

# Review of Causal Discovery Methods Based on Graphical Models

---

Clark Glymour, Kun Zhang y Peter Spirtes

Departamento de Filosofía, Carnegie Mellon University, Pittsburgh, PA, Estados Unidos

1. Introducción
2. Modelos Causales Gráficos Dirigidos.
3. Conclusiones

# Introducción

---

Este artículo trata sobre la aplicación científica de un tipo de representación de relaciones causales, modelos causales gráficos (DGCM) dirigidos y métodos computarizados para encontrar verdaderas representaciones causales de ese tipo a partir de datos, ya sean observacionales, experimentales o ambos.

Comenzaremos con los llamados métodos basados en restricciones y basados en puntajes para el descubrimiento causal.

- PC asume que no hay un factor de confusión (causa común directa no observada de dos variables medidas), y su información causal descubierta es asintóticamente correcta.
- FCI proporciona resultados asintóticamente correctos incluso en presencia de factores de confusión.

Los algoritmos de PC y FCI producen representaciones gráficas de las clases de equivalencia (independencias).

También existen algoritmos basados en puntuación que tienen como objetivo encontrar la estructura causal mediante la optimización de una función de puntuación correctamente definida. Entre ellos, Greedy Equivalence Search (GES).

Por otra parte, recientemente se ha demostrado que los algoritmos basados en Modelos Causales Funcionales (FCM) correctamente definidos pueden distinguir entre diferentes Gráficos Acíclicos Dirigidos (DAG) en la misma clase de equivalencia.

# Modelos Causales Gráficos Dirigidos.

---

## Componentes de un DGCM:

1. un conjunto de variables, consideradas como "variables aleatorias".
2. un conjunto de aristas dirigidas entre pares de variables.
3. una distribución de probabilidad conjunta sobre los valores posibles de todas las variables.

## Consideraciones.

- Las variables se pueden indexar en el tiempo, formando un conjunto de procesos estocásticos relacionados causalmente.
- Las variables pueden ser no medidas
  - categóricas
  - ordinales
  - continuas
- Pueden haber fuertes variaciones en una variable a la cual se le considera ruido.
- Puede haber errores de medición.



Un DGCM asume que el valor de cada variable es una función determinista de sus causas directas en el grafo y sus perturbaciones no medidas.

La clase de DGCM incluye, pero es más general que, modelos de regresión, modelos de factores, modelos de series de tiempo ARM, modelos de clases latentes y otros.

Las DGCM, que no requieren condiciones iniciales (excepto en series de tiempo) ni condiciones de contorno, contrastan con los sistemas de ecuaciones diferenciales y parciales diferenciales, que también pueden ser representaciones de un sistema de relaciones causales.

Se tiene en cuenta que no todos los modelos gráficos dirigidos tienen interpretaciones causales.

Cuando se supone que la condición de Markov se mantiene para un gráfico causal y su distribución de población asociada, se denomina Asunción de Markov causal.

Es importante tener en cuenta que la separación  $d$  y las propiedades relacionadas proporcionan condiciones necesarias pero no suficientes para las relaciones de independencia condicional en la distribución de probabilidad conjunta sobre los valores de las variables.

## Conclusiones

---

- El algoritmo de PC y FCI, como métodos típicos que dependen de las relaciones de independencia condicional, requieren decisiones sobre la independencia condicional como entrada, lo cual es sencillo en casos lineales.
- Para las relaciones causales lineales, los procedimientos de búsqueda pueden escalar muy bien (por ejemplo, PC y GES pueden manejar fácilmente decenas de miles de variables para gráficos dispersos). Pero la salida puede no ser lo suficientemente informativa en ciertas circunstancias.
- Los métodos basados en modelos de ecuaciones estructurales tienen que recurrir a la forma funcional de la influencia causal y, en general, no pueden manejar los factores de confusión latentes de una manera directa.

- Los modelos causales funcionales no gaussianos o no lineales ayudan a identificar información más detallada del proceso causal; sin embargo, los métodos de búsqueda causal basados en ellos generalmente no se escalan tan bien como los métodos basados en independencia condicional.
- Para estimar LiNGAM, los métodos de estimación de dos pasos y FASK son factibles en miles de variables generadas por un gráfico disperso.
- Los métodos actuales para estimar modelos causales no lineales son factibles solo en docenas de variables.

# Conclusiones

	PC	FCI	GES	LiNGAM/PNL/ANM
¿Supuesto de fidelidad requerida?	NO	YES	Se requiere alguna condición más débil (aún no está totalmente clara)	NO
¿Supuestos específicos sobre la distribución de datos requeridos?	NO	NO	Sí (generalmente asume modelos lineales-gaussianos o distribuciones multinomiales)	YES
¿Manejar adecuadamente los confounders?	NO	YES	NO	NO
Output	Clase de equivalencia de Markov	Grafo ancestral parcial	clase de equivalencia de Markov	DAG, así como el modelo causal (bajo las condiciones de identificabilidad respectivas)