

A Survey on Causal Discovery and Estimation: Theory, Methods, and Applications

Pavan Nithin Mullapudi

Pavannithin123@gmail.com

Senior Applied Scientist

Amazon, Seattle, WA

Abstract

The ability to understand and leverage causality separates true intelligence from mere pattern recognition. This survey examines recent advances in causal discovery and estimation techniques, their theoretical foundations, computational implementations, and cross-domain applications. Causal discovery has evolved from a theoretical construct into a practical toolkit for reliable decision-making across diverse domains. We present a comprehensive overview of modern approaches to causal structure identification and effect estimation, emphasizing both methodological innovations and real-world applications in finance, supply chain management, and healthcare.

Keywords: Machine Learning, Cloud Computing, Semi-Supervised Learning, Revenue Prediction

1. Foundations of Causal Discovery

Understanding the laws that govern phenomena is the core of scientific progress, particularly when modeling the interplay between different aspects in a causal fashion. Unlike traditional statistical analysis that focuses on correlations, causal discovery aims to recover the underlying causal graph from observational data, enabling the identification and estimation of causal effects.

1.1 Structural Causal Models and DAGs

Two widely accepted frameworks exist for causal inference: the structural causal model (SCM) and the potential outcome framework (POF). An SCM consists of two core components: the causal graph and the structural equations. The causal graph is represented as a directed acyclic graph (DAG) where edges signify causal relationships between variables (treatments, outcomes, and covariates). These relationships are quantitatively specified through structural equations.

A causal graph is considered a perfect map for a probability distribution if every conditional independence statement derived from the graph can also be derived from the distribution and vice versa. This concept is central to constraint-based causal discovery algorithms that translate conditional probability independence into graphical separation through statistical tests^{[1][2]}.

1.2 From Association to Causation

Pearl's "ladder of causation" categorizes queries into three levels of increasing complexity:

1. Association (observational statistics)
2. Intervention (experimental manipulation)
3. Counterfactual reasoning

Traditional predictive models operate primarily at the first level, while causal models enable reasoning at all three. This distinction is critical in applications where understanding "why" something happens is as important as predicting "what" will happen^{[1][3]}.

2. Causal Discovery Algorithms

Causal discovery algorithms attempt to recover the underlying causal structure from data. These methods can be broadly categorized into several families based on their underlying assumptions and approaches.

2.1 Algorithmic Landscape

Constraint-Based Algorithms

These methods recover the causal graph by exploiting conditional independence statements obtained through statistical tests. They assume faithfulness of the underlying distribution and translate conditional probability independence into graphical separation. Common algorithms include PC and FCI^{[1][4]}.

Score-Based Approaches

Score-based methods formulate causal discovery as an optimization problem, searching for the graph structure that maximizes a scoring function such as BIC or BGe. Greedy Equivalence Search (GES) is a prominent example^{[2][5]}.

Noise-Based Approaches

These techniques leverage assumptions about the noise distribution in the data-generating process. Notable examples include LiNGAM, which assumes non-Gaussian noise, and quantile causal discovery (bQCD), which uses quantile scoring^{[2][6]}.

Granger Causality

Primarily used for time series data, Granger causality determines whether one series helps predict another beyond what can be predicted using only the history of the second series^{[3][7]}.

Differentiable Methods

Recent innovations recast discrete DAG search as continuous optimization, notably NOTEARS, which enforces acyclicity through a smooth trace-exponential constraint on the adjacency matrix^{[4][8]}.

Table 1: Comparison of Causal Discovery Algorithm Families

Algorithm Family	Key Assumptions	Strengths	Weaknesses	Example Algorithms
Constraint-based	Faithfulness, causal sufficiency	Works with mixed data types	Sensitive to errors in independence tests	PC, FCI
Score-based	Correct model class, scoring metric	Statistically consistent	Computationally expensive for large networks	GES, FGES
Noise-based	Non-Gaussian noise, additivity	Can identify causal direction	Strict distributional assumptions	LiNGAM, bQCD
Granger causality	Temporal precedence, stationarity	Natural for time series	Cannot detect instantaneous effects	VAR-based methods
Differentiable	Continuous representation of DAG	Scalable to many variables	May converge to local minima	NOTEARS, DAG-GNN

2.2 Time Series Causal Discovery

Time series data present unique challenges and opportunities for causal discovery. Standard methods must be adapted to account for temporal dependencies, lagged relationships, and non-stationarity.

The constraint-based Causal Discovery from Heterogeneous Data algorithm has been extended to account for lagged relationships in time-series data, allowing for the identification of complex causal relations between financial assets and variables. This approach is particularly valuable in domains like finance where underlying mechanisms are dynamic and non-stationary^{[3][7]}.

For time series, causal discovery algorithms can be classified as Granger causality methods, constraint-based approaches, noise-based approaches, score-based approaches, logic-based approaches, topology-based approaches, and difference-based approaches. Each family offers distinct advantages depending on the characteristics of the data and the assumptions one is willing to make^{[3][7]}.

3. Estimation and Inference

Once a causal structure is identified, the next step is to estimate the magnitude of causal effects.

3.1 Graph-Aware Effect Estimation

Graph-aware estimation first identifies an admissible covariate set that d-separates all back-door paths from treatment to outcome. Standard regression methods (OLS, random forests) can provide consistent estimates when sample sizes are large^{[1][5]}.

For more robust estimation, doubly robust learners combine outcome modeling with propensity score approaches to achieve consistency when either model is correctly specified. This approach is particularly valuable in healthcare settings where observational data often suffer from selection bias^{[1][6]}.

3.2 Counterfactual Queries

Counterfactual inference follows a three-step procedure:

1. Abduction: Inferring latent noise terms consistent with factual observations
2. Action: Replacing structural equations for intervened variables
3. Prediction: Simulating the new outcome

This framework enables answering "what if" questions about past events, such as "How would the patient have responded if we had administered a different treatment?"^{[1][7]}.

3.3 Uncertainty and Sensitivity

Causal estimates are subject to both sampling variance and bias from violations of assumptions. Bootstrap methods and Bayesian approaches can quantify variance, while sensitivity analyses assess how robust findings are to unmeasured confounding^{[1][5]}.

Researchers are increasingly developing formal methods to evaluate how strong an unmeasured confounder would need to be to invalidate a causal conclusion, providing greater confidence in real-world applications of causal models^{[1][6]}.

4. Validation and Refutation

Robust causal analysis demands systematic validation. Several approaches can help assess the validity of causal models:

1. Placebo treatments: Applying the estimation method to a treatment known to have no effect
2. Data subset refuters: Checking consistency of results across different subpopulations
3. Random common-cause simulation: Adding synthetic confounders to test robustness
4. Adversarial edge removal: Systematically removing edges to identify critical dependencies

These methods go beyond traditional statistical validation to specifically test causal assumptions and help identify potential flaws in the model specification^{[1][5]}.

5. Tooling Ecosystem

Open-source software has significantly lowered the barrier to entry for causal modeling. Several libraries provide comprehensive frameworks for causal discovery and estimation.

5.1 DoWhy and DoWhy-GCM

DoWhy offers a four-step API (model, identify, estimate, refute) that forces explicit causal assumptions and encourages sensitivity checks. DoWhy-GCM extends this with structural equation fitting, enabling high-fidelity counterfactual simulation^{[2][8]}.

5.2 EconML, CausalML, and PyWhy

EconML implements state-of-the-art estimation methods like metalearners, doubly robust learners, and instrumental variable approaches. CausalML focuses on uplift modeling, while PyWhy provides a unified ecosystem for causal tools and benchmarks^{[2][8]}.

5.3 Benchmark Datasets

Benchmark datasets are essential for evaluating causal discovery and estimation methods. Common benchmarks include semi-synthetic datasets like IHDP and ACIC, along with real-world datasets in specific domains^{[2][8]}.

6. Applications Across Domains

6.1 Finance

Causal discovery has become increasingly important in financial markets for understanding the complex relationships between assets and variables. By recovering causal structures from time-series data, analysts can identify directional influences between market factors and optimize investment strategies^{[3][7]}.

The application of causal discovery algorithms to factor investing addresses recent criticisms of correlation-based approaches. By understanding the causal mechanisms driving asset returns, investors can build more robust portfolios that remain stable under changing market conditions^{[3][7]}.

Causal AI identifies the impact of interest rates, inflation, and FX rates on customer creditworthiness, determines the causal drivers of credit risk, and suggests the best interventions for defaulted or at-risk loans. This approach enables precise, data-driven decisions for effective risk management^{[3][7]}.

6.2 Supply-Chain & Logistics

In supply chain management, causality modeling offers significant advantages over traditional correlation-based approaches. As organizations face increasingly complex data environments, understanding the "why" behind events becomes crucial for building resilient systems^{[4][8]}.

A novel approach integrating causal discovery with reinforcement learning has demonstrated effectiveness in root cause attribution of delivery risks. This method accurately identifies key drivers of late deliveries, such as shipping mode and delivery status, providing actionable insights for optimizing supply chain performance^{[4][8]}.

Bayesian Network modeling represents a powerful framework for causality modeling in supply chains. The process involves building a causal graph, estimating causal inference, quantifying impact, and

simulating scenarios for decision support. Applications include supplier risk management, manufacturing excellence, logistics optimization, and inventory optimization^{[4][8]}.

6.3 Healthcare & Bioinformatics

Causal inference in healthcare seeks to build and apply models that learn the effect of interventions on outcomes using diverse data types. This approach is especially valuable for drug safety surveillance where confounding bias may mask adverse reactions^{[5][6]}.

The structural causal model framework provides a comprehensive theory of causality that can handle cases where confounders cannot be measured directly. For example, researchers have used blood type as an instrumental variable to estimate the average survival benefit from liver transplantation^{[5][6]}.

Recent advancements in healthcare delivery have integrated IoT, AI, and blockchain technologies with causal modeling to create more personalized healthcare management systems. This shift from conventional hub-based systems to personalized care relies on accurate causal models to guide interventions and treatment decisions^{[5][6]}.

Table 2: Causal Inference Applications in Healthcare

Application	Method	Key Insight	Source
Drug safety surveillance	Doubly robust estimation	Reduces confounding bias in observational studies	[5][6]
Treatment effect estimation	Instrumental variable analysis	Uses blood type as IV for liver transplant outcomes	[5][6]
Personalized medicine	Causal machine learning	Integrates multi-modal data for intervention planning	[7][8]

7. Ethics and Responsible Causal AI

Causal models magnify impact: prescriptive recommendations can harm if trained on biased data. Researchers must consider fairness constraints, audit trails, and alignment with regulations like the EU AI Act and IEEE S-7000 standards^{[7][8]}.

Transparent disclosure of assumptions and human-in-the-loop validations are paramount when deploying causal models in critical applications like healthcare and finance. This requires not only technical rigor but also engagement with domain experts and stakeholders^{[7][8]}.

8. Future Directions

Several promising research directions are emerging in causal discovery and estimation:

1. **Scaling discovery to larger networks:** Current methods struggle with high-dimensional data. Differentiable approaches show promise for scaling to thousands of variables, but further innovations are needed^{[4][8]}.
2. **Integration with deep learning:** Combining the representational power of neural networks with causal structure offers potential for handling complex, high-dimensional data while maintaining interpretability^{[7][8]}.
3. **Leveraging large language models:** Recent work explores how LLMs can support causal discovery in areas like automated machine learning and materials science^{[7][8]}.
4. **Robust methods for non-stationary data:** Many real-world systems exhibit time-varying causal relationships. Developing methods that can handle such dynamics remains an important challenge^{[3][7]}.

9. Conclusion

Causal discovery and estimation have matured from theoretical constructs to practical tools with applications across diverse domains. By moving beyond correlation to identify true causal relationships, these methods enable more reliable decision-making and deeper understanding of complex systems.

The integration of causal discovery with domain-specific knowledge and advanced computational techniques promises to further extend the reach and impact of causality-based approaches. As the field continues to evolve, we anticipate increasing adoption in critical applications where understanding "why" is as important as predicting "what."

References

1. J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge Univ. Press, 2009.
2. A. Zanga and F. Stella, "A Survey on Causal Discovery: Theory and Practice," *arXiv preprint arXiv:2305.10032*, 2023.
3. "Causal Discovery in Financial Markets: A Framework for Nonstationary Time-Series Data," *arXiv preprint*, Jan. 2024.
4. S. Bo and M. Xiao, "Root Cause Attribution of Delivery Risks via Causal Discovery with Reinforcement Learning," *arXiv preprint*, Jan. 2025.
5. X. Wu et al., "Causal Inference in the Medical Domain: A Survey," *Appl. Intell.*, 2024.
6. "Method Comparison and Estimation of Causal Effects of Insomnia on Health Outcomes," *Sci. Rep.*, 2023.
7. N. Pawlowski et al., "Causal Machine Learning for Healthcare and Precision Medicine," *R. Soc. Open Sci.*, 2022.
8. "Causal Inference with Complex Treatments: A Survey," *arXiv preprint*, Jul. 2024.