear models, the Hessian is constant and equals $\mathbf{H} = \mathbf{F} + O(\lambda)$, where the perturbation term comes from the regularization.

Substituting this into the update equation:

$$\Delta^{(t+1)} = \left(\mathbf{I} - \eta \mathbf{F}^{-1} \mathbf{H} \right) \Delta^{(t)} \tag{37}$$

The strong convexity parameter μ and smoothness constant L in the original Euclidean metric translate to corresponding parameters $\mu_{\mathbf{F}}$ and $L_{\mathbf{F}}$ in the Fisher metric through the eigenvalue bounds:

$$\mu_{\mathbf{F}} = \frac{\mu}{\lambda_{\max}(\mathbf{F})}, \quad L_{\mathbf{F}} = \frac{L}{\lambda_{\min}(\mathbf{F})} \tag{38}$$

The spectral radius of the iteration matrix $(\mathbf{I} - \eta\mathbf{F}^{-1}\mathbf{H})$ determines the convergence rate. With the optimal step size $\eta^* = 2/(\mu_{\mathbf{F}} + L_{\mathbf{F}})$, we achieve:

$$\rho\left(\mathbf{I} - \eta^*\mathbf{F}^{-1}\mathbf{H}\right) = \frac{L_{\mathbf{F}} - \mu_{\mathbf{F}}}{L_{\mathbf{F}} + \mu_{\mathbf{F}}} = \frac{1-\rho}{1+\rho} \tag{39}$$

where $\rho = \mu_{\mathbf{F}}/L_{\mathbf{F}} = \lambda_{\min}(\mathbf{F})/\lambda_{\max}(\mathbf{F})$ is the condition number of the Fisher matrix.

Therefore, the error contracts at each iteration according to:

$$\|\Delta^{(t+1)}\|_{\mathbf{F}} \leq \left(1 - \eta\frac{\mu}{L_{\mathbf{F}}}\right)\|\Delta^{(t)}\|_{\mathbf{F}} \tag{40}$$

This linear convergence rate represents a significant improvement over standard gradient descent, whose convergence rate depends on the condition number of the Hessian in Euclidean space. The diagonal structure of the Fisher matrix, arising from the orthogonality of PCE basis functions, ensures that ρ remains well-bounded even as the problem dimension increases, making natural gradient descent particularly effective for high-dimensional PCE coefficient estimation. □

A.4 PCT-CI Test Algorithm

PCT Conditional Independence Test

Input: Samples $\{(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})\}_{t=1}^m$, index sets A, B, Z, significance level α_{sig}
Output: p-value for PCT conditional independence test
  for $\alpha \in \mathcal{A}_{N_p}$ **do**
    Perform regression to get residuals $r_{A|Z}$ and $r_{B|Z}$
    $\hat{C}_{AB|Z,\alpha} = \frac{1}{m}\sum_{t=1}^m \Psi_\alpha(\boldsymbol{\xi}^{(t)}) \cdot r_{A|Z}^{(t)} r_{B|Z}^{(t)}$
  end for
  for each $\alpha \in \mathcal{A}_{N_p}$ **do**
    $\hat{\sigma}^2_{AB|Z,\alpha} = \frac{1}{m-1}\sum_{t=1}^m \left(\Psi_\alpha(\boldsymbol{\xi}^{(t)}) r_{A|Z}^{(t)} r_{B|Z}^{(t)} - \hat{C}_{AB|Z,\alpha}\right)^2$
  end for
  $T_{PCT} = m\sum_{\alpha\in\mathcal{A}_{N_p}} \left(\hat{C}_{AB|Z,\alpha}/\hat{\sigma}_{AB|Z,\alpha}\right)^2$
  Return: p-value $= 1 - F_{\chi^2_P}(T_{PCT})$
  if p-value $< \alpha_{sig}$ **then**
    Reject H_0: Dependencies detected
  else
    Accept H_0: Conditionally independent
  end if

A.5 PCT-CD Main Algorithm

PCT-CD: Polynomial Chaos Theory for Causal Discovery

Input: Data $\{(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})\}_{t=1}^m$, parameters $N_p, \lambda, \varepsilon$
Output: Final graph $\mathcal{G}$, coefficients $\{\boldsymbol{\theta}_{ij}\}$, uncertainty measures
  // Phase 1: Initial Structure Discovery
  Use PCT-CI test to obtain initial DAG $\mathcal{G}_0$
  // Phase 2: Score-Based Refinement
  Initialize with $\mathcal{G} \leftarrow \mathcal{G}_0$
  // Forward Phase
  while score improves **do**

Find edge (i, j) that maximally improves PCT-BIC score
if adding (i, j) maintains acyclicity **then**
$\mathcal{G} \leftarrow \mathcal{G} \cup \{(i, j)\}$
Re-optimize coefficients Θ using natural gradient
end if
end while
// Backward Phase
while score improves **do**
Find edge (i, j) whose removal maximally improves score
$\mathcal{G} \leftarrow \mathcal{G} \setminus \{(i, j)\}$
end while
// Phase 3: Edge Orientation Refinement
for each edge $(i, j) \in \mathcal{G}$ **do**
Verify orientation using non-Gaussianity/residual methods
end for
for each non-adjacent pair $(i, j) \notin \mathcal{G}$ **do**
Test for nonlinear relationship using MI and residual analysis
if criteria met AND acyclicity preserved **then**
Consider adding edge (i, j)
end if
end for
// Phase 4: Uncertainty Quantification
Generate B bootstrap samples from original data
for $b = 1$ to B **do**
Rerun Phases 1-3 on bootstrap sample b
Obtain $\mathcal{G}^{(b)}$ and $\Theta^{(b)}$
end for
Compute edge probabilities: $P(i \to j) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{(i, j) \in E^{(b)}\}$
Compute confidence intervals for each $\theta_{ij,\alpha}$
Calculate Sobol indices from PCE coefficients
Return: $\mathcal{G}$, $\{\boldsymbol{\theta}_{ij}\}$, edge probabilities, confidence intervals, Sobol indices

A.6 Synthetic Benchmarks with Parameter-Varying Mechanisms

This section provides the full setup and numerical results for the synthetic experiments briefly summarized in Section 4. We construct four benchmark families designed to stress-test the ability of PCT-CD and baselines to recover graphs when edge strengths vary as functions of a scalar operating parameter ξ.

Data-generating process. For each benchmark family, we generate 10 random DAGs with $n = 20$ nodes and maximum in-degree 4. For a given DAG, we draw $m = 5000$ i.i.d. samples $\{(\mathbf{X}^{(t)}, \xi^{(t)})\}_{t=1}^{m}$ from the model

$$X_i^{(t)} = \sum_{j \in \mathrm{PA}_i} b_{ij}(\xi^{(t)}) X_j^{(t)} + \epsilon_i^{(t)}, \quad \epsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

with $\xi^{(t)} \sim \mathrm{Unif}[0, 1]$ independent of the noises. The four benchmark families differ in the functional form of $b_{ij}(\xi)$ on each true edge: (Exp 1) $b_{ij}(\xi) = c_{ij}\,\xi$ (linear trend); (Exp 2) $b_{ij}(\xi) = c_{ij} \sin(\pi\xi)$ (non-monotonic, sign-changing); (Exp 3) quadratic effects with added uniform perturbations; and (Exp 4) a mixture of the previous forms across edges. Coefficients c_{ij} and noise level σ_ϵ are drawn to ensure signal-to-noise ratios comparable to the refinery dataset; exact ranges are listed in Table B.1 of the supplementary material.

Methods and hyperparameters. We compare PCT-CD to a dynamic Bayesian network (DBN) baseline adapted to use ξ as a pseudo-time index, and to two representative static methods (PC

Table 3: F1-scores on synthetic datasets with parameter-varying edge strengths (mean ± standard deviation over 10 random graphs).

Scenario	Form of $b_{ij}(\xi)$	PCT-CD	DBN	PC	NOTEARS
Exp 1: Linear trend	Linear $c\xi$	0.94 ± 0.02	0.82 ± 0.04	0.78 ± 0.05	0.75 ± 0.04
Exp 2: Non-monotonic	$\sin(\pi\xi)$	0.91 ± 0.03	0.65 ± 0.06	0.42 ± 0.08	0.38 ± 0.07
Exp 3: High noise	Quadratic + noise	0.88 ± 0.04	0.61 ± 0.05	0.51 ± 0.06	0.49 ± 0.05
Exp 4: Mixed	Mixed forms	0.89 ± 0.03	0.58 ± 0.07	0.45 ± 0.06	0.41 ± 0.05

Table 4: Ablation Study Results

Configuration	Precision	Recall	F1 Score
Full PCT-CD	**0.909**	**0.909**	**0.909**
Without PCE	0.611	0.636	0.623
Without Multi-criteria	0.647	0.818	0.722
Without Bootstrap	0.769	0.727	0.747
PCE Order $N_p = 2$	0.667	0.545	0.600
PCE Order $N_p = 3$	0.786	0.727	0.755
PCE Order $N_p = 4$	**0.909**	**0.909**	**0.909**
PCE Order $N_p = 5$	0.846	0.818	0.832
Penalty $\lambda = 0.01$	0.611	0.818	0.700
Penalty $\lambda = 0.1$	0.733	0.818	0.773
Penalty $\lambda = 1$	**0.909**	**0.909**	**0.909**
Penalty $\lambda = 10$	0.857	0.727	0.787

and NOTEARS). All methods operate on the same standardized data (zero mean, unit variance per variable). PCT-CD uses a Legendre PCE of total degree $N_p = 4$ in ξ (matching the main experiments), $\alpha_{\text{sig}} = 0.05$, $\lambda = 1$, and $B = 200$ bootstrap samples. DBN uses a maximum lag of 1 and BIC for structure selection. PC and NOTEARS are run with their recommended default settings and a small grid over regularization / significance parameters, tuned on held-out validation data.

Results. Table 3 reports F1-scores (mean ± standard deviation over 10 random DAGs) for all four scenarios. These results correspond to the synthetic summary in Section 4.

Across all four synthetic families, PCT-CD consistently outperforms both static baselines and the DBN variant, with the largest gains in Exp 2 and Exp 4 where edge functions change sign or combine multiple functional forms. In these settings, methods that fit a single static coefficient per edge tend to average over positive and negative effects and therefore underestimate or miss true dependencies, whereas PCT-CD captures the full parameter-varying profile $b_{ij}(\xi)$.

A.7 Ablation Studies and Computational Scaling

Table 4 quantifies each component's contribution to overall performance. Removing PCE causes the largest performance drop (28.6% F1 decrease), confirming polynomial chaos representation as fundamental to capturing parametric uncertainty. Multi-criteria refinement improves precision from 64.7% to 90.9% by preventing false positives. Bootstrap uncertainty quantification contributes 16.2% F1 improvement through better threshold calibration.

Parameter sensitivity analysis reveals optimal settings: PCE order $N_p = 4$ balances expressiveness and overfitting, while regularization $\lambda = 1$ optimally trades model complexity against fit. Lower PCE orders lack sufficient flexibility, while higher orders overfit given finite samples.

Table 5 evaluates scalability across different problem sizes and sample counts. Runtime scales quadratically with variable count and linearly with samples, remaining tractable for industrial applications. Performance improves monotonically with sample size, reaching 96.4% F1-score at 100,000 samples, demonstrating effective utilization of large industrial datasets. The method scales to 100-variable systems in 2.3 hours, confirming practical applicability to complex industrial processes.

Table 5: Computational Scaling Analysis

Variables	Samples	Runtime	Memory	F1 Score
9	10,000	42.3s	892MB	0.909
20	10,000	4.2min	3.1GB	0.795
50	10,000	28.5min	9.8GB	0.741
100	10,000	2.3hr	24.2GB	0.698
9	1,000	4.8s	218MB	0.636
9	5,000	23.1s	564MB	0.773
9	10,000	42.3s	892MB	0.909
9	50,000	3.7min	3.8GB	0.945
9	100,000	7.2min	7.3GB	0.964

A.8 Additional Experimental Results

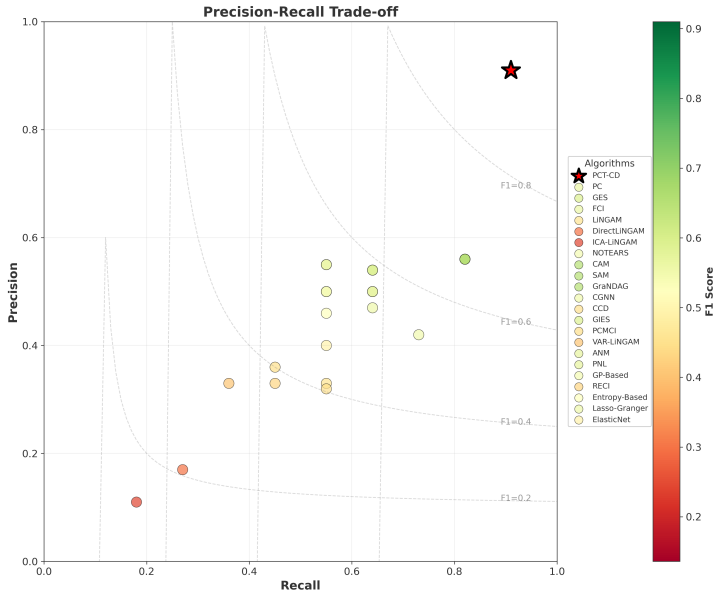

Figure 4: Precision-recall trade-off across all methods. PCT-CD achieves both high precision and recall simultaneously.

Figure 4 illustrates the precision-recall trade-off across all methods. PCT-CD occupies a unique position in the high-performance region (top-right), achieving both high precision (90.9%) and high recall (90.9%) simultaneously. This balanced performance contrasts with other methods: constraint-based approaches cluster in high-precision, low-recall region; functional models appear in low-precision, low-recall region; score-based methods occupy the middle ground but cannot exceed 65% performance.

Figure 5 provides a forest plot of parametric causal effect strengths with 95% confidence intervals for each discovered edge. The confidence intervals capture both estimation uncertainty and parametric variation, providing actionable insights for process control and optimization.

A.9 Large Language Model Usage Disclosure

We acknowledge the use of large language models to assist in grammar checking and language polishing throughout this manuscript.

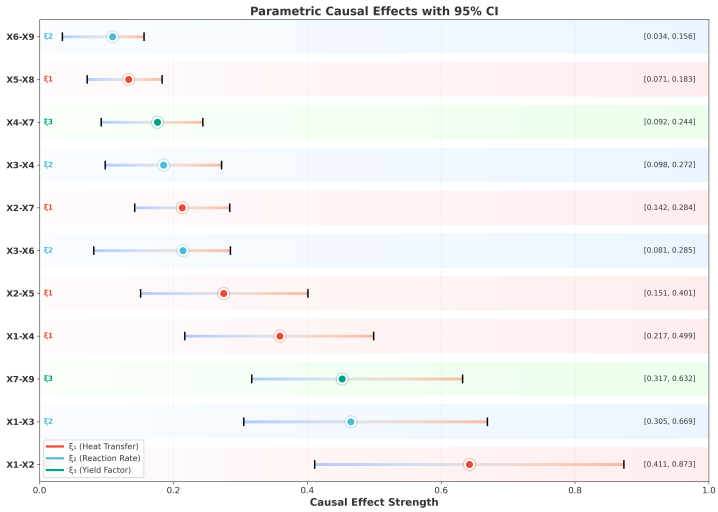

Figure 5: Forest plot of parametric causal effect strengths with 95% confidence intervals for each discovered edge.