Variable elimination, graph reduction and efficient g-formula

F. Richard Guo
May 4th 2022, Fellows Seminar, Simons Institute

Statistical Laboratory, University of Cambridge
Ema Perković
*Washington*

Andrea Rotnitzky
*Torcuato Di Tella*
Suppose data comes from a causal DAG $G$ over vertices $V$. Treatment $A$ (discrete), outcome $Y$. **No unobserved confounders.**
Motivating example

Suppose data comes from a causal DAG $\mathcal{G}$ over vertices $\mathbf{V}$. Treatment $A$ (discrete), outcome $Y$. No unobserved confounders.

$$p = \prod_{v} p(v \mid \text{Pa}(v))$$

$$= p(a \mid i)p(y \mid a, o)p(i \mid w_1)p(o \mid w_1)$$
$$\times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).$$
Motivating example

Suppose data comes from a causal DAG $\mathcal{G}$ over vertices $\mathbf{V}$. Treatment $A$ (discrete), outcome $Y$. No unobserved confounders.

\[
p = \prod_v p(v \mid \text{Pa}(v)) \\
= p(a \mid i)p(y \mid a, o)p(i \mid w_1)p(o \mid w_1) \\
\times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).
\]

\[
p(\text{do}(A = a)) = \prod_{v \neq a} p(v \mid \text{Pa}(v)) \\
= p(y \mid a, o)p(i \mid w_1)p(o \mid w_1) \\
\times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).
\]
Suppose we are interested in the counterfactual mean $\mathbb{E} Y(a)$. 
g-formula

Suppose we are interested in the counterfactual mean \( \mathbb{E} Y(a) \).

\[ \mathbb{E} Y(a) = \psi_a(P; G) \equiv \sum_{v \not\in \{A\}} y \left\{ \prod_{v \neq a} P(v \mid Pa(v, G)) \right\} \bigg|_{A=a}. \]

g-formula (Robins, 1986)
aka manipulated distribution formula (Spirtes, Glymour, and Scheines, 2000), truncated factorization formula (Pearl, 2000)
Suppose we are interested in the counterfactual mean $\mathbb{E} Y(a)$.

**g-formula** (Robins, 1986)
aka *manipulated distribution formula* (Spirtes, Glymour, and Scheines, 2000), *truncated factorization formula* (Pearl, 2000)

\[
\mathbb{E} Y(a) = \Psi_a(P; G) \equiv \sum_{y} \left\{ \prod_{v \neq a} P(v | Pa(v, G)) \right\} \bigg| _{A=a}.
\]

\[
\Psi_a(P; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} yp(y \mid A = a, o)p(i \mid w_1)p(o \mid w_1)
\]
\[
\times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).
\]
Suppose we are interested in the counterfactual mean $\mathbb{E} Y(a)$.

**g-formula** (Robins, 1986)

aka *manipulated distribution formula* (Spirtes, Glymour, and Scheines, 2000), *truncated factorization formula* (Pearl, 2000)

\[
\mathbb{E} Y(a) = \Psi_a(P; G) \equiv \sum_{y \in Y(a)} \left\{ \prod_{v \notin A} P(v \mid Pa(v, G)) \right\} \bigg|_{A=a}.
\]

\[
\Psi_a(P; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} yp(y \mid A = a, o)p(i \mid w_1)p(o \mid w_1) \times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).
\]

Due to the factorization of $P$, in the model, $\mathbb{E} Y(a)$ can be expressed in different forms.
We have back-door formulae

\[ \psi_a^{\text{ADJ}}(P; G) = \mathbb{E}[\mathbb{E}[Y \mid A = a, L]], \]

where adjustment set L can take

\{O\}, \{I, O\}, \{I, W_1, O\}, \{I, O, W_1, W_2\}, \ldots.
Other identifying formulae

We have back-door formulae

$$\Psi_a^{ADJ}(P; G) = \mathbb{E}[\mathbb{E}[Y \mid A = a, L]],$$

where adjustment set $L$ can take

$$\{O\}, \{I, O\}, \{I, W_1, O\}, \{I, O, W_1, W_2\}, \ldots.$$

Which formula should I use?
We can compare the large-sample performance of plugin estimators of formulae. (Suppose all variables take only finitely many levels.)
We can compare the **large-sample performance** of plugin estimators of formulae. (Suppose all variables take only finitely many levels.)

**g-formula is efficient (MLE):**

\[
\psi_a(P_n; G) = \sum_{y,o,i,w_1,w_2,w_3,w_4} y P_n(y | A = a, o) P_n(i | w_1)
\]

\[
\times P_n(o | w_1) P_n(w_1 | w_2, w_3) P_n(w_3 | w_4) P_n(w_2) P_n(w_4).
\]
We can compare the **large-sample performance** of plugin estimators of formulae. (Suppose all variables take only finitely many levels.)

- **g-formula** is efficient (MLE):

\[
\Psi_a(P_n; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} y P_n(y \mid A = a, o) P_n(i \mid w_1) \\
\times P_n(o \mid w_1) P_n(w_1 \mid w_2, w_3) P_n(w_3 \mid w_4) P_n(w_2) P_n(w_4).
\]

- **Adjustment formulae** are not efficient:

\[
\Psi_a^{ADJ}(P_n; G) = \sum_L P_n(L) \sum_y y P_n(y \mid A = a, L)
\]
We can compare the large-sample performance of plugin estimators of formulae. (Suppose all variables take only finitely many levels.)

**g-formula is efficient (MLE):**

\[
\Psi_a(P_n; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} y P_n(y | A = a, o) P_n(i | w_1) \times P_n(o | w_1) P_n(w_1 | w_2, w_3) P_n(w_3 | w_4) P_n(w_2) P_n(w_4).
\]

**adjustment formulae are not efficient:**

\[
\Psi_{a}^{ADJ}(P_n; G) = \sum_{L} P_n(L) \sum_{y} y P_n(y | A = a, L)
\]

Yet, a simpler g-formula is also efficient!

\[
\Psi_a(P_n; G^*) = \sum_{y, o, w_2, w_3} y P_n[y | A = a, o] \times P_n(o | w_2, w_3) P_n(w_2) P_n(w_3).
\]
Another example

1. "back-door" (Pearl, 1993)

\[
A(P;G) = X \quad \text{o} \quad E[Y | A = a, m(p)]
\]

2. "front-door" (Pearl, 1995)

\[
A(P;G) = X \quad m(\quad X \quad A = a_0) \quad p(A = a_0)
\]

or the g-formula (Robins, 1986)

\[
A(P;G) = X \quad m, o \quad E[Y | m, o] \quad p(m | A = a)
\]

Which one should be preferred?

"Another area which is neglected in my opinion ... given an estimand find the best way of decomposing to estimate it" — Judea Pearl (OCIS, Nov 17, 2020).
1. “back-door” (Pearl, 1993)

$$\psi^\text{ADJ}_a(P; G) = \sum_o \mathbb{E}[Y \mid A = a, o] p(o),$$
Another example

1. “back-door” (Pearl, 1993)

\[ \psi_a^{\text{ADJ}}(P; G) = \sum_o \mathbb{E}[Y \mid A = a, o]p(o), \]

2. “front-door” (Pearl, 1995)

\[ \psi_a^{\text{FRONT}}(P; G) = \sum_m \left\{ \sum_{a} \mathbb{E}[Y \mid m, A = a']p(A = a') \right\} p(m \mid A = a), \]
Another example

1. “back-door” (Pearl, 1993)

\[ \Psi_a^{\text{ADJ}}(P; \mathcal{G}) = \sum_o \mathbb{E}[Y \mid A = a, o]p(o), \]

2. “front-door” (Pearl, 1995)

\[ \Psi_a^{\text{FRONT}}(P; \mathcal{G}) = \sum_m \left\{ \sum_a \mathbb{E}[Y \mid m, A = a']p(A = a') \right\} p(m \mid A = a), \]

3. or the g-formula (Robins, 1986)

\[ \Psi_a(P; \mathcal{G}) = \sum_{m,o} \mathbb{E}[Y \mid m, o]p(o)p(m \mid A = a). \]
Another example

1. “back-door” (Pearl, 1993)

\[ \Psi_a^{ADJ}(P; G) = \sum_o \mathbb{E}[Y \mid A = a, o]p(o), \]

2. “front-door” (Pearl, 1995)

\[ \Psi_a^{FRONT}(P; G) = \sum_m \left\{ \sum_a \mathbb{E}[Y \mid m, A = a']p(A = a') \right\} p(m \mid A = a), \]

3. or the g-formula (Robins, 1986)

\[ \Psi_a(P; G) = \sum_{m, o} \mathbb{E}[Y \mid m, o]p(o)p(m \mid A = a). \]

Which one should be preferred?
Another example

1. “back-door” (Pearl, 1993)
\[ \psi_a^{\text{ADJ}}(P; G) = \sum_o \mathbb{E}[Y \mid A = a, o]p(o), \]

2. “front-door” (Pearl, 1995)
\[ \psi_a^{\text{FRONT}}(P; G) = \sum_m \left\{ \sum_a \mathbb{E}[Y \mid m, A = a']p(A = a') \right\} p(m \mid A = a), \]

3. or the g-formula (Robins, 1986)
\[ \psi_a(P; G) = \sum_{m,o} \mathbb{E}[Y \mid m, o]p(o)p(m \mid A = a). \]

Which one should be preferred? “Another area which is neglected in my opinion ... given an estimand find the best way of decomposing to estimate it” — Judea Pearl (OCIS, Nov 17, 2020).
Semiparametric efficiency

The semiparametric efficiency bound is defined with respect to $M(G, V) := \{P(V) : P \text{ factorizes according to } G\}$.

The efficiency bound for estimating $E_Y(a)$ at $P(a)$ (assuming positivity) with respect to $M(G, V)$ is characterized by the efficient influence function $\varepsilon^D, P(V; G)$.

Observation: for some graphs, certain variables “cancel out” from $\varepsilon^D, P(V; G)$.
The semiparametric efficiency bound is defined with respect to

$$\mathcal{M}(\mathcal{G}, \mathbf{V}) := \{P(\mathbf{V}) : P \text{ factorizes according to } \mathcal{G}\}.$$
The semiparametric efficiency bound is defined with respect to

\[ \mathcal{M}(G, \mathbf{V}) := \{P(\mathbf{V}) : P \text{ factorizes according to } G\}. \]

The efficiency bound for estimating \( \mathbb{E} Y(a) \) at \( P \) (assuming positivity) with respect to \( \mathcal{M}(G, \mathbf{V}) \) is characterized by the efficient influence function \( \chi_{\text{eff},P}(\mathbf{V}; G) \).
The **semiparametric** efficiency bound is defined with respect to

\[ \mathcal{M}(\mathcal{G}, \mathcal{V}) := \{ P(\mathcal{V}) : P \text{ factorizes according to } \mathcal{G} \} . \]

- The efficiency bound for estimating \( \mathbb{E} Y(a) \) at \( P \) (assuming positivity) with respect to \( \mathcal{M}(\mathcal{G}, \mathcal{V}) \) is characterized by the **efficient influence function** \( \chi_{\text{eff}, P}(\mathcal{V} ; \mathcal{G}) \).

- Observation: for some graphs, certain variables “cancels out” from \( \chi_{\text{eff}, P}(\mathcal{V} ; \mathcal{G}) \).
Lemma (Rotnitzky and Smucler, 2020) Let \( \mathcal{M} \) be a semiparametric model on vector \( \mathbf{V} \). Suppose \( \mathbf{V}' \) is a subvector of \( \mathbf{V} \), such that

1. \( \psi(P) \) depends on \( P \) only through margin \( P(\mathbf{V}') \)

2. and \( \chi_{\text{eff},P}(\mathbf{V}; \mathcal{M}) \) only depends on \( \mathbf{V} \) through \( \mathbf{V}' \) for every \( P \in \mathcal{M}(\mathbf{V}) \),

then

\[
\chi_{\text{eff},P}(\mathbf{V}; \mathcal{M}) = \chi_{\text{eff},P}(\mathbf{V}'; \mathcal{M}') \quad P\text{-a.s. for every } P \in \mathcal{M},
\]

where \( \mathcal{M}' \equiv \{ P(\mathbf{V}') : P(\mathbf{V}) \in \mathcal{M} \} \) is the induced marginal model over \( \mathbf{V}' \).
Lemma (Rotnitzky and Smucler, 2020) Let $\mathcal{M}$ be a semiparametric model on vector $\mathbf{V}$. Suppose $\mathbf{V}'$ is a subvector of $\mathbf{V}$, such that

1. $\Psi(P)$ depends on $P$ only through margin $P(\mathbf{V}')$
2. and $\chi_{\text{eff},P}(\mathbf{V};\mathcal{M})$ only depends on $\mathbf{V}$ through $\mathbf{V}'$ for every $P \in \mathcal{M}(\mathbf{V})$,

then

$$\chi_{\text{eff},P}(\mathbf{V};\mathcal{M}) = \chi_{\text{eff},P}(\mathbf{V}';\mathcal{M}') \quad P\text{-a.s. for every } P \in \mathcal{M},$$

where $\mathcal{M}' \equiv \{P(\mathbf{V}') : P(\mathbf{V}) \in \mathcal{M}\}$ is the induced marginal model over $\mathbf{V}'$.

Definition Given graph $\mathcal{G}$ over $\mathbf{V}$, we say subset $\mathbf{U} \subset \mathbf{V}$ is uninformative for estimating $\mathbb{E} Y(a)$ if

1. $\mathbb{E} Y(a)$ is identified from $P(\mathbf{V} \setminus \mathbf{U})$,
2. and $\chi_{\text{eff},P}(\mathbf{V};\mathcal{G})$ does not depend on $\mathbf{U}$ $P$-a.e. for all $P \in \mathcal{M}(\mathcal{G},\mathbf{V})$. 
Lemma (Rotnitzky and Smucler, 2020) Let $\mathcal{M}$ be a semiparametric model on vector $\mathbf{V}$. Suppose $\mathbf{V'}$ is a subvector of $\mathbf{V}$, such that

1. $\Psi(P)$ depends on $P$ only through margin $P(\mathbf{V'})$
2. and $\chi_{\text{eff},P}(\mathbf{V};\mathcal{M})$ only depends on $\mathbf{V}$ through $\mathbf{V'}$ for every $P \in \mathcal{M}(\mathbf{V})$,

then

$$
\chi_{\text{eff},P}(\mathbf{V};\mathcal{M}) = \chi_{\text{eff},P}(\mathbf{V'};\mathcal{M}') \quad P\text{-a.s. for every } P \in \mathcal{M},
$$

where $\mathcal{M}' \equiv \{P(\mathbf{V'}) : P(\mathbf{V}) \in \mathcal{M}\}$ is the induced marginal model over $\mathbf{V'}$.

Definition Given graph $\mathcal{G}$ over $\mathbf{V}$, we say subset $\mathbf{U} \subset \mathbf{V}$ is uninformative for estimating $\mathbb{E} Y(a)$ if

1. $\mathbb{E} Y(a)$ is identified from $P(\mathbf{V} \setminus \mathbf{U})$,
2. and $\chi_{\text{eff},P}(\mathbf{V};\mathcal{G})$ does not depend on $\mathbf{U}$ $P$-a.e. for all $P \in \mathcal{M}(\mathcal{G}, \mathbf{V})$.

Irreducible informative set $\mathbf{V}^*(\mathcal{G}) \equiv \{\text{smallest } \mathbf{V'} : \mathbf{V} \setminus \mathbf{V'} \text{ is uninformative}\}$. 
1. **Variable elimination**: identify informative variables $V^*(G)$.
2. **Graph reduction**: characterize the marginal model over $V^*(G)$.
3. Derive a **simpler g-formula**.
Taxonomy of vertices

We write $u \rightarrow v$ if $u$ is an ancestor of $v$. (We always suppose $A \rightarrow Y$.)

- $N(G) = \{v : v \not\rightarrow Y\}$
- $I(G) = \{v : v \not= A : v \rightarrow Y$ only through $A\}$
- $W(G) = \{v / I(G) : A \not\rightarrow v, v \rightarrow Y\}$
- $M(G) = \{v : A \rightarrow v \not\rightarrow Y\}$

- baseline covariates
- mediators
We write $u \mapsto v$ if $u$ is an ancestor of $v$. (We always suppose $A \mapsto Y$.)
We write $u \rightarrow v$ if $u$ is an ancestor of $v$. (We always suppose $A \rightarrow Y$.)
Taxonomy of vertices

We write $u \rightarrow v$ if $u$ is an ancestor of $v$. (We always suppose $A \rightarrow Y$.)

- $N(G) = \{v : v \not\rightarrow Y\}$
- non-ancestors of $Y$
We write \( u \mapsto v \) if \( u \) is an ancestor of \( v \). (We always suppose \( A \mapsto Y \).)

- \( N(G) = \{ v : v \not\mapsto Y \} \)
  - non-ancestors of \( Y \)
- \( I(G) = \{ v \neq A : v \mapsto Y \text{ only through } A \} \)
  - conditional instruments
We write $u \mapsto v$ if $u$ is an ancestor of $v$. (We always suppose $A \mapsto Y$.)

- $N(G) = \{ v : v \not\mapsto Y \}$
  - non-ancestors of $Y$
- $I(G) = \{ v \not= A : v \mapsto Y$ only through $A$\}
  - conditional instruments
- $W(G) = \{ v \not\in I(G) : A \not\mapsto v, v \mapsto Y \}$
  - baseline covariates
Taxonomy of vertices

We write \( u \rightarrow v \) if \( u \) is an ancestor of \( v \). (We always suppose \( A \rightarrow Y \).)

- \( N(G) = \{ v : v \nleftrightarrow Y \} \)  
  - non-ancestors of \( Y \)
- \( I(G) = \{ v \neq A : v \rightarrow Y \text{ only through } A \} \)  
  - conditional instruments
- \( W(G) = \{ v \notin I(G) : A \nleftrightarrow v, v \rightarrow Y \} \)  
  - baseline covariates
- \( M(G) = \{ v : A \rightarrow v \rightarrow Y \} \)  
  - mediators
We write $u \rightarrow v$ if $u$ is an ancestor of $v$. (We always suppose $A \rightarrow Y$.)

- $\mathbf{N}(\mathcal{G}) = \{v : v \not\leftrightarrow Y\}$
  - non-ancestors of $Y$
- $\mathbf{I}(\mathcal{G}) = \{v \neq A : v \leftrightarrow Y \text{ only through } A\}$
  - conditional instruments
- $\mathbf{W}(\mathcal{G}) = \{v \notin \mathbf{I}(\mathcal{G}) : A \not\leftrightarrow v, v \leftrightarrow Y\}$
  - baseline covariates
- $\mathbf{M}(\mathcal{G}) = \{v : A \leftrightarrow v \leftrightarrow Y\}$
  - mediators

A special subset of $\mathbf{W}$ plays an important role:

$$\mathbf{O}(\mathcal{G}) := \text{Pa}(\mathbf{M} \cup \{Y\}) \setminus (\text{De}(\mathbf{M} \cup \{Y\}) \cup \{A\})$$

is the optimal adjustment set (Henckel, Perković, and Maathuis, 2022; Rotnitzky and Sandrock, 2020).

($\mathbf{O}$ consists of direct parents of $\mathbf{M}$ or $Y$ that do not block causal paths.)
Using conditional independences on $G$, Rotnitzky and Smucler (2020) showed

$$
\chi_{\text{eff}, P}(V, G) = \sum_{j=1}^{J} (\mathbb{E}[b(O) \mid W_j, \text{Pa}(W_j)] - \mathbb{E}[b(O) \mid \text{Pa}(W_j)])
$$

$$
+ \sum_{k=1}^{K+1} (\mathbb{E}[AY/\pi(O) \mid M_k, \text{Pa}(M_k)] - \mathbb{E}[AY/\pi(O) \mid \text{Pa}(M_k)]),
$$

where $M_{K+1} \equiv Y$, $b(O) = \mathbb{E}[Y \mid A = 1, O]$, $\pi(O) = P(A = 1 \mid O)$. 

N and I are uninformative
Using conditional independences on $G$, Rotnitzky and Smucler (2020) showed

$$
\chi_{\text{eff}, P}(\mathbf{V}, G) = \sum_{j=1}^{J} \left( \mathbb{E}[b(\mathbf{O}) \mid W_j, \text{Pa}(W_j)] - \mathbb{E}[b(\mathbf{O}) \mid \text{Pa}(W_j)] \right)
+ \sum_{k=1}^{K+1} \left( \mathbb{E}[A Y / \pi(\mathbf{O}) \mid M_k, \text{Pa}(M_k)] - \mathbb{E}[A Y / \pi(\mathbf{O}) \mid \text{Pa}(M_k)] \right),
$$

where $M_{K+1} \equiv Y$, $b(\mathbf{O}) = \mathbb{E}[Y \mid A = 1, \mathbf{O}]$, $\pi(\mathbf{O}) = P(A = 1 \mid \mathbf{O})$.

**Corollary** This implies that

$$
\mathbf{N}(G) \equiv \{\text{non-ancestors of } Y\} \quad \text{and} \quad \mathbf{I}(G) \equiv \{\text{indirect ancestors of } Y\}
$$

are uninformative.
N and I are uninformative

Using conditional independences on \( G \), Rotnitzky and Smucler (2020) showed

\[
\chi_{\text{eff}, P}(V, G) = \sum_{j=1}^{J} (\mathbb{E}[b(O) | W_j, \text{Pa}(W_j)] - \mathbb{E}[b(O) | \text{Pa}(W_j)]) \\
+ \sum_{k=1}^{K+1} (\mathbb{E}[AY / \pi(O) | M_k, \text{Pa}(M_k)] - \mathbb{E}[AY / \pi(O) | \text{Pa}(M_k)])
\]

where \( M_{K+1} \equiv Y \), \( b(O) = \mathbb{E}[Y | A = 1, O] \), \( \pi(O) = P(A = 1 | O) \).

**Corollary** This implies that

\[
N(G) \equiv \{ \text{non-ancestors of } Y \} \quad \text{and} \quad I(G) \equiv \{ \text{indirect ancestors of } Y \}
\]

are uninformative.

Yet, \( N(G) \) and \( I(G) \) are defined with respect to \( G \) rather than \( M(G, V) \).
Causal Markov equivalence

Two DAGs \( G \) and \( G_0 \) are Markov equivalent if they define the same set of models.

\[ M(G, V) = M(G_0, V) \]

Markov equivalence class (MEC) (Andersson, Madigan, and Perlman, 1997; Verma and Pearl, 1991)

\( G' \rightarrow G \rightarrow G_0 \)

is the same adjacency and unshielded colliders (unshielded collider: \( a \rightarrow b \) with \( a \) and \( b \) non-adjacent).

But Markov equivalence does not preserve the causal interpretation.
Two DAGs $G$ and $G'$ are **Markov equivalent** if they define the same set of models

$$G \sim G' \iff M(G, V) = M(G', V).$$
Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are **Markov equivalent** if they define the same set of models

$$\mathcal{G} \sim \mathcal{G}' \iff \mathcal{M}(\mathcal{G}, \mathcal{V}) = \mathcal{M}(\mathcal{G}', \mathcal{V}).$$

**Markov equivalence class (MEC)** (Andersson, Madigan, and Perlman, 1997; Verma and Pearl, 1991)

$\mathcal{G} \sim \mathcal{G}' \iff \mathcal{G}$ and $\mathcal{G}'$ share the same adjacencies and unshielded colliders

(unshielded collider: $a \rightarrow \circ \leftarrow b$ with $a$ and $b$ non-adjacent.)
Causal Markov equivalence

Two DAGs $G$ and $G'$ are **Markov equivalent** if they define the same set of models

$$G \simeq G' \iff M(G, V) = M(G', V).$$

**Markov equivalence class (MEC)** (Andersson, Madigan, and Perlman, 1997; Verma and Pearl, 1991)

$G \simeq G'$ $\iff$ $G$ and $G'$ share the same **adjacencies** and **unshielded colliders**

(unshielded collider: $a \rightarrow o \leftarrow b$ with $a$ and $b$ non-adjacent.)

But Markov equivalence does not preserve the causal interpretation:

\[ \begin{array}{c}
\text{A} \rightarrow Y \\
\Downarrow \\
\text{O} \\
\Downarrow \\
\text{Y} \end{array} \quad \simeq \quad \begin{array}{c}
\text{A} \rightarrow Y \\
\Downarrow \\
\text{O} \\
\Downarrow \\
\text{Y} \end{array} \]
Causal Markov equivalence (with respect to the effect of $A$ on $Y$):

$$\mathcal{G} \sim \mathcal{G}' \iff \mathcal{G} \sim \mathcal{G}' \text{ and } \Psi(P, \mathcal{G}) = \Psi(P, \mathcal{G}') \text{ for all } P \in \mathcal{M}_\mathcal{G}.$$
Causal Markov equivalence (with respect to the effect of $A$ on $Y$):

\[ G \overset{c}{\sim} G' \iff G \sim G' \text{ and } \Psi(P, G) = \Psi(P, G') \text{ for all } P \in \mathcal{M}_G. \]

The causal Markov equivalence class \textbf{(c-MEC)} is characterized by Guo and Perković (2021) (with MPDAGs).
Causal Markov equivalence (with respect to the effect of $A$ on $Y$):

$$G \sim G' \iff G \sim G' \text{ and } \Psi(P, G) = \Psi(P, G') \text{ for all } P \in \mathcal{M}_G.$$ 

The causal Markov equivalence class (c-MEC) is characterized by Guo and Perković (2021) (with MPDAGs).

Theorem $\bigcup_{G' \sim G} I(G') \cup N(G')$ is uninformative.
Causal Markov equivalence (with respect to the effect of $A$ on $Y$):

$$\mathcal{G} \sim \mathcal{G}' \iff \mathcal{G} \sim \mathcal{G}' \text{ and } \Psi(P, \mathcal{G}) = \Psi(P, \mathcal{G}') \text{ for all } P \in \mathcal{M}_\mathcal{G}.$$ 

The causal Markov equivalence class (c-MEC) is characterized by Guo and Perković (2021) (with MPDAGs).

**Theorem** \( \bigcup_{\mathcal{G}'} \mathcal{I}(\mathcal{G}') \cup \mathcal{N}(\mathcal{G}') \) is uninformative.

It can be further shown that there exists $\mathcal{G} \sim \mathcal{G}$ such that

$$\mathcal{I}(\mathcal{G}) = \bigcup_{\mathcal{G}' \sim \mathcal{G}} \mathcal{I}(\mathcal{G}'), \quad \mathcal{N}(\mathcal{G}) = \bigcup_{\mathcal{G}' \sim \mathcal{G}} \mathcal{N}(\mathcal{G}').$$
Causal Markov equivalence

Causal Markov equivalence (with respect to the effect of $A$ on $Y$):

$$G \sim G' \iff G \sim G' \quad \text{and} \quad \Psi(P, G) = \Psi(P, G') \quad \text{for all} \quad P \in \mathcal{M}_G.$$ 

The causal Markov equivalence class \textbf{c-MEC} is characterized by Guo and Perković (2021) (with MPDAGs).

**Theorem** $\bigcup_{G' \sim G} I(G') \cup N(G')$ is uninformative.

It can be further shown that there exists $\bar{G} \sim G$ such that

$$I(\bar{G}) = \bigcup_{G' \sim G} I(G'), \quad N(\bar{G}) = \bigcup_{G' \sim G} N(G').$$

Vertices can be determined uninformative by moving \textbf{within} the causal Markov equivalence class (flipping edges).
I and $W_4$ are uninformative
I and $W_4$ are uninformative
I and $W_4$ are uninformative
\( \chi_{\text{eff},P}(V, G) \) depends on \( W_1 \) thought the terms

\[
\mathbb{E}\{b_a(O) \mid W_1, W_2, W_3\} + b_a(O) - \mathbb{E}\{b_a(O) \mid W_1\}
\]

due to \( O \perp_G W_2, W_3 \mid W_1 \), where \( b_a(O) \equiv \mathbb{E}[Y \mid A = a, O] \).
\( \chi_{\text{eff}, P}(V, G) \) depends on \( W_1 \) thought the terms

\[
\mathbb{E}\{b_a(O) \mid W_1, W_2, W_3\} + b_a(O) - \mathbb{E}\{b_a(O) \mid W_1\}
\]

due to \( O \perp_G W_2, W_3 \mid W_1 \), where \( b_a(O) \equiv \mathbb{E}[Y \mid A = a, O] \).

**Uninformative** \( W_1 \) cannot be detected this way!
Conspired cancellation

Fix $W_j$ $\cap O L e t W_j Ch(W_j)$ be topologically ordered as $

\{W_j^1, \ldots, W_j^r\}$.

Then the EIF only depends on $W_j$ through the terms:

$$+

E[b(O)|W_j, Pa(W_j)] + E[b(O)|W_j^1, Pa(W_j^1)] + E[b(O)|W_j^2, Pa(W_j^2)] \cdots + E[b(O)|W_j^r, Pa(W_j^r)]$$

To make this happen, we can posit the following graphical $W$-criterion:

1. $W_j \nsubseteq G O | W_j^r, Pa(W_j^r) \cap \{W_j\}$,

2. and for $m = 1, \ldots, r$:
   (i) $W_j^m \neq W_j^m (\text{children are chained})$
   (ii) $Pa(W_j^m) \checkmark Pa(W_j^m) \cap \{W_j^m\} (\text{parent sets are decreasing})$
   (iii) $Pa(W_j^m) \cap Pa(W_j^m) \nsubseteq G O | Pa(W_j^m) (\text{left-over piece is separated from O})$

As imilargraphic M-criterion applies to mediator $M_{i2} M_{i2}$. 

18
Fix $W_j \equiv W_{j_0} \in W \setminus O$. Let $W \cap \text{Ch}(W_j)$ be topo-sorted as $\{W_{j_1}, \ldots, W_{j_r}\}$. 
Fix $W_j \equiv W_{j_0} \in \mathbf{W} \setminus \mathbf{O}$. Let $\mathbf{W} \cap \text{Ch}(W_j)$ be topo-sorted as \{\(W_{j_1}, \ldots, W_{j_r}\}\).

Then the EIF only depends on $W_j$ through the terms:

\[ + \mathbb{E}[b(O) \mid W_j, \text{Pa}(W_j)] + \mathbb{E}[b(O) \mid W_{j_1}, \text{Pa}(W_{j_1})] + \mathbb{E}[b(O) \mid W_{j_2}, \text{Pa}(W_{j_2})] + \cdots + \mathbb{E}[b(O) \mid W_{j_r}, \text{Pa}(W_{j_r})] \]

\[ - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_1})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_2})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_3})] - \cdots \]
Conspired cancellation

Fix $W_j \equiv W_{j0} \in W \setminus O$. Let $W \cap Ch(W_j)$ be topo-sorted as $\{W_{j1}, \ldots, W_{jr}\}$.

Then the EIF only depends on $W_j$ through the terms:

$$+ \mathbb{E}[b(O) \mid W_j, Pa(W_j)] + \mathbb{E}[b(O) \mid W_{j1}, Pa(W_{j1})] + \mathbb{E}[b(O) \mid W_{j2}, Pa(W_{j2})] + \cdots + \mathbb{E}[b(O) \mid W_{jr}, Pa(W_{jr})]$$

$$- \mathbb{E}[b(O) \mid Pa(W_{j1})] - \mathbb{E}[b(O) \mid Pa(W_{j2})] - \mathbb{E}[b(O) \mid Pa(W_{j3})] - \cdots$$

To make this happen, we can posit the following graphical $W$-criterion:

1. $W_j \supseteq G \cap W_{jr}, Pa(W_{jr}) \cap \{W_{jr}\}$,
2. and for $m = 1, \ldots, r$:
   (i) $W_{jm} \neq W_{jm+1}$ (children are chained)
   (ii) $Pa(W_{jm}) \nsupseteq Pa(W_{jm+1}) \cap \{W_{jm+1}\}$ (parent sets are decreasing)
   (iii) $Pa(W_{jm}) \cap Pa(W_{jm+1}) \supseteq G \cap Pa(W_{jm})$ (left-over piece is separated from $O$)

As similar graphical $M$-criterion applies to mediator $M_i$. 


Conspired cancellation

Fix $W_j \equiv W_{j_0} \in W \setminus O$. Let $W \cap \text{Ch}(W_j)$ be topo-sorted as $\{W_{j_1}, \ldots, W_{j_r}\}$.

Then the EIF only depends on $W_j$ through the terms:

$$+ \mathbb{E}[b(O) \mid W_j, \text{Pa}(W_j)] + \mathbb{E}[b(O) \mid W_{j_1}, \text{Pa}(W_{j_1})] + \mathbb{E}[b(O) \mid W_{j_2}, \text{Pa}(W_{j_2})] \cdots + \mathbb{E}[b(O) \mid W_{j_r}, \text{Pa}(W_{j_r})]$$

$$- \mathbb{E}[b(O) \mid \text{Pa}(W_{j_1})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_2})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_3})] \cdots$$

To make this happen, we can posit the following graphical $W$-criterion:

1. $W_j \perp_g O \mid W_{j_r}, \text{Pa}(W_{j_r}) \setminus \{W_j\}$,

2. and for $m = 1, \ldots, r$:
   
   (i) $W_{j_{m-1}} \rightarrow W_{j_m}$ (children are chained)
   
   (ii) $\text{Pa}(W_{j_m}) \subseteq \text{Pa}(W_{j_{m-1}}) \cup \{W_{j_{m-1}}\}$ (parent sets are decreasing)
   
   (iii) $\text{Pa}(W_{j_{m-1}}) \setminus \text{Pa}(W_{j_m}) \perp_g O \mid \text{Pa}(W_{j_m})$ (left-over piece is separated from $O$)
Conspired cancellation

Fix $W_j \equiv W_{j_0} \in W \setminus O$. Let $W \cap \text{Ch}(W_j)$ be topo-sorted as $\{W_{j_1}, \ldots, W_{j_r}\}$.

Then the EIF only depends on $W_j$ through the terms:

$$+ \mathbb{E}[b(O) \mid W_j, \text{Pa}(W_j)] + \mathbb{E}[b(O) \mid W_{j_1}, \text{Pa}(W_{j_1})] + \mathbb{E}[b(O) \mid W_{j_2}, \text{Pa}(W_{j_2})] \ldots + \mathbb{E}[b(O) \mid W_{j_r}, \text{Pa}(W_{j_r})]$$

$$- \mathbb{E}[b(O) \mid \text{Pa}(W_{j_1})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_2})] - \mathbb{E}[b(O) \mid \text{Pa}(W_{j_3})] \ldots$$

To make this happen, we can posit the following graphical $W$-criterion:

1. $W_j \perp_g O \mid W_{j_r}, \text{Pa}(W_{j_r}) \setminus \{W_j\}$,

2. and for $m = 1, \ldots, r$:

   (i) $W_{j_{m-1}} \rightarrow W_{j_m}$ (children are chained)

   (ii) $\text{Pa}(W_{j_m}) \subseteq \text{Pa}(W_{j_{m-1}}) \cup \{W_{j_{m-1}}\}$ (parent sets are decreasing)

   (iii) $\text{Pa}(W_{j_{m-1}}) \setminus \text{Pa}(W_{j_m}) \perp_g O \mid \text{Pa}(W_{j_m})$ (left-over piece is separated from $O$)

A similar graphical $M$-criterion applies to mediator $M_i \in M$.  

**Theorem** The set of informative variables is given by

\[ V^*(G) = \{A, Y\} \cup O \]

\[ \cup \{W_j \in W \setminus O : W_j \text{ fails the W-criterion}\} \]

\[ \cup \{M_i \in M : M_i \text{ fails the M-criterion}\}. \]
Theorem} The set of informative variables is given by

\[ V^*(\mathcal{G}) = \{A, Y\} \cup O \cup \{W_j \in W \setminus O : W_j \text{ fails the } W\text{-criterion}\} \cup \{M_i \in M : M_i \text{ fails the } M\text{-criterion}\}. \]

Proof sketch:

1. \(A, Y, O\) are informative.
Theorem The set of informative variables is given by

\[ V^*(G) = \{A, Y\} \cup O \]
\[ \cup \{W_j \in W \setminus O : W_j \text{ fails the } W\text{-criterion}\} \]
\[ \cup \{M_i \in M : M_i \text{ fails the } M\text{-criterion}\}. \]

Proof sketch:

1. \(A, Y, O\) are informative.

2. \(W_j/M_i\) satisfies the W/M-criterion \(\implies\) \(W_j/M_i\) is uninformative by conditional independence.
**Theorem** The set of informative variables is given by

\[
V^*(G) = \{A, Y\} \cup O \\
\cup \{W_j \in W \setminus O : W_j \text{ fails the } W\text{-criterion}\} \\
\cup \{M_i \in M : M_i \text{ fails the } M\text{-criterion}\}.
\]

Proof sketch:

1. \(A, Y, O\) are informative.
2. \(W_j/M_i\) satisfies the \(W/M\)-criterion \(\implies W_j/M_i\) is uninformative
   - by conditional independence.
3. \(W_j/M_i\) fails the \(W/M\)-criterion \(\implies W_j/M_i\) is informative
   - by constructing certain \(P \in \mathcal{M}(G, V)\) such that \(\chi_{\text{eff}, P}(V, G)\) depends on \(W_j/M_i\).
How do we represent the following marginal model of a DAG?

\[
\mathcal{M}(\mathcal{G}, V^*) \equiv \{P(V^*): P \in \mathcal{M}(\mathcal{G}, V)\},
\]

where \(V^* \equiv V^*(\mathcal{G})\).

Marginal models of a DAG can be complicated.
Latent projection

One popular approach is the latent projection (Verma and Pearl, 1991).
One popular approach is the **latent projection** (Verma and Pearl, 1991).

Suppose $\mathcal{G} = (\mathbf{V} \cup \mathbf{U}, \mathbf{E})$ for observed $\mathbf{V}$ and latent $\mathbf{U}$.

1. Whenever there is a path of the form $\bowtie \xrightarrow{u_1} \cdots \xrightarrow{u_2} \mathbf{v}$, add $\bowtie \xrightarrow{\mathbf{w}} \mathbf{v}$ (if not already present).

2. Whenever there is a path of the form $\mathbf{w} \xleftarrow{u_1} \cdots \xleftarrow{u_2} \mathbf{v}$, add $\mathbf{w} \xleftarrow{\mathbf{v}}$.

Results in an ADMG (not always a DAG). Conceptually strange.
One popular approach is the **latent projection** (Verma and Pearl, 1991).

Suppose $\mathcal{G} = (\mathbf{V} \cup \mathbf{U}, \mathbf{E})$ for observed $\mathbf{V}$ and latent $\mathbf{U}$.

1. Whenever there is a path of the form $w u_1 \ldots u_2 v$, add $w \rightarrow v$ (if not already present).
2. Whenever there is a path of the form $w u_1 \ldots u_2 v$, add $w \leftarrow v$.

Results in an ADMG (not always a DAG). Conceptually strange.
One popular approach is the **latent projection** (Verma and Pearl, 1991).

Suppose $\mathcal{G} = (\mathbf{V} \cup \mathbf{U}, \mathbf{E})$ for observed $\mathbf{V}$ and latent $\mathbf{U}$.

1. Whenever there is a path of the form $w_u \ldots u_v$, add $w \rightarrow v$ (if not already present).
2. Whenever there is a path of the form $w_u \ldots u_v$, add $w \leftarrow v$.

Results in an ADMG (not always a DAG). Conceptually strange.
Graph reduction algorithm

\[
\begin{align*}
V^* & \leftarrow \{A, Y\} \cup W \cup M \\
G^* & \leftarrow G(V^*) \text{ by projecting out } N \text{ and } I \text{ with latent projection} \\
\text{for } \nu \in V^* \text{ do} & \\
\quad \text{if } (\nu \in W \setminus O \text{ and } \nu \text{ satisfies the W-criterion) or (} \nu \in M \text{ and } \nu \text{ satisfies the M-criterion) then} & \\
\quad \quad & V^* \leftarrow V^* \setminus \{\nu\} \\
\quad \quad & G^* \leftarrow G_{-\nu}^* \\
\quad \end{if} & \\
\text{end for} & \\
\text{return } G^* & 
\end{align*}
\]
Graph reduction algorithm

\[ V^* \leftarrow \{A, Y\} \cup W \cup M \]
\[ G^* \leftarrow G(V^*) \text{ by projecting out } N \text{ and } I \text{ with latent projection} \]

for \( v \in V^* \) do
    if (\( v \in W \setminus O \) and \( v \) satisfies the \( W \)-criterion) or (\( v \in M \) and \( v \) satisfies the \( M \)-criterion) then
        \[ V^* \leftarrow V^* \setminus \{v\} \]
        \[ G^* \leftarrow G^*_{-v} \]
    end if
end for
return \( G^* \)

\( G^*_{-v} \) saturates edges from \( \text{Pa}(v) \) to \( \text{Ch}(v) \) and those within \( \text{Ch}(v) \), before removing \( v \).
Graph reduction algorithm

\[
\begin{align*}
V^* & \leftarrow \{A, Y\} \cup W \cup M \\
G^* & \leftarrow G(V^*) \text{ by projecting out } N \text{ and } I \text{ with latent projection} \\
\text{for } \nu \in V^* \text{ do} & \\
\quad \text{if } (\nu \in W \setminus O \text{ and } \nu \text{ satisfies the } W\text{-criterion}) \text{ or } (\nu \in M \text{ and } \nu \text{ satisfies the } M\text{-criterion}) \text{ then} & \\
\quad \quad V^* & \leftarrow V^* \setminus \{\nu\} \\
\quad \quad G^* & \leftarrow G^*_{-\nu} \\
\quad \text{end if} & \\
\text{end for} & \\
\text{return } G^* & \\
\end{align*}
\]

\(G^*_{-\nu}\) saturates edges from \(Pa(\nu)\) to \(Ch(\nu)\) and those within \(Ch(\nu)\), before removing \(\nu\).
Graph reduction algorithm

\[ V^* \leftarrow \{A, Y\} \cup W \cup M \]
\[ G^* \leftarrow G(V^*) \] by projecting out \( N \) and \( I \) with latent projection
\[
\text{for } v \in V^* \text{ do}
\]
\[
\text{if } (v \in W \setminus O \text{ and } v \text{ satisfies the } W\text{-criterion}) \text{ or } (v \in M \text{ and } v \text{ satisfies the } M\text{-criterion}) \text{ then}
\]
\[ V^* \leftarrow V^* \setminus \{v\} \]
\[ G^* \leftarrow G_{-v}^* \]
\[
\text{end if}
\]
\[
\text{end for}
\]
\[
\text{return } G^*
\]

\( G_{-v}^* \) saturates edges from Pa(\( v \)) to Ch(\( v \)) and those within Ch(\( v \)), before removing \( v \).

\[
\begin{array}{c}
W_i \quad W_k \\
\downarrow \quad \downarrow \\
\{W_j\} \quad \{W_j\} \\
\downarrow \quad \downarrow \\
W_{i_1} \quad W_{i_2} \quad W_{i_3} \\
\end{array}

\Rightarrow

\begin{array}{c}
W_i \quad W_k \\
\downarrow \quad \downarrow \\
\{W_j\} \quad \{W_j\} \\
\downarrow \quad \downarrow \\
W_{i_1} \quad W_{i_2} \quad W_{i_3} \\
\end{array}
Graph reduction algorithm

\[ V^* \leftarrow \{A, Y\} \cup W \cup M \]
\[ G^* \leftarrow G(V^*) \text{ by projecting out } N \text{ and } I \text{ with latent projection} \]

\[ \text{for } v \in V^* \text{ do} \]
\[ \quad \text{if } (v \in W \setminus O \text{ and } v \text{ satisfies the } W\text{-criterion}) \text{ or } (v \in M \text{ and } v \text{ satisfies the } M\text{-criterion}) \text{ then} \]
\[ \quad \quad V^* \leftarrow V^* \setminus \{v\} \]
\[ \quad \quad G^* \leftarrow G_{-v} \]
\[ \quad \text{end if} \]
\[ \text{end for} \]
\[ \text{return } G^* \]

\[ G_{-v} \] saturates edges from Pa(\(v\)) to Ch(\(v\)) and those within Ch(\(v\)), before removing \(v\).
Graph reduction algorithm

$$V^* \leftarrow \{A, Y\} \cup W \cup M$$

$$G^* \leftarrow G(V^*)$$ by projecting out $N$ and $I$ with latent projection

for $v \in V^*$ do

    if ($v \in W \setminus O$ and $v$ satisfies the $W$-criterion) or ($v \in M$ and $v$ satisfies the $M$-criterion) then

        $V^* \leftarrow V^* \setminus \{v\}$
        $G^* \leftarrow G_{-v}$

    end if

end for

return $G^*$

$G_{-v}$ saturates edges from $Pa(v)$ to $Ch(v)$ and those within $Ch(v)$, before removing $v$.
Graph reduction algorithm

\[ V^* \leftarrow \{A, Y\} \cup W \cup M \]
\[ G^* \leftarrow G(V^*) \text{ by projecting out } N \text{ and } I \text{ with latent projection} \]
for \( \nu \in V^* \) do
  if (\( \nu \in W \setminus O \) and \( \nu \) satisfies the \( W \)-criterion) or (\( \nu \in M \) and \( \nu \) satisfies the \( M \)-criterion) then
    \[ V^* \leftarrow V^* \setminus \{\nu\} \]
    \[ G^* \leftarrow G_{-\nu} \]
  end if
end for
return \( G^* \)

\( G_{-\nu} \) saturates edges from \( \text{Pa}(\nu) \) to \( \text{Ch}(\nu) \) and those within \( \text{Ch}(\nu) \), before removing \( \nu \).
Graph reduction algorithm

Theorem

The reduced graph $G'$ is a DAG on vertices $V'$ with the following properties.

1. $G'$ does not depend on the order that vertices are visited in the Algorithm.
2. $M(G, V') = M(G', V')$. (when uninformative vars are continuous)
3. $a(P; G) = a(P; G')$ for every $P \in M(G, V)$.
4. $e_P(V, G) = e_P(V', G')$ $P$-a.e. for every $P \in M(G, V)$.
Graph reduction algorithm

Theorem
The reduced graph $G^\ast$ is a DAG on vertices $V^\ast = \{V \ast (G)\}$ with the following properties.

1. $G^\ast$ does not depend on the order that vertices are visited in the algorithm.
2. $M(G, V^\ast) = M(G^\ast, V^\ast)$. (when uninformative vars are continuous)
3. $a(P; G) = a(P; G^\ast)$ for every $P \in M(G, V)$.
4. $e_P(V, G) = e_P(V, G^\ast)$ $P$-a.e. for every $P \in M(G, V)$.
Theorem

The reduced graph $G^\ast$ is a DAG on vertices $V^\ast$ with the following properties.

1. $G^\ast$ does not depend on the order that vertices are visited in the algorithm.

2. $M(G, V^\ast) = M(G^\ast, V^\ast)$ (when uninformative vars are continuous) (based on mDAGs of Evans, 2016)

3. $a(P; G) = a(P; G^\ast)$ for every $P \in M(G, V)$.

4. $e(\cdot), P(V, G) = e(\cdot), P(V^\ast, G^\ast)$ $P$-a.e. for every $P \in M(G, V)$.
Theorem
The reduced graph $G^*$ is a DAG on vertices $V^*$ with the following properties.

1. $G^*$ does not depend on the order that vertices are visited in the algorithm.
2. $M(G, V^*) = M(G^*, V^*)$. (when uninformative vars are continuous)
3. $a(P; G) = a(P; G^*)$ for every $P \in M(G, V)$.
4. $\mathbb{E}(P(V, G)) = \mathbb{E}(P(V, G^*))$ $P$-a.e. for every $P \in M(G, V)$.
Theorem The reduced graph $G^*$ is a DAG on vertices $V^* \equiv V^*(G)$ with the following properties.
Theorem The reduced graph $G^*$ is a DAG on vertices $V^* \equiv V^*(G)$ with the following properties.

1. $G^*$ does not depend on the order that vertices are visited in the Algorithm.
Graph reduction algorithm

Theorem The reduced graph $G^*$ is a DAG on vertices $V^* \equiv V^*(G)$ with the following properties.

1. $G^*$ does not depend on the order that vertices are visited in the Algorithm.

2. $\mathcal{M}(G, V^*) = \mathcal{M}(G^*, V^*)$. (based on mDAGs of Evans, 2016)
Theorem The reduced graph $G^*$ is a DAG on vertices $V^* \equiv V^*(G)$ with the following properties.

1. $G^*$ does not depend on the order that vertices are visited in the Algorithm.
2. $\mathcal{M}(G, V^*) = \mathcal{M}(G^*, V^*)$. (\(\text{\ding{118}}\) when uninformative vars are continuous)
   (\(\text{\ding{118}}\) based on mDAGs of Evans, 2016)
3. $\psi_a(P; G) = \psi_a(P; G^*)$ for every $P \in \mathcal{M}(G, V)$. 

\[G\]

\[G^*\]
**Theorem** The reduced graph $\mathcal{G}^*$ is a DAG on vertices $\mathbf{V}^* \equiv \mathbf{V}^*(\mathcal{G})$ with the following properties.

1. $\mathcal{G}^*$ does not depend on the order that vertices are visited in the Algorithm.

2. $\mathcal{M}(\mathcal{G}, \mathbf{V}^*) = \mathcal{M}(\mathcal{G}^*, \mathbf{V}^*)$. (\(\uparrow\) when uninformative vars are continuous) (\(\uparrow\) based on mDAGs of Evans, 2016)

3. $\Psi_\alpha(P; \mathcal{G}) = \Psi_\alpha(P; \mathcal{G}^*)$ for every $P \in \mathcal{M}(\mathcal{G}, \mathbf{V})$.

4. $\chi_{\text{eff}, P}(\mathbf{V}, \mathcal{G}) = \chi_{\text{eff}, P}(\mathbf{V}^*, \mathcal{G}^*)$ $P$-a.e. for every $P \in \mathcal{M}(\mathcal{G}, \mathbf{V})$. 

Corollary

A \( a(P; G) \) is the irreducible "efficient" g-formula in the sense that

\[ | a(P_n; G) - a(P_n; G^\ast) | = o(n^{-1/2}) \text{ as } n \to \infty \]

under iid sampling of \( P \) with initial samples space.

A Y O

W_2 W_3 W_4

G a(P; G) =

X y, o, i, w_1, w_2, w_3, w_4

yp(y | A = a, o) \cdot p(i | w_1) \cdot p(o | w_1) \cdot \\
\cdot p(w_1 | w_2, w_3) \cdot p(w_3 | w_4) \cdot p(w_2) \cdot p(w_4).
Corollary \( \Psi_a(P; G^*) \) is the irreducible “efficient” g-formula in the sense that

\[
|\Psi_a(P_n; G^*) - \Psi_a(P_n; G)| = o_p(n^{-1/2}) \quad \text{as } n \to \infty
\]

under iid sampling of \( P \in \mathcal{M}(G, V) \) with a finite sample space.
**Corollary** $\Psi_a(P; G^*)$ is the irreducible “efficient” g-formula in the sense that

$$|\Psi_a(P_n; G^*) - \Psi_a(P_n; G)| = o_p(n^{-1/2})$$

as $n \to \infty$

under iid sampling of $P \in \mathcal{M}(G, \mathcal{V})$ with a finite sample space.

$$\Psi_a(P; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} yp(y | A = a, o)p(i | w_1)p(o | w_1)$$

$$\times p(w_1 | w_2, w_3)p(w_3 | w_4)p(w_2)p(w_4).$$
**Corollary** \( \Psi_a(P; G^*) \) is the irreducible “efficient” g-formula in the sense that

\[
|\Psi_a(P_n; G^*) - \Psi_a(P_n; G)| = o_p(n^{-1/2}) \quad \text{as} \quad n \to \infty
\]

under iid sampling of \( P \in \mathcal{M}(G, \mathcal{V}) \) with a finite sample space.

\[
\Psi_a(P; G) = \sum_{y, o, i, w_1, w_2, w_3, w_4} yp(y \mid A = a, o)p(i \mid w_1)p(o \mid w_1) \\
\times p(w_1 \mid w_2, w_3)p(w_3 \mid w_4)p(w_2)p(w_4).
\]

\[
\Psi_a(P; G^*) = \sum_{y, o, w_2, w_3} yP[y \mid A = a, o] \\
\times P(o \mid w_2, w_3)P(w_2)P(w_3).
\]
Examples (i)

\[ G = G^* \]
The $g$-formula

$$\Psi_a(P; G) = \sum_{m,o} \mathbb{E}[Y | m, o] p(o) p(m | a)$$

is efficient.
Examples (i)

The g-formula

$$\psi_a(P; G) = \sum_{m, o} \mathbb{E}[Y \mid m, o] p(o) p(m \mid a)$$ is efficient.

Neither the back-door

$$\psi_a^{\text{ADJ}}(P; G) = \sum_o \mathbb{E}[Y \mid a, o] p(o)$$

nor the front-door

$$\psi_a^{\text{FRONT}}(P; G) = \sum_m \left\{ \sum_a \mathbb{E}[Y \mid m, a'] p(a') \right\} p(m \mid a)$$

is efficient.
Examples (ii.a)

\[
a(P; G_1) = \mathbb{E}[Y | A = a, O = o] \cdot P(o).
\]
Examples (ii.a)

\[ a(P; G^{\uparrow 1}) = X_o E[Y | A = a, O = o] P(o). \]
\[
\Psi_a(P; G_1^*) = \sum_o \mathbb{E}[Y | A = a, O = o]P(o).
\]
Examples (ii.b)
Examples (ii.b)

\[
\Psi_a(P; G_2^*) = \sum_{M} \mathbb{E}[Y \mid A = a, M] \sum_{O} P(M \mid O, a)P(O).
\]

\[G_2^* = G_2\]
Examples (ii.c)

\[ a \left( P, G^{\etr} \right) = \mathbb{E} \left[ Y \middle| A = a \right]. \]
Examples (ii.c)
Examples (ii.c)

$$\psi_a(P; G_3^*) = \mathbb{E}[Y \mid A = a].$$
Examples (iii)
Examples (iii)

\[
G(a(P; G) = X \quad M_1 \quad E[Y | M_1] \quad X \quad O_1, O_2 \quad P(O_1) \quad P(O_2). \quad M_1 \quad (M_2) \quad (M_3) \quad Y
\]

\[
G'(a) = A \quad (\tilde{I}) \quad O_1 \quad O_2 \quad \Rightarrow \quad A \quad M_1 \quad (M_2) \quad (M_3) \quad Y
\]
Examples (iii)

\[ P(X|M_1) = P(Y|M_1) \]

\[ G \]

\[ \Rightarrow \]

\[ \Rightarrow \]

\[ (1) \]

\[ O_1 \quad O_2 \]

\[ A \quad M_1 \quad M_2 \quad M_3 \quad Y \]

\[ O_1 \quad O_2 \]

\[ A \quad M_1 \quad M_2 \quad M_3 \quad Y \]
$$\Psi_a(P; G^*) = \sum_{M_1} E[Y | M_1] \sum_{O_1, O_2} P(M_1 | O_1, O_2, A = a) P(O_1) P(O_2).$$
Examples (iv)
Try simplifying your causal DAG with R package **reduceDAG** available from https://unbiased.co.in

```r
library(dagitty)
library(reduceDAG)

# Define the DAG

# Exposure
# A
# Mediator
# M
# Outcome
# Y
# Observed
# O

g <- dagitty('dag {
  A [pos="0,2", exposure]
  M [pos="1,1"]
  Y [pos="2,2", outcome]
  O [pos="1,0"]
  A -> M -> Y
  A -> Y
  O -> M
}
)

cat(gFormula(g))
# sum_{M,Y} Y P(Y \mid A=a,M) \sum_{O} P(M \mid A=a,O) P(O)

h <- reduceDAG(g, verbose=TRUE)
# Uninformative variables {M} are eliminated.
# Reduced g-formula:
# sum_{O,Y} Y P(Y \mid A=a,O) P(O)
```
We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.
We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.

- For some graphs, certain variables are uninformative for optimal estimation in large samples.
We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.

- For some graphs, certain variables are uninformative for optimal estimation in large samples.
- We graphically characterized the set of irreducible informative variables $V^*$. 

R package `reduceDAG`. 

32
We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.

- For some graphs, certain variables are uninformative for optimal estimation in large samples.
- We graphically characterized the set of irreducible informative variables $V^*$.
- The marginal model over $V^*$ is represented by a DAG $G^*$.
  - A polynomial time algorithm for constructing $G^*$.
We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.

- For some graphs, certain variables are uninformative for optimal estimation in large samples.
- We graphically characterized the set of irreducible informative variables $V^*$.
- The marginal model over $V^*$ is represented by a DAG $G^*$.
  - A polynomial time algorithm for constructing $G^*$.
- For optimal estimation, $G^*$ is all you need.
  - $G^*$ prescribes the simplest g-formula that is efficient.
  - $G^*$ could inform data collection and estimation strategies.
Conclusion

We have studied estimating the counterfactual mean (or the average treatment effect) of a point intervention given a causal DAG.

- For some graphs, certain variables are uninformative for optimal estimation in large samples.
- We graphically characterized the set of irreducible informative variables $V^*$. 
- The marginal model over $V^*$ is represented by a DAG $G^*$. 
  - A polynomial time algorithm for constructing $G^*$.
- For optimal estimation, $G^*$ is all you need. 
  - $G^*$ prescribes the simplest g-formula that is efficient.
  - $G^*$ could inform data collection and estimation strategies.
- R package `reduceDAG`. 
Thanks!

arXiv: 2202.11994
R package: reduceDAG
References


Guo, F. Richard and Emilija Perković (2021). “Minimal enumeration of all possible total effects in a Markov equivalence class”. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 


**W-criterion** Suppose \( \{ W_j \} \cup \text{Ch}(W_j) \cap W \) is topologically sorted as \( \{ W_{j_0} \equiv W_j, W_{j_1}, \ldots, W_{j_r} \} \). Then \( W_j \in W \setminus O \) is uninformative if and only if

1. \( W_j \perp_{O \mid O \setminus W_{j_r}, \text{Pa}(W_{j_r}) \setminus \{ W_j \}} \),
2. and for \( m = 1, \ldots, r \):
   (i) \( W_{j_{m-1}} \rightarrow W_{j_m} \) (children are chained)
   (ii) \( \text{Pa}(W_{j_m}) \subseteq \text{Pa}(W_{j_{m-1}}) \cup \{ W_{j_{m-1}} \} \) (parent sets are decreasing)
   (iii) \( \text{Pa}(W_{j_{m-1}}) \setminus \text{Pa}(W_{j_m}) \perp_{O \mid O \setminus \text{Pa}(W_{j_m})} \) (left-over piece is separated from \( O \))

**M-criterion** Suppose \( \{ M_i \} \cup \text{Ch}(M_i) \cap M \) is topologically sorted as \( \{ M_{i_0} \equiv M_i, M_{i_1}, \ldots, M_{i_k} \} \). Then \( M_i \in M \) is uninformative if and only if

1. \( M_i \perp_{\{ A, Y \} \cup O_{\text{min}} \mid M_{i_k}, \text{Pa}(M_{i_k}) \setminus \{ M_i \}} \),
2. and for \( l = 1, \ldots, k \):
   (i) \( M_{i_{l-1}} \rightarrow M_i \) (children are chained)
   (ii) \( \text{Pa}(M_{i_l}) \subseteq \text{Pa}(M_{i_{l-1}}) \cup \{ M_{i_{l-1}} \} \) (parent sets are decreasing)
   (iii) \( \text{Pa}(M_{i_{l-1}}) \setminus \text{Pa}(M_{i_l}) \perp_{\{ A, Y \} \cup O_{\text{min}} \mid \text{Pa}(M_{i_l})} \) (left-over piece is separated from \( A, Y, O_{\text{min}} \))
**Nonparametric model** $\mathcal{M}_0(\mathbf{V}) \equiv \{\text{all laws over vector } \mathbf{V}\}$.

**Identifying formula** Fix a model $\mathcal{M}(\mathbf{V}) \subseteq \mathcal{M}_0(\mathbf{V})$ and a functional $\gamma(P) : \mathcal{M}(\mathbf{V}) \to \mathbb{R}$.

Functional $\chi(P) : \mathcal{M}_0(\mathbf{V}) \to \mathbb{R}$ is an identifying formula for $\gamma(P)$ if $\chi(P) = \gamma(P)$ for every $P \in \mathcal{M}(\mathbf{V})$.

**Efficient identifying formula** Consider a semiparametric model $\mathcal{M}(\mathbf{V}) \subseteq \mathcal{M}_0(\mathbf{V})$ and a regular functional $\gamma : \mathcal{M}(\mathbf{V}) \to \mathbb{R}$. Let $\gamma_{P,\text{eff}}(\mathbf{V})$ be its efficient influence function with respect to $\mathcal{M}(\mathbf{V})$.

An identifying formula $\chi : \mathcal{M}_0(\mathbf{V}) \to \mathbb{R}$ for functional $\gamma$ is called efficient if $\chi_{P,NP}(\mathbf{V}) = \gamma_{P,\text{eff}}(\mathbf{V})$ $P$-almost-everywhere for every $P \in \mathcal{M}(\mathbf{V})$. 