Graphical causal models: An algebraic perspective

Luis David García–Puente

Department of Mathematics and Statistics
Sam Houston State University

2013 Southern Regional Council on Statistics Summer Research Conference
Department of Bioinformatics and Biostatistics
University of Louisville
June 3, 2013
Consider a game consisting of coin flips where earlier coin outcomes affect
the biases of later coins in a prescribed way.

(Imagine I have some clear, heavy plastic that I can stick to the later coins
to give them any bias I want, on the fly.)
**Example "DACB"**

We can specify a coin biasing game with a diagram of how the coins influence each other, i.e., a **graph** on the coin names with a list of biases called a **conditional probability table (CPT)**, e.g.:

```
50%[0] + 50%[1]  

D  |  A
---|---
50%[0] + 60%[1]  

C  

DA -> C
[00] -> 90%[0] + 10%[1]
[01] -> 80%[0] + 20%[1]
[10] -> 70%[0] + 30%[1]

C -> B
[0] -> 20%[0] + 80%[1]
[1] -> 80%[0] + 20%[1]
```
Example "DACB"

Here, the first two coin flips are from two different coins, and the outcomes (0 or 1) are labelled $D$ and $A$.

\[
\begin{align*}
50\%[0] & + 50\%[1] \\
D & \\
\uparrow & \\
40\%[0] & + 60\%[1] \\
A \\
\downarrow & \\
C & \\
\uparrow & \\
DA \rightarrow C & \\
[00] & \rightarrow 90\%[0] + 10\%[1] \\
[01] & \rightarrow 80\%[0] + 20\%[1] \\
[10] & \rightarrow 70\%[0] + 30\%[1] \\
\downarrow & \\
B & \\
\uparrow & \\
C \rightarrow B & \\
[0] & \rightarrow 20\%[0] + 80\%[1] \\
[1] & \rightarrow 80\%[0] + 20\%[1]
\end{align*}
\]
**Example "DACB"**

Based on the outcome $DA$, a bias is chosen for another coin, which we flip and label its outcome $C$. Similarly $C$ determines a bias for the $B$ coin.

- **D**
  - $50\% [0] + 50\% [1]$

- **A**
  - $40\% [0] + 60\% [1]$

- **C**
  - $DA \rightarrow C$
    - $[00] \rightarrow 90\% [0] + 10\% [1]$
    - $[01] \rightarrow 80\% [0] + 20\% [1]$
    - $[10] \rightarrow 70\% [0] + 30\% [1]$
    - $[11] \rightarrow 60\% [0] + 40\% [1]$

- **B**
  - $C \rightarrow B$
    - $[0] \rightarrow 20\% [0] + 80\% [1]$
    - $[1] \rightarrow 80\% [0] + 20\% [1]$
Thus, a *coin-biasing game* is specified by data \((V, G, \Theta)\), where:

- \(V\) is a set of **binary random variables**, 
- \(G\) is a **directed acyclic graph** (DAG) called the **structure**, whose vertices are the variables, and 
- \(\Theta\) is a **conditional probability table** (CPT) specifying the values 

\[ P(V_i = v \mid \text{parents}(V_i) = \overline{w}), \]

for all \(i, v,\) and \(\overline{w}\).

Note: Without the binary restriction, this is the definition of a **Bayesian network** or **Bayes net** [J. Pearl, 1985].
Now suppose the “DACB” game is running inside a box, but we do not know its structure graph $G$ or the CPT parameters $\Theta$. Each time it runs, it prints us out a receipt showing the value of the variables $A, B, C,$ and $D$, in that order, but nothing else:

Say we got 10,000 such receipts, from which we estimate a probability table for the 16 possible outcomes...
From this probability table we can infer any correlational relationships we want. How about causality?

Stats 101 quiz:
From the probabilities alone, can we infer the causal structure of the game? What extra information is needed?
From this probability table we can infer any correlational relationships we want. How about causality?
From this probability table we can infer any correlational relationships we want. How about causality?

From the probabilities alone, can we infer $G$, the causal structure of the game?
From this probability table we can infer any correlational relationships we want. How about causality?

From the probabilities alone, can we infer $G$, the causal structure of the game? What extra information is needed?
The **probability data alone** is enough information to reliably distinguish the **causal structure** $G$ of the “DACB” game from other structures on 4 binary variables.
Nothing more is needed!

The probability data alone is enough information to reliably distinguish the causal structure $G$ of the “DACB” game from other structures on 4 binary variables.

The reason is that, by arising from $G$, the 16 probabilities $p_{0000}, p_{0001}, \ldots, p_{1111}$ are forced to satisfy a system of 13 polynomial equations $f_j = 0$ which encode [SECRET!] properties readable from the graph that do not depend on the CPT $\Theta$ [Pistone, Riccomagno, Wynn, 2001].

These equations are almost never satisfied by coin-biasing games arising from other graphs that are not subgraphs of $G$, and coin-biasing games arising from $G$ almost never satisfy conditional independence properties of its proper subgraphs.
For now, the take-away is:

Causality and Algebraic Geometry
Andrew Critch, UC Berkeley
critch@math.berkeley.edu

Coin- and die-biasing games

10,000 receipts

ABCD=1100
ABCD=1000
ABCD=1100
ABCD=0100
ABCD=1101
ABCD=...
In short, **causality** is the extent to which we can employ directed graphical models to predict and control real-world phenomena, i.e., it is how well we can **pretend nature is a die-biasing game**.

**Definition (J. Pearl, 2000; awarded the 2011 Turing Prize)**

A (fully specified) **causal theory** is defined by an ordered triple \((V, G, \Theta)\): a set of random variables, a DAG on those variables, and a compatible CPT. If not all of \(V\), often a subset \(O \subset V\) of **observed variables** is also specified, and the others are called **hidden variables**.
A joint probability distribution $P$ on the random variables $V$ is generated by $(G, \Theta)$ in the obvious way (like a die-biasing game),

$$P(v_1, \ldots, v_n) = \prod_i P(v_i | \text{parents}(v_i)).$$

With this framework in place, we can say that

- **causal hypotheses** are partial specifications of causal theories. For example, perhaps only $(V, G)$ is described, or only part of $G$.

- **causal inference** is the problem of recovering information about $(G, \Theta)$ from the probabilities $P$ or other partial information.
Empirically, causal hypotheses make two kinds of predictions:

- **Interventional predictions** - how you expect the system to respond if you start controlling parts of it, and

- **Observational predictions** - things you can see without manipulating the system.

In here, we will only focus on **observational predictions**.
For a given causal structure $G$ on a set of observed variables $V$ and unobserved variables $U$, there is a semi-algebraic set, or semivariety, of probability distributions $M_G$ which can arise from some parameter assignment $(CPT, \Theta)$ to $G$.

It turns out that $M_G$ is often a proper subset of the set $\Delta_V$ of all possible probability distributions on $V$. 
Exercise: for each of the following pairs of coin biasing graphs, try to determine if they are observationally equivalent, i.e., give rise to the same semivariety of distributions:
Exercise: for each of the following pairs of coin biasing graphs, try to determine if they are observationally equivalent, i.e., give rise to the same semivariety of distributions:

Answers:
Observational predictions

How can we tell them apart? The simplest observational predictions made by a causal theory that can be used to distinguish it are conditional independence statements implied by its graph $G$. These can be thought of as equational constraints on its semivariety of distributions.

“$A \perp \perp C \mid B$” is read

“A is independent of $C$ conditional on $B$”, and intuitively means

“When $B$ is known, learning $A$ gives no new information about $C$.”

In terms of probabilities, this is

$$P(A \mid B) \cdot P(C \mid B) = P(A, C \mid B)$$
Example: In a world where mating happens completely randomly...

\[ A \perp C \mid B, \text{ but } A \not\perp C \]

\[ A \perp C \mid B, \text{ but } A \not\perp C \]

\[ A \not\perp C \mid B, \text{ but } A \perp C. \]
Many mathematical subtleties arise in trying to infer the underlying graph of a causal process. Even die-biasing games on 3 and 4 random variables have causally important mathematical properties that are not immediately intuitive:

- Not every causal structure $G$ can be recovered uniquely from the outputs of a die-biasing game on it. Instead, DAGs come in small equivalence classes with other DAGs that are “observationally indistinguishable” from them.

- At least 3 random variables are required to test any causal relationship, observationally. (I.e, on two variables, only the DAG with no edge can be recovered, so $A \rightarrow B$ is indistinguishable from $B \rightarrow A$.)
The causal inference problem

Subtleties
The **observational equivalence class** of a causal diagram is determined by its “v-pattern”, i.e., the occurrence of induced subgraphs of the form
\[ \bullet \rightarrow \bullet \leftarrow \bullet \] (see Pearl, 2000):

![](image-url)
In other words, two causal structures on a set of variables $V$ have the same *semivariety* if and only if they have the same v-pattern.
The DACB graph has the nice property that it is the only graph in its observational equivalence class, so it uniquely determines its semivariety, and in fact its variety (the Zariski closure of its semivariety).
In summary, causal hypothesis:

- can still be made mathematically precise;

- imply testable predictions in form of interventions and conditional independences; and

- under the right circumstances can be reliably inferred from probabilities observed without interventions (controlled experiments).

The statistical package PCalg for R (Kalisch, Maechler, Colombo) incorporates algorithms for inferring DAG structures from data on causal structures of discrete and Gaussian variables.
When models involve hidden variables, even if they are not observationally equivalent, \textit{conditional independences} alone are not often enough to tell them apart, and finer algebraic invariants or inequalities are needed to distinguish their \textit{semivarieties}.
The following two models on binary variables satisfy the same conditional independencies (none!), but their semivarieties are respectively 9 and 15 dimensional:
The following two models on binary variables satisfy the same conditional independencies (none!), but their semivarieties are respectively 9 and 15 dimensional:

What is a good example of this phenomenon, on binary variables, where both models involve a hidden variable?
CANDIDATE 1:

These two models are observationally equivalent, even if we are allowed to view $H$. 

Diagram: 

[Diagram showing two models with nodes H, A, B, C, D, with arrows indicating relationships between nodes]
These models can be distinguished by **conditional independencies** among observed variables, namely, whether $A \perp \perp B$. 
These models are different, and cannot be distinguished by conditional independencies, so let’s examine this example.
The **semivariety** of this model lives in \( \Delta^{15} \subset \mathbb{C}^{15} \subset \mathbb{P}^{15} \), and its Zariski closure is the **first secant variety** of the **Segre embedding** of \( \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \). By [Raicu, 2010], this variety is cut out by the \( 3 \times 3 \) minors of the \( 4 \times 4 \) flattenings of the \( 2 \times 2 \times 2 \times 2 \) tensor \( (p_{abcd}) \). This ideal is minimally generated by 32 of the minors, which look like this:

\[
\begin{align*}
p_{0111}p_{1010}p_{1101} & - p_{0110}p_{1011}p_{1101} - p_{0111}p_{1001}p_{1110} + p_{0101}p_{1011}p_{1110} + p_{0110}p_{1001}p_{1111} - p_{0101}p_{1010}p_{1111}.
\end{align*}
\]
Model #1 ("naive Bayes")

In other words, by known results from algebraic geometry we can already write down 32 cubic equations which all the distributions arising from this model must satisfy, and which generate all other such equations.
This model is equivalent to the model on the right, whose semivariety Zariski closure is the \textbf{first secant variety} of the \textbf{Segre embedding} of $\mathbb{P}^3 \times \mathbb{P}^1 \times \mathbb{P}^1$, whose equations we also know by [Raicu, 2010]!
Model #2:

Finer algebraic invariants

Model # 2: The second model is equivalent to the model on the right, whose
semivariety Zariski closure is the first secant variety of the Segre
embedding of $P^3 \times P^1 \times P^1$, whose equations we also know by
$[Raicu, 2010]$!

Namely, they are the $3 \times 3$ minors of the unique $4 \times 4$ flattening of the
$4 \times 2 \times 2$ tensor $(p_{(ab)cd})$. 

![Diagram](image-url)
How can we use invariants like these, or related algebraic tools, to discover the relevance of hidden variables in natural datasets, like social network data, or medical survey data? How much data would we need? How much noise could we tolerate?
BIBLIOGRAPHY

- Pearl (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann