Efficient Least Squares for Estimating Total Causal Effects

Richard Guo, Emilija Perković

Pacific Causal Inference Conference, 2020

Department of Statistics, University of Washington, Seattle

Highlights

Highlights

- We consider estimating a total causal effect from **observational data**.

Highlights

- We consider estimating a total causal effect from **observational data**.
- We assume

Highlights

- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.

Highlights

- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.
  - **Causal sufficiency**: no unobserved confounding, no selection bias.

Highlights

- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.
  - **Causal sufficiency**: no unobserved confounding, no selection bias.
- The causal DAG is known up to a Markov equivalence class with additional background knowledge.

Highlights

- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.
  - **Causal sufficiency**: no unobserved confounding, no selection bias.
- The causal DAG is known up to a Markov equivalence class with additional background knowledge.
- We present a least squares estimator that is

Highlights

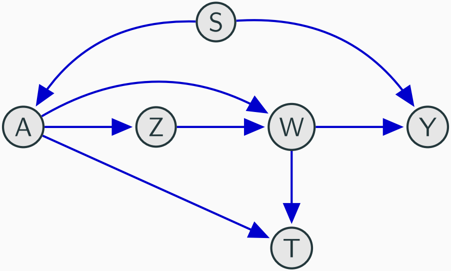
- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.
  - **Causal sufficiency**: no unobserved confounding, no selection bias.
- The causal DAG is known up to a Markov equivalence class with additional background knowledge.
- We present a least squares estimator that is
  - **Complete**: applicable whenever the effect is identified,

Highlights

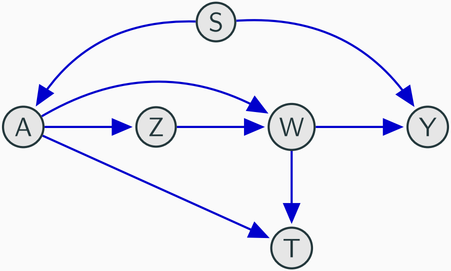
- We consider estimating a total causal effect from **observational data**.
- We assume
  - **Linearity**: data is generated from a linear structural equation model.
  - **Causal sufficiency**: no unobserved confounding, no selection bias.
- The causal DAG is known up to a Markov equivalence class with additional background knowledge.
- We present a least squares estimator that is
  - **Complete**: applicable whenever the effect is identified,
  - **Efficient**: relative to a large class of estimators,

  which is the first of its kind in the literature ...

Causal DAG, linear SEM

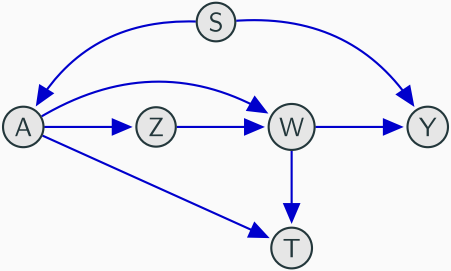


Causal DAG, linear SEM



Suppose $\mathcal{D}$ is the underlying causal DAG. $\mathcal{D}$ is **unknown**.

Causal DAG, linear SEM

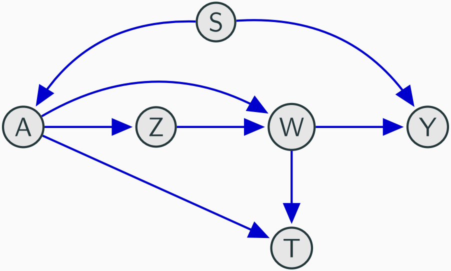


Suppose $\mathcal{D}$ is the underlying causal DAG. $\mathcal{D}$ is **unknown**.

Suppose data is generated by a linear structural equation model (SEM)

$$X_v = \sum_{u: u \to v} \gamma_{uv} X_u + \epsilon_u, \quad \mathbb{E}\, \epsilon_u = 0, \quad 0 < \operatorname{var} \epsilon_u < \infty.$$

Causal DAG, linear SEM



Suppose $\mathcal{D}$ is the underlying causal DAG. $\mathcal{D}$ is **unknown**.

Suppose data is generated by a linear structural equation model (SEM)

$$X_v = \sum_{u:u \to v} \gamma_{uv} X_u + \epsilon_u, \quad \mathbb{E}\,\epsilon_u = 0, \quad 0 < \operatorname{var} \epsilon_u < \infty.$$

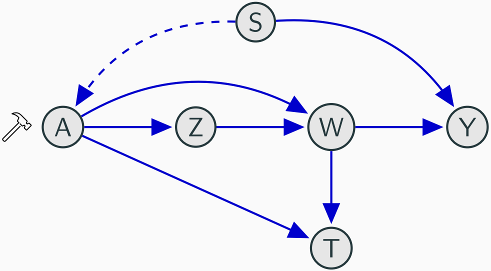
Under causal sufficiency, the errors are **mutually independent** (no $i \leftrightarrow j$ in the path diagram).

Total effect

Suppose we want to estimate the total (causal) effect of A on Y.

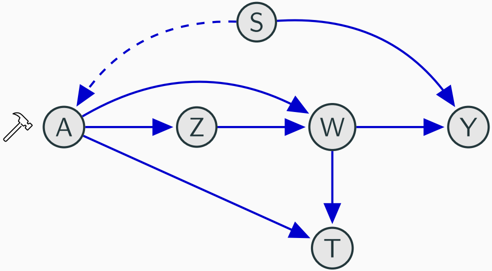
Total effect

Suppose we want to estimate the total (causal) effect of A on Y.



Total effect

Suppose we want to estimate the total (causal) effect of A on Y.

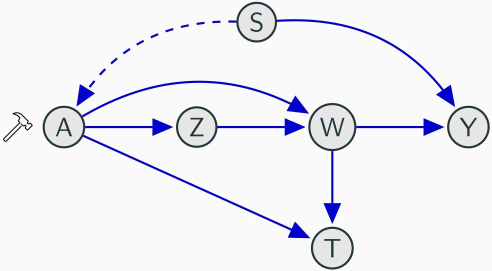


☞ The total effect τ_{AY} is defined as the slope of $x_a \mapsto \mathbb{E}[X_Y|\text{do}(X_A = x_a)]$, given by a sum-product of Wright (1934):

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[X_Y|\text{do}(X_A = x_a)] = (\gamma_{AZ}\gamma_{ZW} + \gamma_{AW})\gamma_{WY}.$$

Total effect

Suppose we want to estimate the total (causal) effect of A on Y.



☞ The total effect τ_{AY} is defined as the slope of $x_a \mapsto \mathbb{E}[X_Y | \mathrm{do}(X_A = x_a)]$, given by a sum-product of Wright (1934):

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[X_Y | \mathrm{do}(X_A = x_a)] = (\gamma_{AZ}\gamma_{ZW} + \gamma_{AW})\gamma_{WY}.$$

Here we consider point intervention ($|A| = 1$) for simplicity. For a joint intervention ($|A| > 1$), total effect can be similarly defined.

Markov equivalence, CPDAG

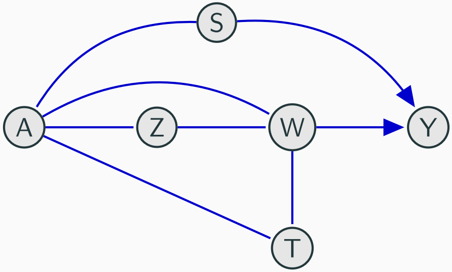
Markov equivalence, CPDAG

Without making further assumptions, the causal DAG $\mathcal{D}$ can only be identified from observed distribution up to a **Markov equivalence class**.

Markov equivalence, CPDAG

Without making further assumptions, the causal DAG $\mathcal{D}$ can only be identified from observed distribution up to a **Markov equivalence class**.

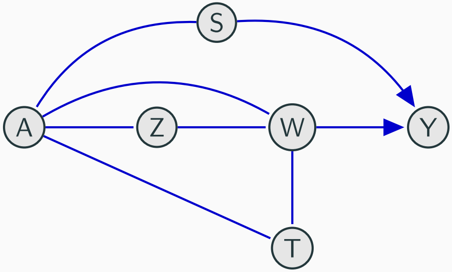
The Markov equivalence class of $\mathcal{D}$ is uniquely represented by a CPDAG/essential graph $\mathcal{C}$.



Markov equivalence, CPDAG

Without making further assumptions, the causal DAG $\mathcal{D}$ can only be identified from observed distribution up to a **Markov equivalence class**.

The Markov equivalence class of $\mathcal{D}$ is uniquely represented by a CPDAG/essential graph $\mathcal{C}$.



☞ Knowing only $\mathcal{C}$ is often **insufficient** to identify the total effect.

Identifiability from a partially directed graph

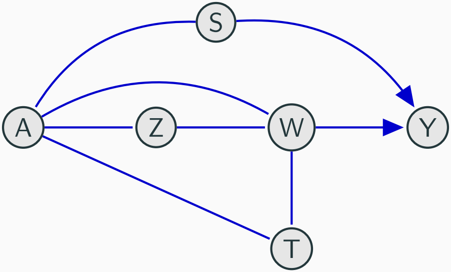
Theorem (Perković, 2020)

The total effect τ_{AY} is identified from a maximally oriented partially directed acyclic graph $\mathcal{G}$ **if and only if** there is no proper, possibly causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge.

Identifiability from a partially directed graph

Theorem (Perković, 2020)

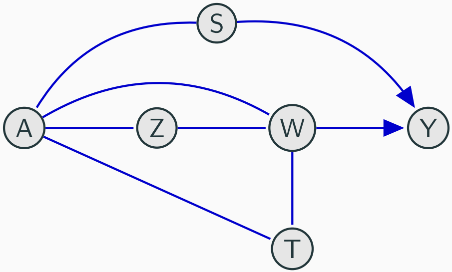
The total effect τ_{AY} is identified from a maximally oriented partially directed acyclic graph $\mathcal{G}$ **if and only if** there is no proper, possibly causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge.



Identifiability from a partially directed graph

Theorem (Perković, 2020)

The total effect τ_{AY} is identified from a maximally oriented partially directed acyclic graph $\mathcal{G}$ **if and only if** there is no proper, possibly causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge.



☞ In the unidentified case, see also the IDA algorithms (Maathuis, Kalisch, and Bühlmann, 2009; Nandy, Maathuis, and Richardson, 2017) that enumerates possible total effects.

Background knowledge, MPDAG

However, often we have additional knowledge that can help towards identification.

Background knowledge, MPDAG

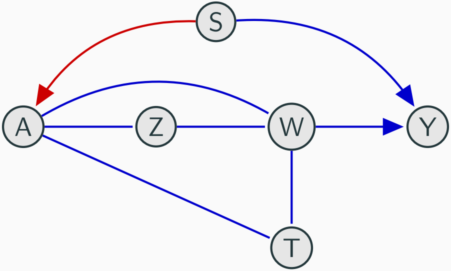
However, often we have additional knowledge that can help towards identification.

☞ Suppose we know that S **temporally preceeds** A.

Background knowledge, MPDAG

However, often we have additional knowledge that can help towards identification.

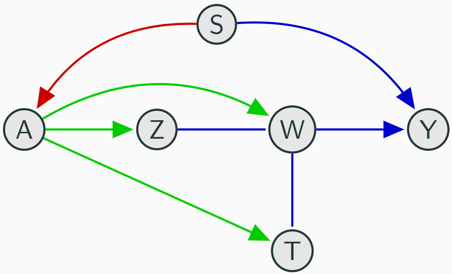
☞ Suppose we know that S **temporally preceeds** A.



Background knowledge, MPDAG

However, often we have additional knowledge that can help towards identification.

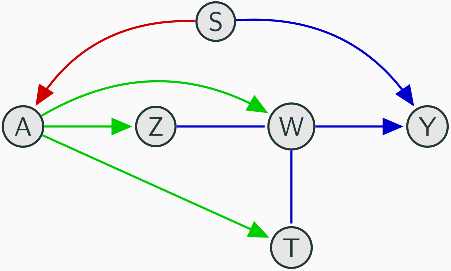
☞ Suppose we know that S **temporally preceeds** A.



Background knowledge, MPDAG

However, often we have additional knowledge that can help towards identification.

☞ Suppose we know that S **temporally preceeds** A.

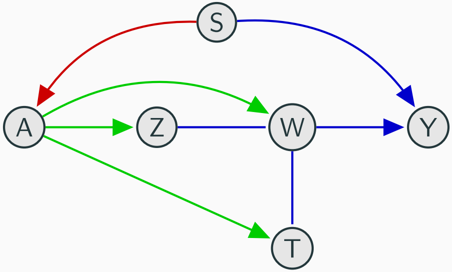


The green orientations are further **implied** by the rules of Meek (1995).

Background knowledge, MPDAG

However, often we have additional knowledge that can help towards identification.

☞ Suppose we know that S **temporally preceeds** A.



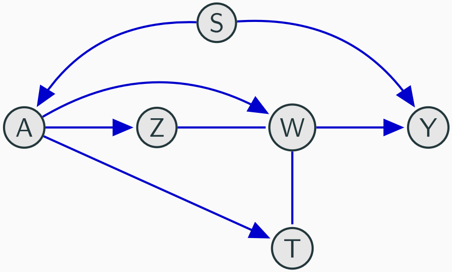
The green orientations are further **implied** by the rules of Meek (1995).

☞ In this example, τ_{AY} is **identified** from the resulting maximally oriented partially directed acyclic graph (MPDAG) $\mathcal{G}$.

Adjustment estimator

Our task is to estimate τ_{AY} from n iid observational sample generated by a linear SEM associated with causal DAG $\mathcal{D}$, given that

$$\mathcal{D} \in [\mathcal{G}] \text{ for MPDAG } \mathcal{G}, \quad \tau_{AY} \text{ is identifiable from } \mathcal{G}.$$

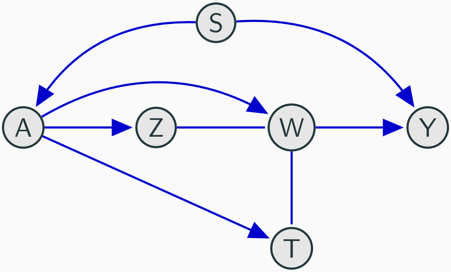


MPDAG $\mathcal{G}$

Adjustment estimator

Our task is to estimate τ_{AY} from n iid observational sample generated by a linear SEM associated with causal DAG $\mathcal{D}$, given that

$$\mathcal{D} \in [\mathcal{G}] \text{ for MPDAG } \mathcal{G}, \quad \tau_{AY} \text{ is identifiable from } \mathcal{G}.$$



MPDAG $\mathcal{G}$

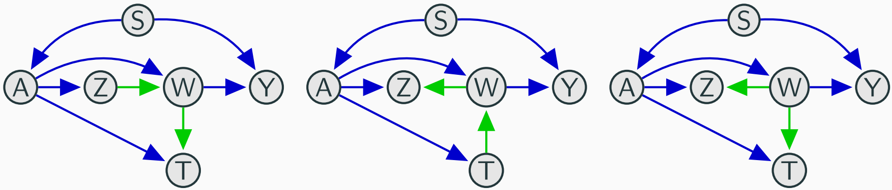
☞ **Adjustment estimator**: $\hat{\tau}_{AY}^{\text{adj}}$ is the least squares coefficient of A from $Y \sim A + S$.

Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.

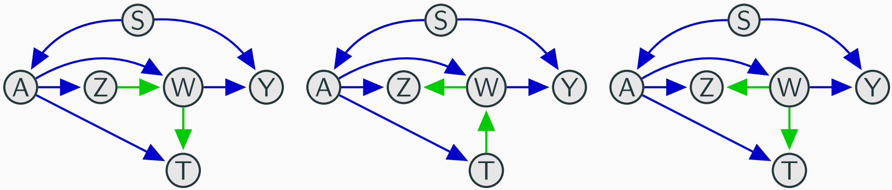
Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.



Adjustment estimator

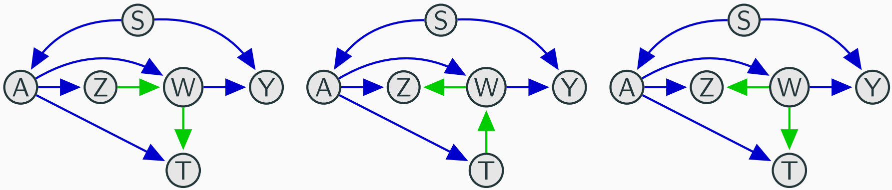
Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.



Adjustment estimator

Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.

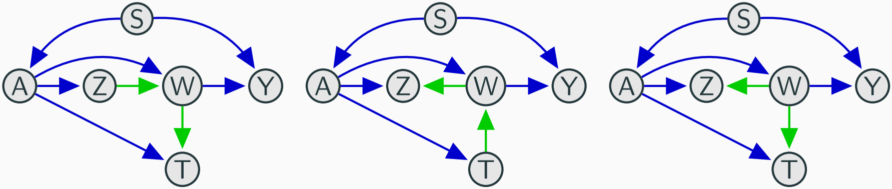


Adjustment estimator

- may not exist when $|A| > 1$.

Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.

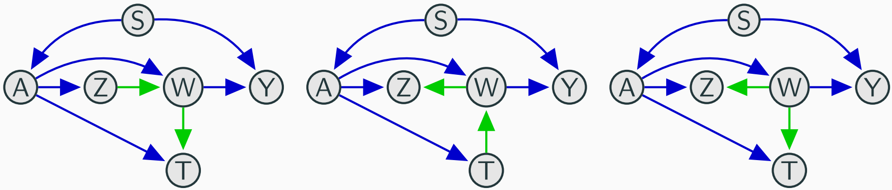


Adjustment estimator

- may not exist when $|A| > 1$.
- may not be unique.

Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.

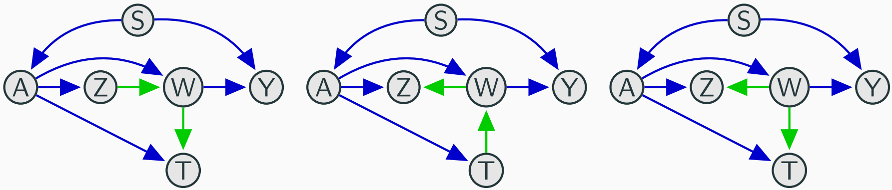


Adjustment estimator

- may not exist when $|A| > 1$.
- may not be unique.
  - The most efficient adjustment estimator is recently characterized by Henckel, Perković, and Maathuis (2019) and Witte et al. (2020).

Adjustment estimator

Adjustment $Y \sim A + S$ can be justified by looking at the elements of $[\mathcal{G}]$.

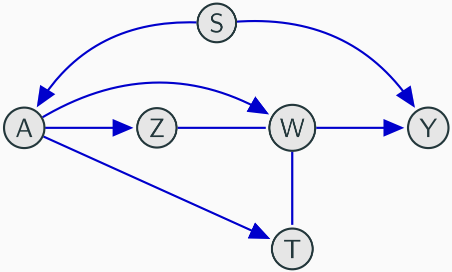


Adjustment estimator

- may not exist when $|A| > 1$.
- may not be unique.
  - The most efficient adjustment estimator is recently characterized by Henckel, Perković, and Maathuis (2019) and Witte et al. (2020).
- not efficient.

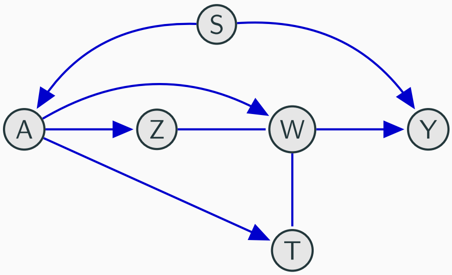
Our proposal: $\mathcal{G}$-regression estimator

We achieve efficient estimation by exploiting the "additional" conditional independences in $\mathcal{G}$ in this over-identified setting.



Our proposal: $\mathcal{G}$-regression estimator

We achieve efficient estimation by exploiting the "additional" conditional independences in $\mathcal{G}$ in this over-identified setting.

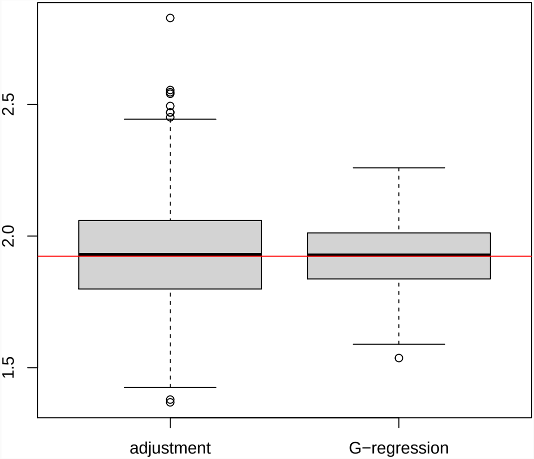


☞ $\mathcal{G}$**-regression estimator**

$$\hat{\tau}^{\mathcal{G}}_{AY} = \hat{\lambda}_{AW}\hat{\lambda}_{WY},$$

where $\hat{\lambda}_{AW}$, $\hat{\lambda}_{WY}$ are taken from $W \sim A$ and $Y \sim W + S$ respectively.

Our proposal: $\mathcal{G}$-regression estimator



$n = 100$, t_5 errors.

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-**regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$**-regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

$\mathcal{G}$-regression estimator is

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \text{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-**regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

$\mathcal{G}$-regression estimator is

1. **complete**,

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-**regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

$\mathcal{G}$-regression estimator is

1. **complete**,
2. the **most efficient** estimator among all consistent, regular estimators that only depend on the **first two moments** of data.

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-**regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

$\mathcal{G}$-regression estimator is

1. **complete**,
2. the **most efficient** estimator among all consistent, regular estimators that only depend on the **first two moments** of data.

- How to derive this estimator?

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. **$\mathcal{G}$-regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

$\mathcal{G}$-regression estimator is

1. **complete**,
2. the **most efficient** estimator among all consistent, regular estimators that only depend on the **first two moments** of data.

- How to derive this estimator?

1. Find the MLE under Gaussian errors.

$\mathcal{G}$-regression estimator

Define the set of vertices $D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-**regression estimator** is

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

where $\hat{\Lambda}^{\mathcal{G}}$ is a $|V| \times |V|$ matrix consisting of least squares coefficients for each "bucket".

Theorem

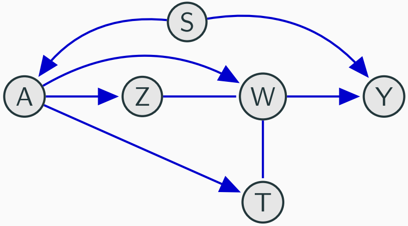
$\mathcal{G}$-regression estimator is

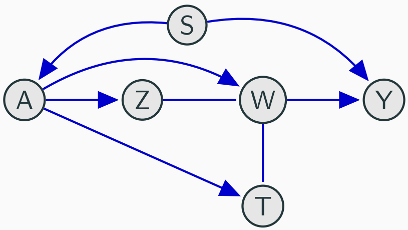
1. **complete**,
2. the **most efficient** estimator among all consistent, regular estimators that only depend on the **first two moments** of data.

- How to derive this estimator?

1. Find the MLE under Gaussian errors.
2. Show that this MLE is "efficient" even when errors are non-Gaussian.

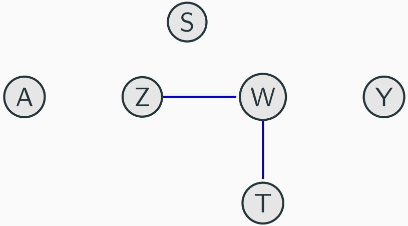
Buckets, reparametrization and Gaussian MLE





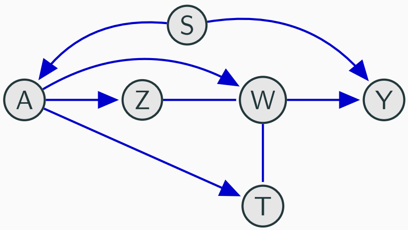
Let "buckets" be the maximal connected components of the undirected part of $\mathcal{G}$.

Buckets, reparametrization and Gaussian MLE



Let "buckets" be the maximal connected components of the undirected part of $\mathcal{G}$.

Buckets, reparametrization and Gaussian MLE

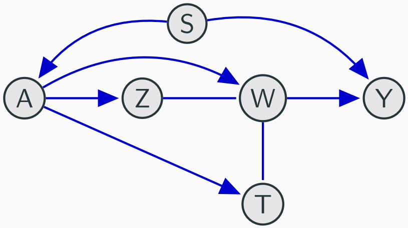


Let "buckets" be the maximal connected components of the undirected part of $\mathcal{G}$.

Further, buckets can be topologically ordered by the directed part of $\mathcal{G}$:

$$B_1 = \{S\},\ B_2 = \{A\},\ B_3 = \{Z, W, T\},\ B_4 = \{Y\}.$$

Buckets, reparametrization and Gaussian MLE



Let "buckets" be the maximal connected components of the undirected part of $\mathcal{G}$.

Further, buckets can be topologically ordered by the directed part of $\mathcal{G}$:

$$B_1 = \{S\},\ B_2 = \{A\},\ B_3 = \{Z, W, T\},\ B_4 = \{Y\}.$$

Lemma: Restrictive property

For each bucket B_i, vertices in B_i have the same set of external parents, denoted as $\mathrm{Pa}(B_i)$.

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\mathsf{T}}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\intercal}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

▶ Two nice things happen under this reparametrization:

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\intercal}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

▶ Two nice things happen under this reparametrization:

1. With $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$, τ_{AY} can be identified as

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

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\intercal}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

▶ Two nice things happen under this reparametrization:

1. With $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$, τ_{AY} can be identified as

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

☞ The bucket-wise **error distribution** is **nuisance**.

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\intercal}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

▶ Two nice things happen under this reparametrization:

1. With $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$, τ_{AY} can be identified as

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

☞ The bucket-wise **error distribution** is **nuisance**.

2. Under Gaussian errors, the MLE for each $\Lambda_{\mathrm{Pa}(B_k),B_k}$ is just the least squares coefficients of $B_k \sim \mathrm{Pa}(B_k)$.

Buckets, reparametrization and Gaussian MLE

The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda^{\mathsf{T}}_{\mathrm{Pa}(B_k),B_k} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$

- Λ: $|V| \times |V|$ upper-triangular matrix corresponding to directed edges between buckets.
- ε_{B_k}: errors associated with bucket B_k, independent across buckets.

▶ Two nice things happen under this reparametrization:

1. With $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$, τ_{AY} can be identified as

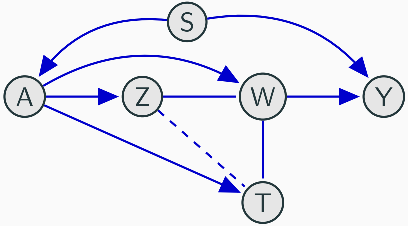
$$\tau_{AY} = \Lambda_{A,D} \left[(I - \Lambda_{D,D})^{-1} \right]_{D,Y}.$$

☞ The bucket-wise **error distribution** is **nuisance**.

2. Under Gaussian errors, the MLE for each $\Lambda_{\mathrm{Pa}(B_k),B_k}$ is just the least squares coefficients of $B_k \sim \mathrm{Pa}(B_k)$. ☞ $\mathcal{G}$-regression.

Buckets, reparametrization and Gaussian MLE

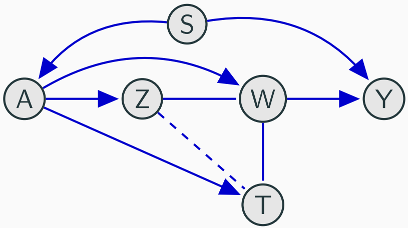
The second property is a special case of “seemingly unrelated regression” due to the **restrictive property**.



$$(X_Z, X_W, X_T) = (\lambda_{AZ}, \lambda_{AW}, \lambda_{AT})X_A + \varepsilon_{B_3},$$
$$\varepsilon_{B_3} \sim \mathcal{N}(\mathbf{0}, \Omega_3), \quad (\Omega_3)_{ZT \cdot W} = 0.$$

Buckets, reparametrization and Gaussian MLE

The second property is a special case of "seemingly unrelated regression" due to the **restrictive property**.



$$(X_Z, X_W, X_T) = (\lambda_{AZ}, \lambda_{AW}, \lambda_{AT}) X_A + \varepsilon_{B_3},$$
$$\varepsilon_{B_3} \sim \mathcal{N}(\mathbf{0}, \Omega_3), \quad (\Omega_3)_{ZT\cdot W} = 0.$$

☞ See also Anderson and Olkin (1985, §5) and Amemiya (1985, §6.4) for this phenomenon.

Efficiency theory

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

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

Efficiency theory

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

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

The efficiency theory entails two parts.

Efficiency theory

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

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

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**

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

where $\bar{\mathcal{G}}$ is the saturated version of $\mathcal{G}$.

Efficiency theory

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

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

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**

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

    where $\bar{\mathcal{G}}$ is the saturated version of $\mathcal{G}$.
  - ☞ This generalizes a result from Drton (2018).

Efficiency theory

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

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

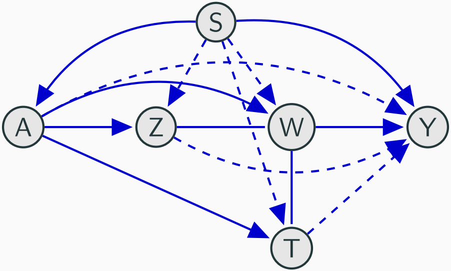
The efficiency theory entails two parts.

- ☞ 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**

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

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

Efficiency theory



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

$$B_1 = \{S\},\ B_2 = \{A\},\ B_3 = \{Z, W, T\},\ B_4 = \{Y\}.$$

Proof sketch

Proof sketch

1. Suppose $|A| = 1$. 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.

Proof sketch

1. Suppose $|A| = 1$. 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**.

Proof sketch

1. Suppose $|A| = 1$. 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**.
3. Compute acov of $\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right)$ via asymptotic linear expansions.

Proof sketch

1. Suppose $|A| = 1$. 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**.
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
$$\begin{aligned}\mathrm{avar}(\hat{\tau}) &= \begin{pmatrix}\frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}})_k}\\ \frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}^c})_k}\end{pmatrix}^{\intercal} \mathrm{acov}\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right) \begin{pmatrix}\frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}})_k}\\ \frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}^c})_k}\end{pmatrix}\\ &\leq \sup_{\partial\hat{\tau}/\partial(\hat{\Lambda}_{k,\mathcal{G}^c})_k} \begin{pmatrix}\frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}})_k}\\ \frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}^c})_k}\end{pmatrix}^{\intercal} \mathrm{acov}\left((\hat{\Lambda}_{k,\mathcal{G}})_k, (\hat{\Lambda}_{k,\mathcal{G}^c})_k\right) \begin{pmatrix}\frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}})_k}\\ \frac{\partial\hat{\tau}}{\partial(\hat{\Lambda}_{k,\mathcal{G}^c})_k}\end{pmatrix}.\end{aligned}$$

Simulation results

An instance is simulated by the following steps.

Simulation results

An instance is simulated by the following steps.

1. Draw $\mathcal{D}$ from a random graph ensemble.

Simulation results

An instance is simulated by the following steps.

1. Draw $\mathcal{D}$ from a random graph ensemble.
2. Take $\mathcal{G} = \text{CPDAG}(\mathcal{D})$.

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 SEM with random coefficients and a random error type (normal, t, logistic, uniform).

Simulation results

An instance is simulated by the following steps.

1. Draw $\mathcal{D}$ from a random graph ensemble.
2. Take $\mathcal{G} = \mathrm{CPDAG}(\mathcal{D})$.
3. Simulate data from a linear SEM with random coefficients and a random error type (normal, t, logistic, uniform).
4. Pick (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.

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 SEM with random coefficients and a random error type (normal, t, logistic, uniform).
4. Pick (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

☞ We compare to the following estimators in the literature:

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 SEM with random coefficients and a random error type (normal, t, logistic, uniform).
4. Pick (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

☞ We compare to the following estimators in the literature:

- `adj.O`: optimal adjustment estimator (Henckel, Perković, and Maathuis, 2019),

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 SEM with random coefficients and a random error type (normal, t, logistic, uniform).
4. Pick (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

☞ We compare to the following estimators in the literature:

- `adj.O`: optimal adjustment estimator (Henckel, Perković, and Maathuis, 2019),
- `IDA.M`: joint-IDA estimator based on modifying Cholesky decompositions (Nandy, Maathuis, and Richardson, 2017),

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 SEM with random coefficients and a random error type (normal, t, logistic, uniform).
4. Pick (A, Y) such that τ_{AY} is identified from $\mathcal{G}$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

☞ We compare to the following estimators in the literature:

- `adj.O`: optimal adjustment estimator (Henckel, Perković, and Maathuis, 2019),
- `IDA.M`: joint-IDA estimator based on modifying Cholesky decompositions (Nandy, Maathuis, and Richardson, 2017),
- `IDA.R`: joint-IDA estimator based on recursive regressions (Nandy, Maathuis, and Richardson, 2017).

Simulation results

Table 1: Percentage of identified instances not estimable using contending estimators. All instances are estimable with $\mathcal{G}$-regression.

Estimator	$\lvert A \rvert$	$\lvert V \rvert = 20$	$\lvert V \rvert = 50$	$\lvert V \rvert = 100$
`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

Table 2: Geometric average of squared errors relative to $\mathcal{G}$-regression, computed from estimable instances.

$\lvert A \rvert$	$\lvert V \rvert = 20$		$\lvert V \rvert = 50$		$\lvert V \rvert = 100$	
	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
`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

Final remarks

Final remarks

- **Details**: arxiv.org/abs/2008.03481

Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** `eff`^{2}: github.com/richardkwo/eff2

Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** `eff`^{2}: github.com/richardkwo/eff2
- **Why restricting to the first two moments?**
  This is a large class of estimators, containing all the estimators we know from the literature ...

Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** `eff`^{2}: github.com/richardkwo/eff2
- **Why restricting to the first two moments?**
  This is a large class of estimators, containing all the estimators we know from the literature ...

  Also, this is a tradeoff between theory and practice. The problem is a generalized, multivariate location-shift regression model (Bickel et al., 1993; Tsiatis, 2006). Theoretically, a semiparametric efficient estimator can be constructed by estimating the error score and then solving estimating equations. But the resulting estimator seems unstable for practical purposes (Tsiatis, 2006).

Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** `eff`^{2}: github.com/richardkwo/eff2
- **Why restricting to the first two moments?**
  This is a large class of estimators, containing all the estimators we know from the literature ...

  Also, this is a tradeoff between theory and practice. The problem is a generalized, multivariate location-shift regression model (Bickel et al., 1993; Tsiatis, 2006). Theoretically, a semiparametric efficient estimator can be constructed by estimating the error score and then solving estimating equations. But the resulting estimator seems unstable for practical purposes (Tsiatis, 2006).
- **Beyond linear SEMs?**
  It worth considering generalization along the lines of Rotnitzky and Smucler (2019).

References

Amemiya, Takeshi (1985). *Advanced Econometrics*. Harvard University Press.

Anderson, Theodore Wilbur and Ingram Olkin (1985). "Maximum-likelihood estimation of the parameters of a multivariate normal distribution". In: *Linear algebra and its applications* 70, pp. 147–171.

Bickel, Peter J. et al. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Vol. 4. Baltimore: Johns Hopkins University Press.

References ii

Drton, Mathias (2018). "Algebraic problems in structural equation modeling". In: *The 50th Anniversary of Gröbner Bases*. Mathematical Society of Japan, pp. 35–86.

Henckel, Leonard, Emilija Perković, and Marloes H. Maathuis (2019). "Graphical criteria for efficient total effect estimation via adjustment in causal linear models". In: *arXiv preprint arXiv:1907.02435*.

Maathuis, Marloes H., Markus Kalisch, and Peter Bühlmann (2009). "Estimating high-dimensional intervention effects from observational data". In: *The Annals of Statistics* 37.6A, pp. 3133–3164.

Meek, Christopher (1995). "Causal inference and causal explanation with background knowledge". In: *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pp. 403–410.

References iii

Nandy, Preetam, Marloes H. Maathuis, and Thomas S. Richardson (2017). "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings". In: *The Annals of Statistics* 45.2, pp. 647–674.

Perković, Emilija (2020). "Identifying causal effects in maximally oriented partially directed acyclic graphs". In: *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*.

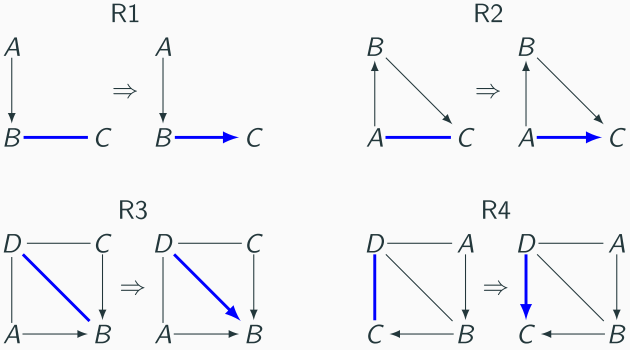
Rotnitzky, Andrea and Ezequiel Smucler (2019). "Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models". In: *arXiv preprint arXiv:1912.00306*.

Tsiatis, Anastasios (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Witte, Janine et al. (2020). "On efficient adjustment in causal graphs". In: *arXiv preprint arXiv:2002.06825*.

Wright, Sewall (1934). "The Method of Path Coefficients". In: *The Annals of Mathematical Statistics* 5.3, pp. 161–215.

Meek's rules



The orientation rules from Meek (1995).