Causal effects in MPDAGs: Identification and efficient estimation

Emilija Perković
joint work with F. Richard Guo

Department of Statistics, University of Washington

Goal

- Estimate the total causal effect of X_A on X_Y

Observational data

Randomized control studies

Goal

- Estimate the total causal effect of X_A on X_Y
  - the change in X_Y due to $do(x_a)$-
  from observational data.

- $do(x_a)$: an intervention that sets variables X_A to x_a.

Observational data

Randomized control studies

Goal

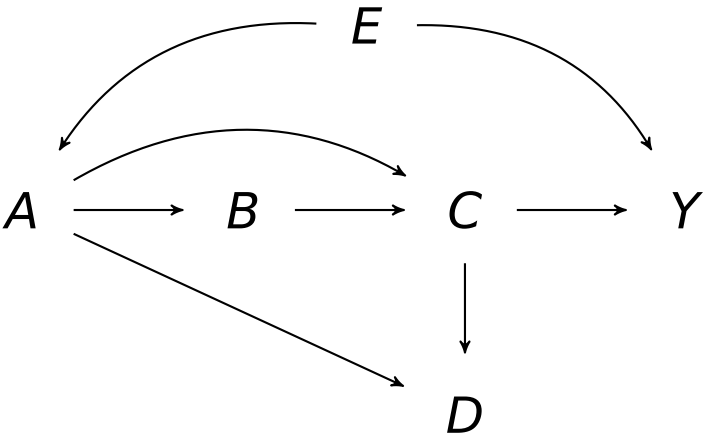
- Estimate the total causal effect of X_A on X_Y
  - the change in X_Y due to $do(x_a)$-
  from observational data.

- $do(x_a)$: an intervention that sets variables X_A to x_a.
  $f(x_y|do(x_a)) \neq f(x_y|x_a)$.

Observational data

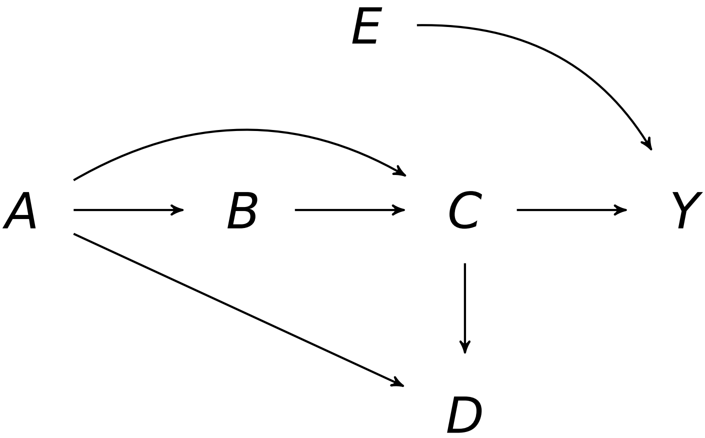
Randomized control studies

Observational Causal DAG



Causal Directed Acyclic Graph (DAG) $\mathcal{D}$.

Interventional Causal DAG

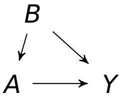


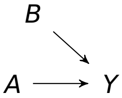
Causal DAG $\mathcal{D}$ **after a "do"-intervention on X_A.**

DAGs and linear SCMs

- $do(x_a)$: an intervention that sets variables X_A to x_a.
- Observational density $f(x_{\mathbf{v}})$, Interventional density $f(x_{\mathbf{v}}|do(x_a))$.
- A DAG $\mathcal{D}$ is causal if for all observational and interventional densities:

$$f(x_{\mathbf{v}}) = \prod_{J \in V} f(x_j|x_{pa(j)}) \quad \text{and} \quad f(x_{\mathbf{v}}|do(x_a)) = \prod_{J \in V \setminus \{A\}} f(x_j|x_{pa(j)})$$





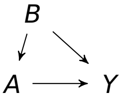
$$f(x_b, x_a, x_y) = f(x_y|x_b, x_a) f(x_a|x_b) f(x_b)$$

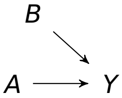
$$f(x_b, x_y|do(x_a)) = f(x_y|x_b, x_a) f(x_b)$$

DAGs and linear SCMs

- $do(x_a)$: an intervention that sets variables X_A to x_a.
- Observational density $f(x_\mathbf{v})$, Interventional density $f(x_\mathbf{v}|do(x_a))$.
- A DAG $\mathcal{D}$ is causal if for all observational and interventional densities:

$$f(x_\mathbf{v}) = \prod_{J \in V} f(x_j | x_{pa(j)}) \quad \text{and} \quad f(x_\mathbf{v}|do(x_a)) = \prod_{J \in V \setminus \{A\}} f(x_j | x_{pa(j)})$$





$$f(x_b, x_a, x_y) = f(x_y|x_b, x_a) f(x_a|x_b) f(x_b) \qquad f(x_b, x_y|do(x_a)) = f(x_y|x_b, x_a) f(x_b)$$

- We also assume that the data is generated by a linear causal model:

$$\begin{aligned} X_B &\leftarrow \epsilon_B \\ X_A &\leftarrow \gamma_{ba} X_B + \epsilon_A \\ X_Y &\leftarrow \gamma_{ay} X_A + \gamma_{by} X_B + \epsilon_Y \end{aligned} \qquad \begin{aligned} X_B &\leftarrow \epsilon_B \\ X_A &\leftarrow x_a \\ X_Y &\leftarrow \gamma_{ay} x_a + \gamma_{by} X_B + \epsilon_Y, \end{aligned}$$

- where for $U \in \mathbf{V}$, $\mathbb{E}\, \epsilon_U = 0, \quad 0 < \mathrm{var}\, \epsilon_U < \infty, \quad \epsilon_U$ are mutually independent.

How to define a causal effect?

Total causal effect

- For simplicity $\mathbf{A} = \{A\}, \mathbf{Y} = \{Y\}$ for the rest of this talk.
- Total causal effect, τ_{AY}:

$$\tau_{AY} = E[X_Y | do(X_A = x_a + 1)] - E[X_Y | do(X_A = x_a)] = \frac{\partial}{\partial x_a} E[X_Y | do(x_a)],$$

Identifiability

- A total causal effect is identifiable from observational data if

$f(x_y | do(x_a))$ can be expressed as a function of $f(x_{\mathbf{v}})$.

How to define a causal effect?

Total causal effect

- For simplicity $\mathbf{A} = \{A\}, \mathbf{Y} = \{Y\}$ for the rest of this talk.
- Total causal effect, τ_{AY}:

$$\tau_{AY} = E[X_Y|do(X_A = x_a + 1)] - E[X_Y|do(X_A = x_a)] = \frac{\partial}{\partial x_a} E[X_Y|do(x_a)],$$

Identifiability

- A total causal effect is identifiable from observational data if

  $f(x_y|do(x_a))$ can be expressed as a function of $f(x_{\mathbf{V}})$.

- Given the causal DAG, every total causal effect is identifiable.

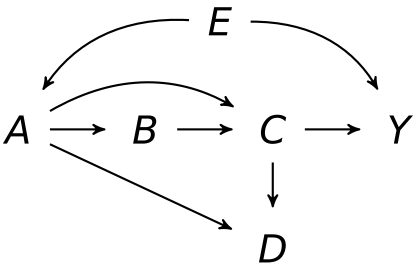
Truncated Factorization, G-formula (Robins '86, Pearl '93, Spirtes '93): $\mathbf{V}' = \mathbf{V} \setminus \{A, Y\}$,

$$f(x_y|do(x_y)) = \int \prod_{I \in \mathbf{V} \setminus \{A\}} f(x_i|x_{pa(i)}) dx_{\mathbf{v}'}.$$

Adjustment (Pearl '93, Shpitser et al '10): $\mathbf{Z}$ is an adjustment set if

$$f(x_y|do(x_a)) = \int f(x_y|x_a, x_{\mathbf{z}}) f(x_{\mathbf{z}}) dx_{\mathbf{z}}$$

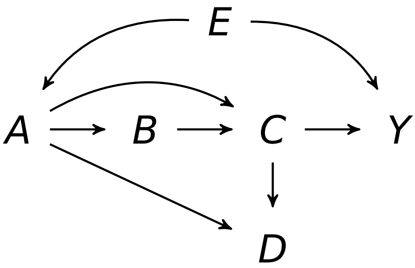
Causal DAG, linear SCM



- Data is generated by:

$$X = \Gamma^\intercal X + \epsilon, \qquad \Gamma = (\gamma_{ij}), \quad I \not\to J \Rightarrow \gamma_{ij} = 0,$$
$$\mathbb{E}\,\epsilon = 0, \quad 0 < \text{var}\,\epsilon_i < \infty, \quad \epsilon_i \text{ are mutually independent.}$$

Causal DAG, linear SCM

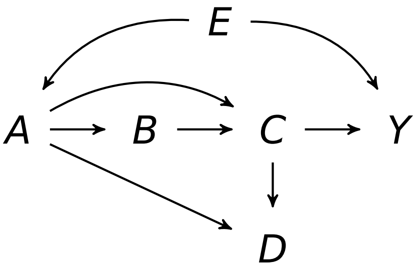


- Data is generated by:

$$X = \Gamma^{\intercal} X + \epsilon, \qquad \Gamma = (\gamma_{ij}), \quad I \not\to J \Rightarrow \gamma_{ij} = 0,$$
$$\mathbb{E}\,\epsilon = 0, \quad 0 < \text{var}\,\epsilon_i < \infty, \quad \epsilon_i \text{ are mutually independent.}$$

- Suppose we are interested in τ_{AY}.

Causal DAG, linear SCM

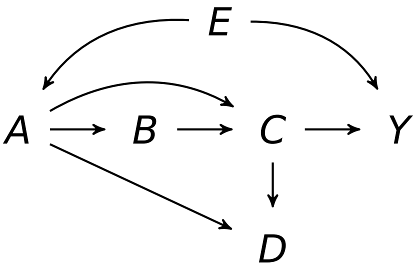


- By the path tracing rules (Wright, 1934)

$$\tau_{AY} =$$

$$= \cdots = \gamma_{cy}(\gamma_{bc}\gamma_{ab} + \gamma_{ac}),$$

Causal DAG, linear SCM

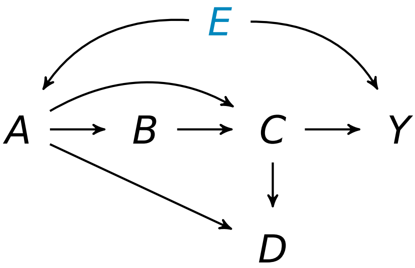


- By the path tracing rules (Wright, 1934) and the g-formula:

$$\begin{aligned}\tau_{AY} &= \frac{\partial}{\partial x_a}\mathbb{E}[X_Y|\text{do}(x_a)] \\ &= \frac{\partial}{\partial x_a}\int \mathbb{E}[X_Y|x_c, x_e]f(x_c|x_a, x_b)f(x_b|x_a)f(x_e)dx_b dx_c dx_e \\ &= \cdots = \gamma_{cy}(\gamma_{bc}\gamma_{ab} + \gamma_{ac}),\end{aligned}$$

- Suggests a **plug-in estimator** - a sum-product of elements of $\widehat{\Gamma}$. Elements of estimated with least squares e.g., γ_{cy}, γ_{ey} from $X_Y \sim X_C + X_E$.

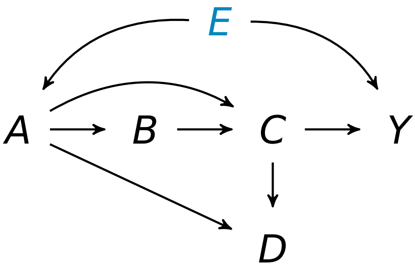
Causal DAG, linear SCM



- Additionally, since $\{E\}$ is an adjustment set

$$\tau_{AY} = \frac{\partial}{\partial x_a}\mathbb{E}[Y|\text{do}(x_a)] = \frac{\partial}{\partial x_a}\int E[X_Y|x_a, x_e]f(x_e)dx_e,$$

Causal DAG, linear SCM

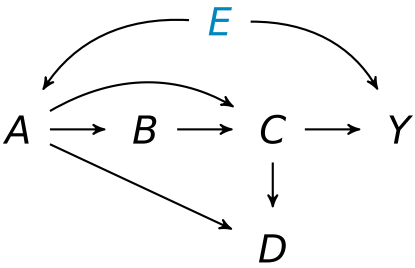


- Additionally, since {E} is an adjustment set

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[Y|\text{do}(x_a)] = \frac{\partial}{\partial x_a} \int E[X_Y|x_a, x_e] f(x_e) dx_e,$$

- Suggests estimating τ_{AY} as the least squares coefficient in $X_Y \sim X_A + X_E$.

Causal DAG, linear SCM

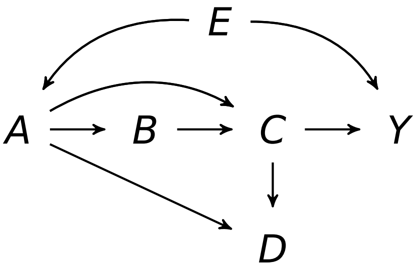


- Additionally, since {E} is an adjustment set

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[Y|\text{do}(x_a)] = \frac{\partial}{\partial x_a} \int E[X_Y|x_a, x_e] f(x_e) dx_e,$$

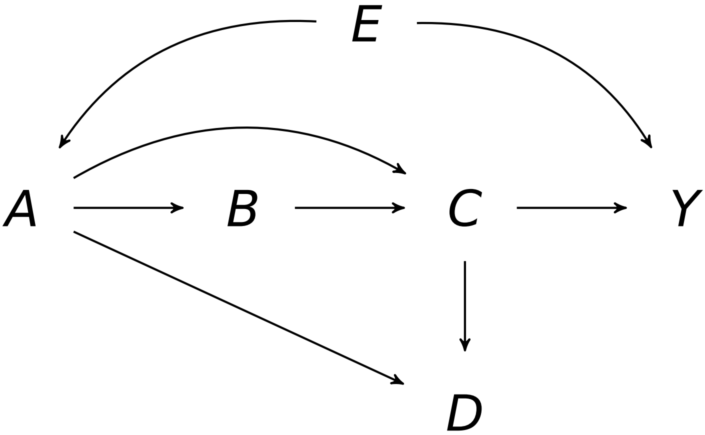
- Suggests estimating τ_{AY} as the least squares coefficient in $X_Y \sim X_A + X_E$.
- **Which estimator is more efficient?**
  And what if we do not know the causal DAG?

Existing Results



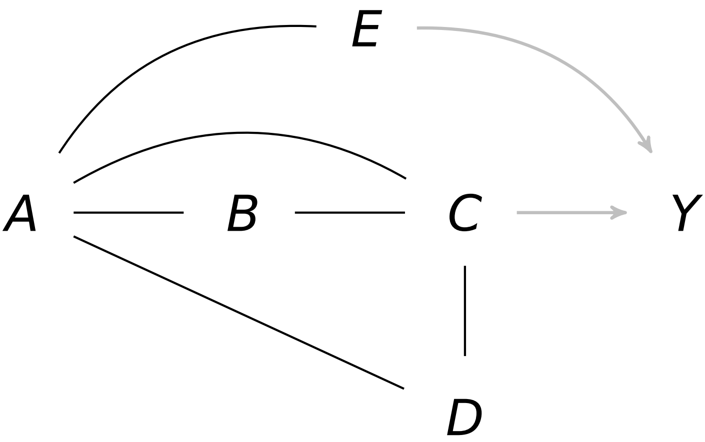
- **Which estimator is more efficient?**
  - Assuming Gaussian errors and given a particular DAG, Hayashi and Kuroki (2014) show that the path tracing plug-in estimator is **more efficient** than covariate adjustment.
  - The path tracing based estimator is the plug-in MLE.
- **What if we do not have the DAG?**

What if we do not know the DAG?



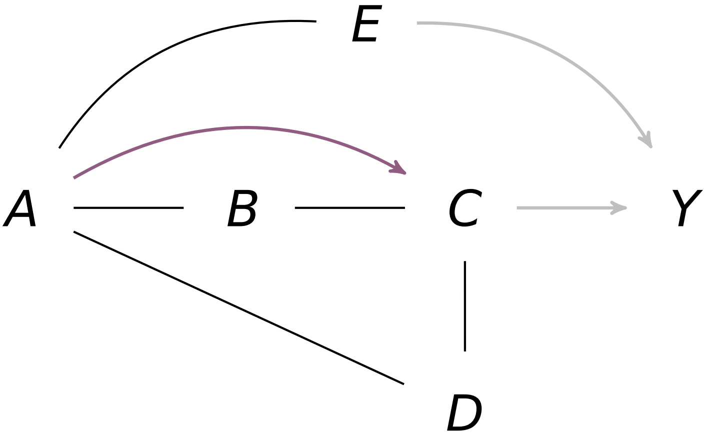
Causal Directed Acyclic Graph (DAG) $\mathcal{D}$.

What if we do not have the DAG?



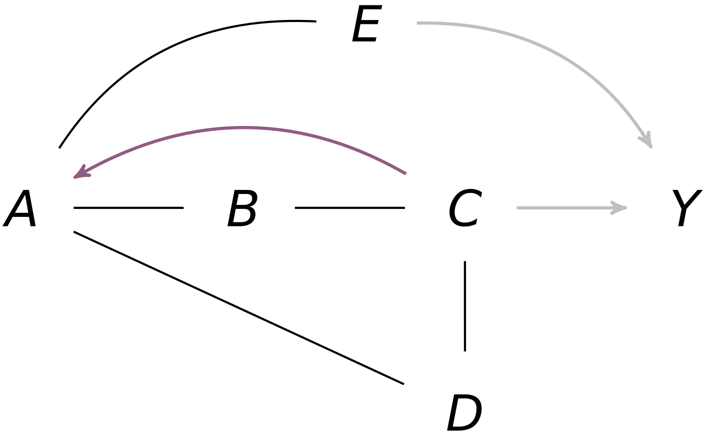
Completed Partially Directed Acyclic Graph (CPDAG).

What if we do not have the DAG?



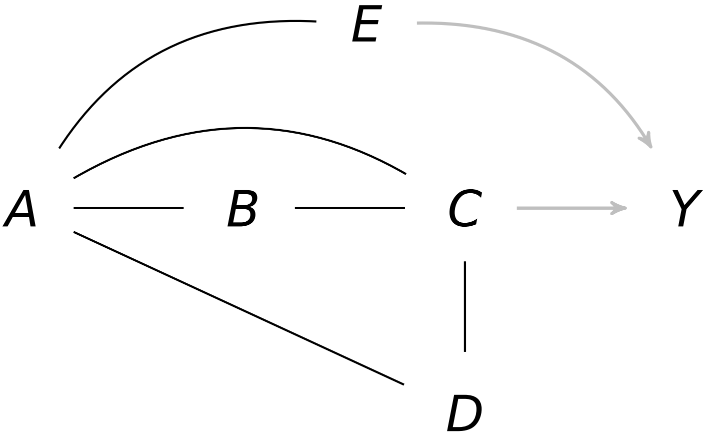
Completed Partially Directed Acyclic Graph (CPDAG).

What if we do not have the DAG?



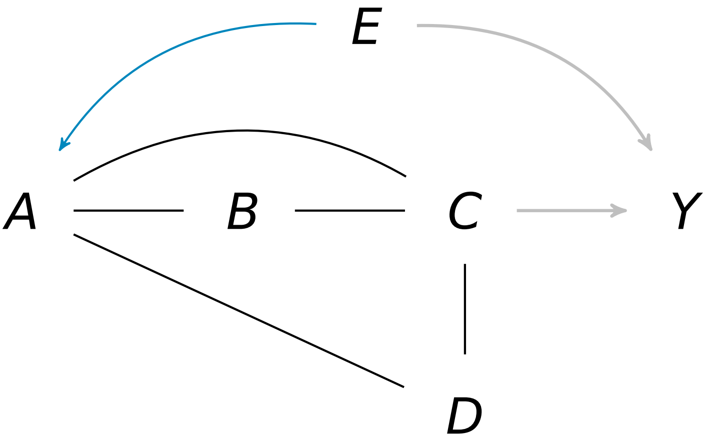
Completed Partially Directed Acyclic Graph (CPDAG).

What if we do not have the DAG?



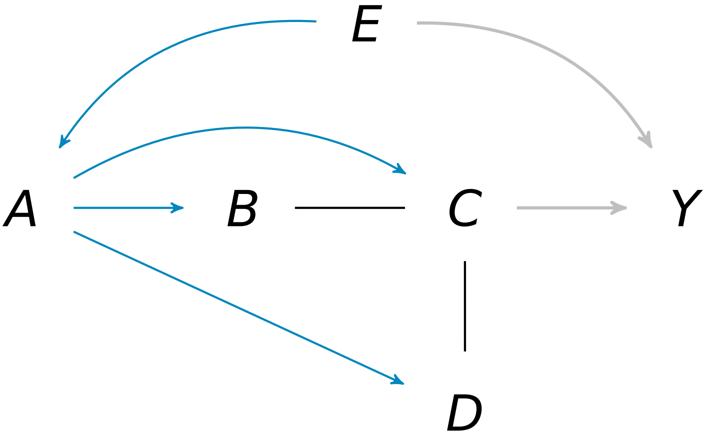
Completed Partially Directed Acyclic Graph (CPDAG).

What if we do not have the DAG?



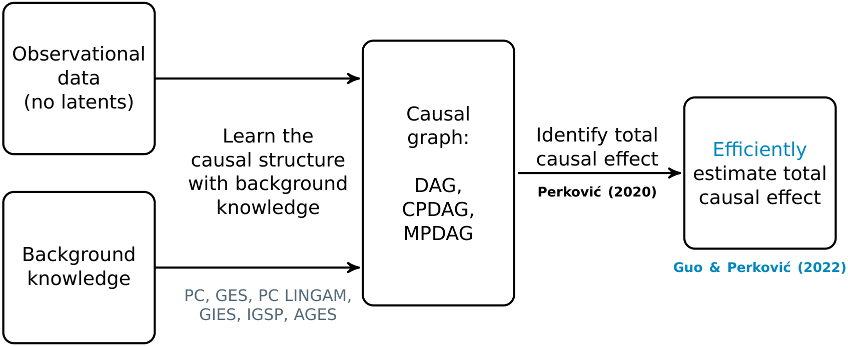
Partially Directed Acyclic Graph (PDAG).

What if we do not have the DAG?



Maximally oriented Partially Directed Acyclic Graph (MPDAG).

Framework



- PC (Spirtes et al, 1993), GES (Chickering, 2002) + Adding background knowledge (Meek, 1995; TETRAD, Scheines et al., 1998), PC LINGAM (Hoyer et al., 2008), GIES (Hauser and Bühlmann, 2012), IGSP (Wang et al., 2017), etc.
- Other framing: start with a DAG and remove some directional information while keeping the orientations closed under Meek orientation rules (Meek, 1995).

Existing Results

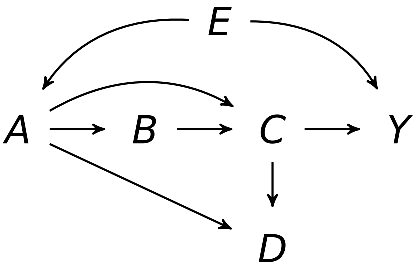
Graphical criterion	DAG	CPDAG	MPDAG
Adjustment (Pearl ’93, Shpitser et al ’10, Perković et al ’15, ’17, ’18)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
G-formula, Truncated Factorization (Robins ’86, Pearl ’93)	$\Leftrightarrow$		
Causal identification formula (Perković ’20)	$\Leftrightarrow$	$\Leftrightarrow$	$\Leftrightarrow$

Existing Results

Graphical criterion	DAG	CPDAG	MPDAG
Adjustment (Pearl '93, Shpitser et al '10, Perković et al '15, '17, '18)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
G-formula, Truncated Factorization (Robins '86, Pearl '93)	$\Leftrightarrow$		
Causal identification formula (Perković '20)	$\Leftrightarrow$	$\Leftrightarrow$	$\Leftrightarrow$

- Henckel et al (2022), Witte et al, (2020), Rotnitzky and Smucler (2020) graphically characterize an **optimal covariate adjustment set** in DAGs, CPDAGs, and MPDAGs.
- However, covariate adjustment is not complete for estimating all identifiable causal effects.
- Can we leverage the **causal identification formula** for a more efficient estimator in CPDAGs and MPDAGs?

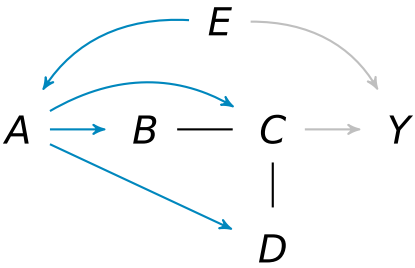
Block-recursive reparametrization



- Data is generated by

$$X = \Gamma^{\intercal} X + \epsilon, \qquad \Gamma = (\gamma_{ij}), \quad I \not\to J \Rightarrow \gamma_{ij} = 0,$$
$$\mathbb{E}\,\epsilon = 0, \quad 0 < \operatorname{var} \epsilon_I < \infty, \quad \epsilon_I \text{ are mutually independent.}$$

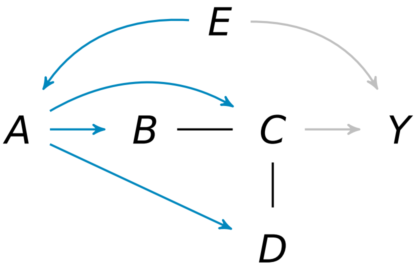
Block-recursive reparametrization



- Data is generated by

$$X = \Gamma^\intercal X + \epsilon, \qquad \Gamma = (\gamma_{ij}), \quad I \not\to J \Rightarrow \gamma_{ij} = 0,$$
$$\mathbb{E}\,\epsilon = 0, \quad 0 < \operatorname{var} \epsilon_I < \infty, \quad \epsilon_I \text{ are mutually independent.}$$

Block-recursive reparametrization

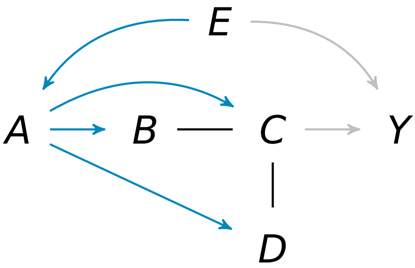


- Data is generated by

$$X = \Gamma^{\intercal} X + \epsilon, \qquad \Gamma = (\gamma_{ij}), \quad I \not\to J \Rightarrow \gamma_{ij} = 0,$$
$$\mathbb{E}\,\epsilon = 0, \quad 0 < \operatorname{var} \epsilon_I < \infty, \quad \epsilon_I \text{ are mutually independent.}$$

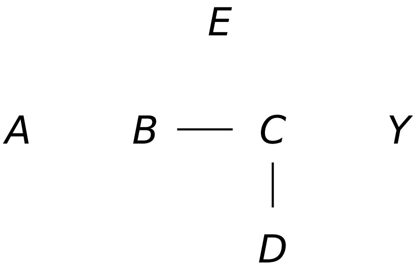
- Γ is **not identifiable**.

Block-recursive reparametrization



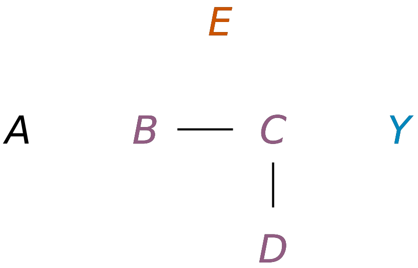
- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

Block-recursive reparametrization



- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

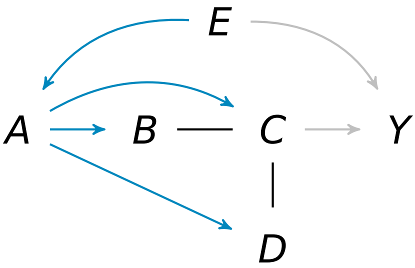
Block-recursive reparametrization



- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

Block-recursive reparametrization

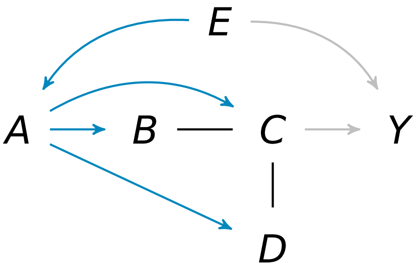


- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

1. The "between bucket" causal effects are **identifiable**. (Perković 2020).

Block-recursive reparametrization

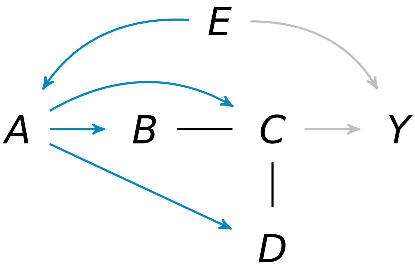


- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

1. The "between bucket" causal effects are **identifiable**. (Perković 2020).
2. **Restrictive property:** Each node in a bucket has the same out-of-bucket parents (Guo and Perković, 2022).

Block-recursive reparametrization



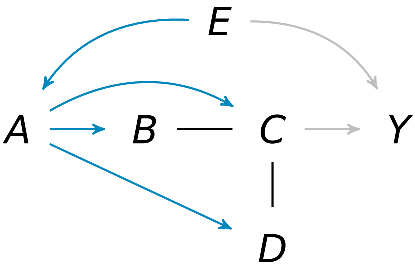
- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

1. The “between bucket” causal effects are **identifiable**. (Perković 2020).
2. **Restrictive property:** Each node in a bucket has the same out-of-bucket parents (Guo and Perković, 2022).

- We use this to reparametrize the SCM.

Block-recursive reparametrization



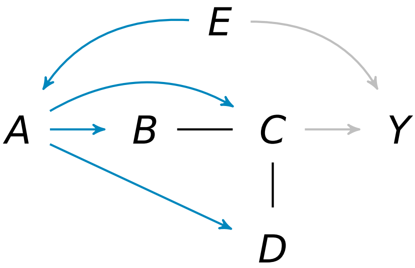
- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

$$X_{\mathbf{B_i}} = \Gamma^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \Gamma^{\intercal}_{\mathbf{B_i}} X_{\mathbf{B_i}} + \epsilon_{\mathbf{B_i}},$$

,

Block-recursive reparametrization



- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

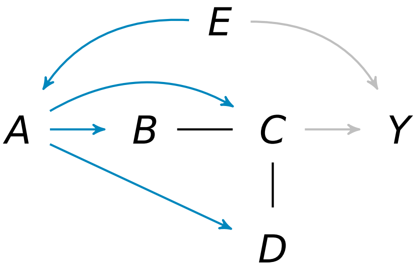
$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

$$X_{\mathbf{B_i}} = \Gamma^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \Gamma^{\intercal}_{\mathbf{B_i}} X_{\mathbf{B_i}} + \epsilon_{\mathbf{B_i}},$$

$$X_{\mathbf{B_i}} = \left(I - \Gamma_{\mathbf{B_i}}\right)^{-\intercal} \Gamma^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \left(I - \Gamma_{\mathbf{B_i}}\right)^{-\intercal} \epsilon_{\mathbf{B_i}}$$

,

Block-recursive reparametrization



- Consider **buckets** (maximal undirected connected components) in MPDAG $\mathcal{G}$:

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

$$X_{\mathbf{B_i}} = \Gamma^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \Gamma^{\intercal}_{\mathbf{B_i}} X_{\mathbf{B_i}} + \epsilon_{\mathbf{B_i}},$$

$$X_{\mathbf{B_i}} = \left(I - \Gamma_{\mathbf{B_i}}\right)^{-\intercal} \Gamma^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \left(I - \Gamma_{\mathbf{B_i}}\right)^{-\intercal} \epsilon_{\mathbf{B_i}}$$

$$= \Lambda^{\intercal}_{\mathrm{pa}(\mathbf{B_i},\mathcal{G}),\mathbf{B_i}} X_{\mathrm{pa}(\mathbf{B_i},\mathcal{G})} + \varepsilon_{\mathbf{B_i}},$$

Block-recursive reparametrization

Proposition (Block-recursive form, Guo and Perković, 2022)

Let $\mathbf{B_1}, \dots, \mathbf{B_K}$ be the ordered bucket decomposition of $\mathbf{V}$ in MPDAG $\mathcal{G}$. Then

$$X = \Lambda^{\intercal} X + \varepsilon, \qquad \Lambda = (\lambda_{ij}),\ J \in \mathbf{B_k},\ I \notin \text{pa}(\mathbf{B_k}, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0,$$

$$\mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{\mathbf{B_k}} \varepsilon_{\mathbf{B_k}}^{\intercal} \succ \mathbf{0}, \quad \varepsilon_{\mathbf{B_k}} \text{ mutually independent,}$$

Block-recursive reparametrization

Proposition (Block-recursive form, Guo and Perković, 2022)

Let $\mathbf{B_1}, \ldots, \mathbf{B_K}$ be the ordered bucket decomposition of $\mathbf{V}$ in MPDAG $\mathcal{G}$. Then

$$X = \Lambda^{\intercal} X + \varepsilon, \qquad \Lambda = (\lambda_{ij}),\ J \in \mathbf{B_k},\ I \notin \mathrm{pa}(\mathbf{B_k}, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0,$$
$$\mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{\mathbf{B_k}} \varepsilon_{\mathbf{B_k}}^{\intercal} \succ \mathbf{0}, \quad \varepsilon_{\mathbf{B_k}} \text{ mutually independent,}$$

Two nice things happen under this re-parametrization:

Block-recursive reparametrization

Proposition (Block-recursive form, Guo and Perković, 2022)

Let $\mathbf{B_1}, \ldots, \mathbf{B_K}$ be the ordered bucket decomposition of $\mathbf{V}$ in MPDAG $\mathcal{G}$. Then

$$X = \Lambda^{\intercal} X + \varepsilon, \qquad \Lambda = (\lambda_{ij}),\ J \in \mathbf{B_k},\ I \notin \mathrm{pa}(\mathbf{B_k}, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0,$$

$$\mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{\mathbf{B_k}} \varepsilon_{\mathbf{B_k}}^{\intercal} \succ \mathbf{0}, \quad \varepsilon_{\mathbf{B_k}} \text{ mutually independent,}$$

Two nice things happen under this re-parametrization:

- For $\mathbf{S} = \mathrm{An}(Y, \mathcal{G}_{\mathbf{V} \setminus \{A\}})$, τ_{AY} can be identified as

$$\tau_{AY} = \Lambda_{A,\mathbf{S}} \left[(I - \Lambda_{\mathbf{S},\mathbf{S}})^{-1} \right]_{\mathbf{S},Y}.$$

  The bucket-wise error distribution is a nuisance.

Block-recursive reparametrization

Proposition (Block-recursive form, Guo and Perković, 2022)

Let $\mathbf{B_1}, \dots, \mathbf{B_K}$ be the ordered bucket decomposition of $\mathbf{V}$ in MPDAG $\mathcal{G}$. Then

$$X = \Lambda^\intercal X + \varepsilon, \qquad \Lambda = (\lambda_{ij}),\ J \in \mathbf{B_k},\ I \notin \mathrm{pa}(\mathbf{B_k}, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0,$$
$$\mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{\mathbf{B_k}} \varepsilon_{\mathbf{B_k}}^\intercal \succ \mathbf{0}, \quad \varepsilon_{\mathbf{B_k}} \text{ mutually independent,}$$

Two nice things happen under this re-parametrization:

- For $\mathbf{S} = \mathrm{An}(Y, \mathcal{G}_{\mathbf{V} \setminus \{A\}})$, τ_{AY} can be identified as

$$\tau_{AY} = \Lambda_{A,\mathbf{S}} \left[(I - \Lambda_{\mathbf{S},\mathbf{S}})^{-1} \right]_{\mathbf{S},Y}.$$

  The bucket-wise error distribution is a nuisance.
- Under Gaussian errors, the MLE for each $\Lambda_{\mathrm{pa}(\mathbf{B_i}, \mathcal{G}), \mathbf{B_i}}$ corresponds to the least squares coefficients from $\mathbf{B_i} \sim \mathrm{pa}(\mathbf{B_i}, \mathcal{G})$. $\rightarrow \mathcal{G}$-regression.

Efficiency

Theorem ($\mathcal{G}$-regression, Guo and Perković, 2022)

If τ_{AY} is identifiable given MPDAG $\mathcal{G}$, the **$\mathcal{G}$-regression estimator** is defined as:

$$\hat{\tau}^{\mathcal{G}}_{AY} := \hat{\Lambda}^{\mathcal{G}}_{A,\mathbf{S}} \left[(I - \hat{\Lambda}^{\mathcal{G}}_{\mathbf{S},\mathbf{S}})^{-1} \right]_{\mathbf{S},Y},$$

where $\mathbf{S} = \mathrm{An}(Y, \mathcal{G}_{\mathbf{V}\setminus\{A\}})$, and $\hat{\Lambda}^{\mathcal{G}}$ is matrix consisting of least squares coefficients for each “bucket” regression.

Then for **any consistent estimator** $\hat{\tau}_{AY}$ of τ_{AY} such that $\hat{\tau}_{AY}$ is a **differentiable function of the sample covariance** it holds that

$$\mathrm{avar}\left(\hat{\tau}_{AY}\right) \geq \mathrm{avar}\left(\hat{\tau}^{\mathcal{G}}_{AY}\right).$$

Efficiency

Theorem ($\mathcal{G}$-regression, Guo and Perković, 2022)

If τ_{AY} is identifiable given MPDAG $\mathcal{G}$, the **$\mathcal{G}$-regression estimator** is defined as:

$$\hat{\tau}_{AY}^{\mathcal{G}} := \hat{\Lambda}_{A,\mathbf{S}}^{\mathcal{G}} \left[(I - \hat{\Lambda}_{\mathbf{S},\mathbf{S}}^{\mathcal{G}})^{-1} \right]_{\mathbf{S},Y},$$

where $\mathbf{S} = \mathrm{An}(Y, \mathcal{G}_{\mathbf{V} \setminus \{A\}})$, and $\hat{\Lambda}^{\mathcal{G}}$ is matrix consisting of least squares coefficients for each "bucket" regression.

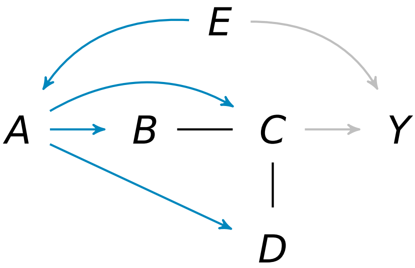
Then for **any consistent estimator** $\hat{\tau}_{AY}$ of τ_{AY} such that $\hat{\tau}_{AY}$ is a **differentiable function of the sample covariance** it holds that

$$\mathrm{avar}\left(\hat{\tau}_{AY}\right) \geq \mathrm{avar}\left(\hat{\tau}_{AY}^{\mathcal{G}}\right).$$

This includes estimators based on:

- covariate adjustment (Henckel et al, 2022, Witte et al, 2020),
- recursive regressions (Nandy et al, 2017, Gupta et al, 2020),
- modified Cholesky decomposition (Nandy et al, 2017).

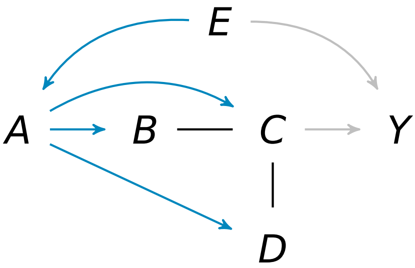
Block-recursive reparametrization



- Causal identification formula and $\mathcal{G}$-regression:

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[X_Y | \text{do}(x_a)]$$

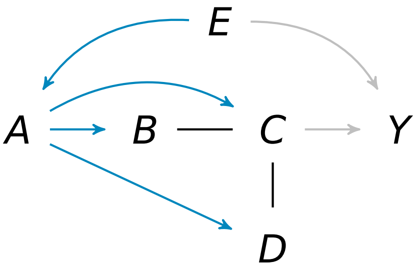
Block-recursive reparametrization



- Causal identification formula and $\mathcal{G}$-regression:

$$\begin{aligned}\tau_{AY} &= \frac{\partial}{\partial x_a}\mathbb{E}[X_Y|\mathrm{do}(x_a)]\\ &= \frac{\partial}{\partial x_a}\int \mathbb{E}[X_Y|x_c, x_e]f(x_c|x_a)f(x_e)dx_c dx_e\end{aligned}$$

Block-recursive reparametrization



- Causal identification formula and $\mathcal{G}$-regression:

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[X_Y | \mathrm{do}(x_a)]$$
$$= \frac{\partial}{\partial x_a} \int \mathbb{E}[X_Y | x_c, x_e] f(x_c | x_a) f(x_e) dx_c dx_e$$

- Suggests a **plug-in estimator** based on least squares regressions $X_Y \sim X_C + X_E, \quad X_C \sim X_A$.

Simulation results

An instance is simulated by the following steps.

1. Draw $\mathcal{D}$ from a random graph ensemble.
2. Take $\mathcal{G} = \text{CPDAG}(\mathcal{D})$.
3. Simulate data from a linear SCM with random error type (normal, t, logistic, uniform).
4. Choose (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $err = \|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

We compare $\mathcal{G}$-regression to the following estimators:

Simulation results

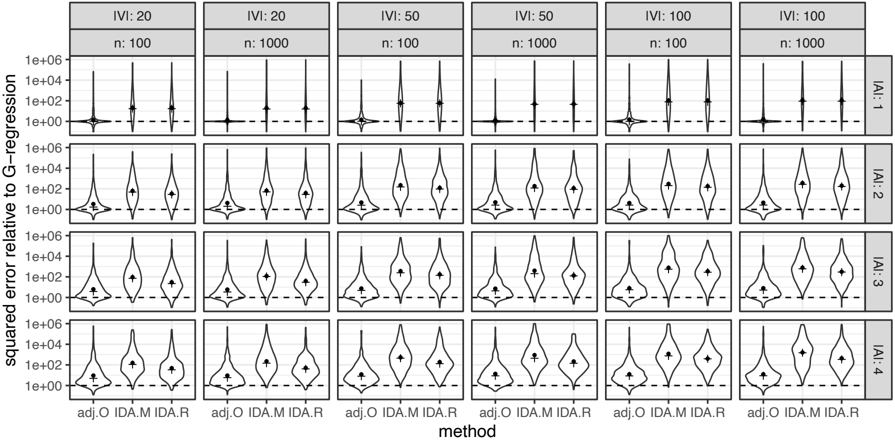
An instance is simulated by the following steps.

1. Draw $\mathcal{D}$ from a random graph ensemble.
2. Take $\mathcal{G} = \text{CPDAG}(\mathcal{D})$.
3. Simulate data from a linear SCM with random error type (normal, t, logistic, uniform).
4. Choose (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $err = \|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

We compare $\mathcal{G}$-regression to the following estimators:

- `adj.O`: optimal adjustment estimator (Henckel et al, 2022), or
- `IDA.M`: joint-IDA estimator based on modifying Cholesky decompositions (Nandy et al, 2017), or
- `IDA.R`: joint-IDA estimator based on recursive regressions (Nandy et al, 2017).

Simulation results



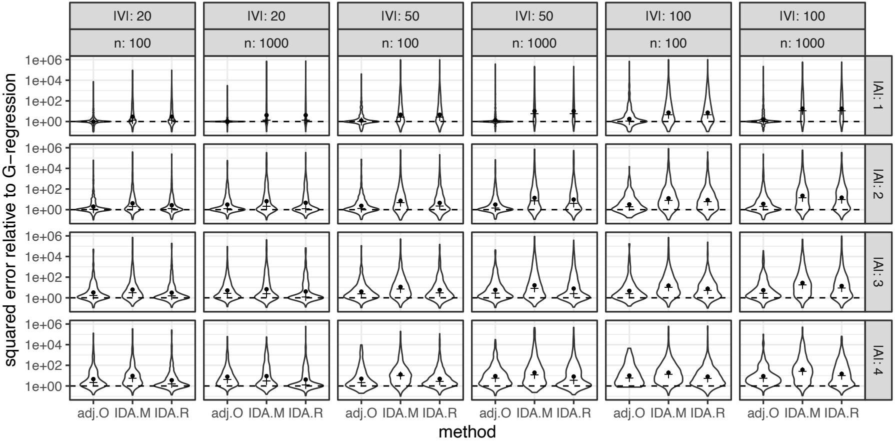
Violin plots displaying relative squared errors $\frac{estimator.err}{\mathcal{G}-reg.err}$ given the true DAG.

Simulation results

Table: Percentage of identified instances not estimable using contending estimators. All instances are estimable with $\mathcal{G}$-regression.

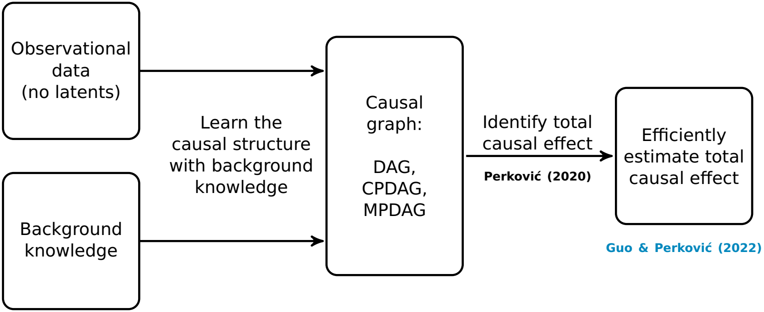
Estimator	$\|\mathbf{A}\|$	$\|\mathbf{V}\| = 20$	$\|\mathbf{V}\| = 50$	$\|\mathbf{V}\| = 100$
`adj.O`	1	0%	0%	0%
	2	17%	10%	5%
	3	30%	18%	15%
	4	36%	29%	22%
`IDA.M`	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%
`IDA.R`	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%

Simulation results



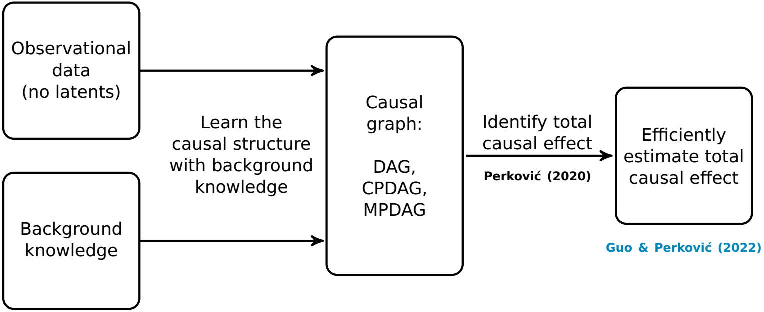
Violin plots displaying relative squared errors $\frac{\mathcal{G}-reg.err}{estimator.err}$ given GES estimated CPDAG.

Final remarks



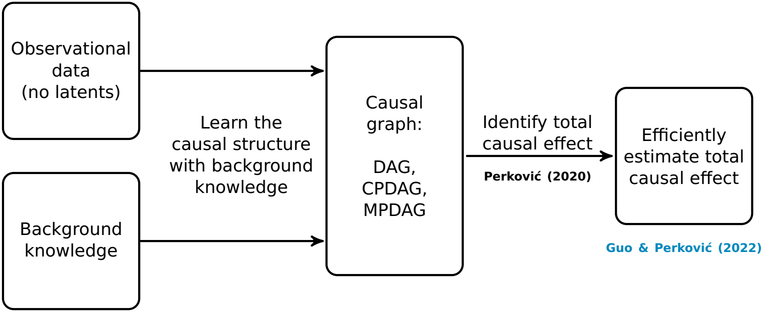
- **R package** `eff`^{2}: github.com/richardkwo/eff2

Final remarks



- **R package** `eff`^{2}: github.com/richardkwo/eff2
- **Beyond linear SCMs?** See Richard's talk on 02/23 at the Berkeley Causal Inference Group.

Final remarks



- **R package** `eff`^{2}: github.com/richardkwo/eff2
- **Beyond linear SCMs?** See Richard's talk on 02/23 at the Berkeley Causal Inference Group.

Thanks!

Simulation results

Table: Geometric average of squared errors relative to $\mathcal{G}$-regression, computed from estimable instances.

$\lvert\mathbf{A}\rvert$	$\lvert\mathbf{V}\rvert = 20$ $n = 100$	$\lvert\mathbf{V}\rvert = 20$ $n = 1000$	$\lvert\mathbf{V}\rvert = 50$ $n = 100$	$\lvert\mathbf{V}\rvert = 50$ $n = 1000$	$\lvert\mathbf{V}\rvert = 100$ $n = 100$	$\lvert\mathbf{V}\rvert = 100$ $n = 1000$
`adj.O`						
1	1.3	1.3	1.4	1.3	1.5	1.5
2	3.4	4.2	4.7	4.9	4.2	4.5
3	6.3	5.9	7.4	7.2	7.8	8.0
4	9.3	9.3	12	14	12	12
`IDA.M`						
1	20	19	61	48	103	108
2	62	65	220	182	293	356
3	93	119	354	396	749	771
4	154	222	533	895	1188	1604
`IDA.R`						
1	20	19	61	48	103	108
2	33	38	121	113	176	199
3	30	39	171	135	342	312
4	48	50	187	214	405	432

Simulation results

Table: Geometric average of squared errors relative to $\mathcal{G}$-regression, computed from estimable instances given GES estimated CPDAG

$\vert\mathbf{A}\vert$	$\vert\mathbf{V}\vert = 20$ $n = 100$	$\vert\mathbf{V}\vert = 20$ $n = 1000$	$\vert\mathbf{V}\vert = 50$ $n = 100$	$\vert\mathbf{V}\vert = 50$ $n = 1000$	$\vert\mathbf{V}\vert = 100$ $n = 100$	$\vert\mathbf{V}\vert = 100$ $n = 1000$
adj.O						
1	1.0	1.0	1.2	1.3	1.8	1.6
2	2.0	3.1	2.4	3.1	3.2	3.7
3	3.3	5.2	4.0	5.9	4.7	5.5
4	4.6	7.9	5.0	9.0	10	8.9
IDA.M						
5	2.9	4.1	4.5	10	7.3	18
6	4.2	6.6	7.3	14	13	22
7	6.2	6.8	12	16	15	28
8	9.5	9.0	13	20	19	37
IDA.R						
9	2.9	4.1	4.5	10	7.3	18
10	2.7	4.6	4.5	9.6	8.5	15
11	3.1	4.1	5.8	7.8	7.6	14
12	3.6	4.2	4.9	8.2	8.1	15

Identification of total causal effect

$\mathbf{S_1}, \ldots, \mathbf{S_K}$ is a partition of $\mathbf{S} = An(Y, \mathcal{G}_{\mathbf{V}\setminus\{A\}})$ induced by $\mathbf{B_1}, \ldots, \mathbf{B_K}$.
Let $\mathbf{F_k} = \{A\} \cap \text{pa}(\mathbf{S_k}, \mathcal{G})$, for all $k \in \{1, \ldots, k\}$. Then

$$P(X_{\mathbf{S}}|\text{do}(x_A)) = \prod_{k=1}^{K} P(X_{\mathbf{S_k}}|X_{\text{pa}(\mathbf{S_k},\mathcal{G})}) = \prod_{k=1}^{K} P(X_{\mathbf{S_k}}|X_{\text{pa}(\mathbf{S_k},\mathcal{G})\setminus\mathbf{F_k}}, X_{\mathbf{F_k}} = x_{\mathbf{F_k}}),$$

where $x_{\mathbf{F_k}}$ is fixed by the $\text{do}(x_A)$ operation.

$$\begin{aligned} X_{\mathbf{S_k}} \mid \left\{X_{\text{pa}(\mathbf{S_k},\mathcal{G})\setminus\mathbf{F_k}}, X_{F_i} = x_{\mathbf{F_k}}\right\} &=_d \Lambda^{\intercal}_{\text{pa}(\mathbf{S_k},\mathcal{G})\setminus\mathbf{F_k},\mathbf{S_k}} X_{\text{pa}(\mathbf{S_k},\mathcal{G})\setminus\mathbf{F_k}} + \Lambda_{\mathbf{F_k},\mathbf{S_k}} x_{\mathbf{F_k}} + \varepsilon_{\mathbf{S_k}} \\ &= \Lambda^{\intercal}_{\text{pa}(\mathbf{S_k},\mathcal{G})\cap\mathbf{S},\mathbf{S_k}} X_{\text{pa}(\mathbf{S_k},\mathcal{G})\cap\mathbf{S}} + \Lambda_{\text{pa}(\mathbf{S_k},\mathcal{G})\cap\{A\},\mathbf{S_k}} x_{\text{pa}(\mathbf{S_k},\mathcal{G})\cap\{A\}} + \varepsilon_{\mathbf{S_k}} \end{aligned}$$

The fact that the display above holds for every $k = 1, \ldots, K$ implies that the joint interventional distribution $P(X_{\mathbf{S}}|\text{do}(x_A))$ satisfies

$$X_{\mathbf{S}} = \Lambda^{T}_{\mathbf{S},\mathbf{S}} X_{\mathbf{S}} + \Lambda^{\intercal}_{A,\mathbf{S}} x_A + \varepsilon_{\mathbf{S}}.$$

It follows that $X_{\mathbf{S}} = (I - \Lambda_{\mathbf{S},\mathbf{S}})^{-\intercal}(\Lambda^{\intercal}_{A,\mathbf{S}} x_A + \varepsilon_{\mathbf{S}})$ and since $Y \in \mathbf{S}$, we have

$$\tau_{AY} = \frac{\partial}{\partial x_A} \mathbb{E}[X_Y \mid \text{do}(x_A)] = \Lambda_{A,\mathbf{S}} \left[(I - \Lambda_{\mathbf{S},\mathbf{S}})^{-1}\right]_{\mathbf{S},Y}.$$

Efficiency theory

Let Σ_n be the sample covariance. Consider the class of estimators

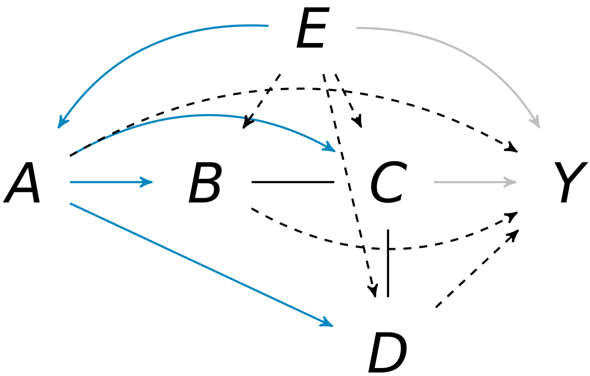
$$\mathcal{T} = \Big\{\hat{\tau}(\Sigma_n) : \mathbb{R}^{|\mathbf{V}|\times|\mathbf{V}|}_{\text{PD}} \to \mathbb{R}^{|\mathbf{A}|} : \hat{\tau}(\Sigma_n) \text{ is a differentiable and consistent estimator of } \tau_{AY}\Big\}.$$

The efficiency theory entails two parts.

- Establish an efficiency bound on $\mathcal{T}$.
  The bound is derived from the gradient condition on $\mathcal{T}$ (as in standard semiparametric efficiency theory) and a **diffeomorphism**

  $$\mathbb{R}^{|\mathbf{V}|\times|\mathbf{V}|}_{\text{PD}} \longleftrightarrow ((\Lambda_{\text{pa}(\mathbf{B_k},\bar{\mathcal{G}}),\mathbf{B_k}}, \Omega_k) : k = 1, \ldots, K) \text{ associated with } \bar{\mathcal{G}},$$

  where $\bar{\mathcal{G}}$ is the saturated version of $\mathcal{G}$.
  This generalizes a result from Drton (2018).
- Verify that $\hat{\tau}^{\mathcal{G}}_{AY}$ achieves this bound.



Saturated $\bar{\mathcal{G}}$ according to buckets.

$$\mathbf{B_1} = \{E\},\ \mathbf{B_2} = \{A\},\ \mathbf{B_3} = \{B, C, D\},\ \mathbf{B_4} = \{Y\}.$$

Proof sketch

1. Suppose $|\mathbf{A}| = 1$. Rewrite $\hat{\tau} \in \mathcal{T}$ as

$$\hat{\tau}(\Sigma_n) = \hat{\tau}\left((\hat{\Lambda}_k)_{k,\mathcal{G}}, (\hat{\Lambda}_k)_{k,\mathcal{G}^c}, (\hat{\Omega}_k)_k\right),$$

   where $(\hat{\Lambda}_k)_{k,\mathcal{G}^c} = (\hat{\Lambda}_k)_{k,\bar{\mathcal{G}}\setminus\mathcal{G}}$ are introduced dashed edges.

2. Consistency of $\hat{\tau}$ implies

$$\frac{\partial \hat{\tau}}{\partial \hat{\Lambda}_{k,\mathcal{G}}} = \frac{\partial \tau_{\mathcal{G}}}{\partial \hat{\Lambda}_{k,\mathcal{G}}} \ (k = 2, \ldots, K), \quad \frac{\partial \hat{\tau}}{\partial \hat{\Omega}_k} = \mathbf{0} \ (k = 1, \ldots, K),$$

   but $\frac{\partial \hat{\tau}}{\partial \hat{\Lambda}_{k,\mathcal{G}^c}}$ is free to vary.

3. Compute acov of $\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right)$ via asymptotic linear expansions.

4. By the delta method, an upper bound can be derived from quadratic form

$$\begin{aligned}
\operatorname{avar}(\hat{\tau}) &= \begin{pmatrix} \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}})_k} \\ \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}^c})_k} \end{pmatrix}^{\top} \operatorname{acov}\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right) \begin{pmatrix} \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}})_k} \\ \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}^c})_k} \end{pmatrix} \\
&\leq \sup_{\partial \hat{\tau} / \partial (\hat{\Lambda}_{k,\mathcal{G}^c})_k} \begin{pmatrix} \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}})_k} \\ \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}^c})_k} \end{pmatrix}^{\top} \operatorname{acov}\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right) \begin{pmatrix} \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}})_k} \\ \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_{k,\mathcal{G}^c})_k} \end{pmatrix}.
\end{aligned}$$