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

• 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...
• We consider estimating a total causal effect from observational data.
• We consider estimating a total causal effect from observational data.
• We assume
• 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
• 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,
• We consider estimating a total causal effect from \textit{observational data}.

• We assume
  • \textbf{Linearity}: data is generated from a linear structural equation model.
  • \textbf{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
  • \textbf{Complete}: applicable whenever the effect is identified,
  • \textbf{Efficient}: relative to a large class of estimators,
which is the first of its kind in the literature ...
Suppose $D$ is the underlying causal DAG. $D$ is unknown. Suppose data is generated by a linear structural equation model (SEM) $X_v = \sum u: u \rightarrow v \gamma_{uv} X_u + \epsilon_u$, $E\epsilon_u = 0$, $0 < \text{var} \epsilon_u < \infty$. Under causal sufficiency, the errors are mutually independent (no $i \leftrightarrow j$ in the path diagram).
Suppose $D$ is the underlying causal DAG. $D$ is unknown.
Suppose $\mathcal{D}$ is the underlying causal DAG. $\mathcal{D}$ is unknown.

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

$$X_v = \sum_{u: u \rightarrow v} \gamma_{uv} X_u + \epsilon_v, \quad \mathbb{E} \epsilon_u = 0, \quad 0 < \text{var} \epsilon_u < \infty.$$
Suppose $\mathcal{D}$ is the underlying causal DAG. $\mathcal{D}$ is unknown.

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

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

Under causal sufficiency, the errors are mutually independent (no $i \leftrightarrow j$ in the path diagram).
Suppose we want to estimate the total (causal) effect of $A$ on $Y$. 

The total effect $\tau_{AY}$ is defined as the slope of $x_a \mapsto E[X \mid do(X_A = x_a)]$, given by a sum-product of Wright (1934):

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

Here we consider point intervention ($|A| = 1$) for simplicity. For a joint intervention ($|A| > 1$), total effect can be similarly defined.
Suppose we want to estimate the total (causal) effect of $A$ on $Y$. The total effect $\tau_{AY}$ is defined as the slope of $x \mapsto \mathbb{E}[X|\text{do}(X=A=x)]$, given by a sum-product of Wright (1934): $\tau_{AY} = \partial_{x} \mathbb{E}[X|\text{do}(X=A=x)] = (\gamma_{AZ}\gamma_{ZW} + \gamma_{AW})\gamma_{WY}$. Here we consider point intervention ($|A| = 1$) for simplicity. For a joint intervention ($|A| > 1$), total effect can be similarly defined.
Suppose we want to estimate the total (causal) effect of $A$ on $Y$.

The total effect $\tau_{AY}$ 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):

$$\tau_{AY} = \frac{\partial}{\partial x_a} \mathbb{E}[X_Y|\text{do}(X_A = x_a)] = (\gamma_{AZ}\gamma_{ZW} + \gamma_{AW})\gamma_{WY}.$$
Suppose we want to estimate the total (causal) effect of $A$ on $Y$.

The total effect $\tau_{AY}$ 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):

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

Here we consider point intervention ($|A| = 1$) for simplicity. For a joint intervention ($|A| > 1$), total effect can be similarly defined.
Without making further assumptions, the causal DAG $D$ can only be identified from observed distribution up to a Markov equivalence class. The Markov equivalence class of $D$ is uniquely represented by a CPDAG/essential graph $C$. Knowing only $C$ is often insufficient to identify the total effect.
Without making further assumptions, the causal DAG $\mathcal{D}$ can only be identified from observed distribution up to a Markov equivalence class.
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 $\mathcal{C}$. 
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.
The total effect $\tau_{AY}$ 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.
Theorem (Perković, 2020)

The total effect $\tau_{AY}$ 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.
Theorem (Perković, 2020)

The total effect $\tau_{AY}$ is identified from a maximally oriented partially directed acyclic graph $G$ if and only if there is no proper, possibly causal path from $A$ to $Y$ in $G$ 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.
However, often we have additional knowledge that can help towards identification.
However, often we have additional knowledge that can help towards identification.

Suppose we know that \( S \text{ temporally precedes } A \).
However, often we have additional knowledge that can help towards identification.

Suppose we know that $S$ temporally precedes $A$. 

The green orientations are further implied by the rules of Meek (1995). In this example, $\tau_{AY}$ is identified from the resulting maximally oriented partially directed acyclic graph (MPDAG) $G$. 

---

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

Suppose we know that $S$ temporally precedes $A$. 

The green orientations are further implied by the rules of Meek (1995). In this example, $\tau_{AY}$ is identified from the resulting maximally oriented partially directed acyclic graph (MPDAG) $G$. 
However, often we have additional knowledge that can help towards identification.

Suppose we know that $S$ temporally precedes $A$.

The green orientations are further implied by the rules of Meek (1995).
However, often we have additional knowledge that can help towards identification.

Suppose we know that $S$ temporally precedes $A$.

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

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

\[ A \rightarrow Z \rightarrow W \rightarrow Y \rightarrow T \]
Our task is to estimate $\tau_{AY}$ from $n$ iid observational sample generated by a linear SEM associated with causal DAG $D$, given that $D \in [G]$ for MPDAG $G$, $\tau_{AY}$ is identifiable from $G$.

Adjustment estimator

$\hat{\tau}_{adj}^{AY}$ is the least squares coefficient of $A$ from $Y \sim A + S$.
Our task is to estimate $\tau_{AY}$ from $n$ iid observational sample generated by a linear SEM associated with causal DAG $D$, given that $D \in [G]$ for MPDAG $G$, $\tau_{AY}$ is identifiable from $G$.

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

MPDAG $G$
Adjustment estimator

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

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

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 $Y \sim A + S$ 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 $[G]$.

Adjustment estimator

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

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

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 $[G]$.

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 $[G]$.

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: $G$-regression estimator

We achieve efficient estimation by exploiting the “additional” conditional independences in $G$ in this over-identified setting.

\[ \hat{\tau}_{AY}^G = \hat{\lambda}_{AW} \hat{\lambda}_{WY}, \]

where $\hat{\lambda}_{AW}$, $\hat{\lambda}_{WY}$ are taken from $W \sim A$ and $Y \sim W + S$ respectively.
Our proposal: $G$-regression estimator

$n = 100$, $t_5$ errors.
Define the set of vertices $D := \text{An}(Y, G_{V\setminus A})$. \textit{G-regression estimator} is

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

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

---

\textbf{Theorem} 

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

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

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

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

**Theorem**

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

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

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

**Theorem**

$G$-regression estimator is

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

$$\hat{r}_{A,Y}^G := \hat{\Lambda}_{A,D}^G \left( (I - \hat{\Lambda}_{D,D}^G)^{-1} \right)_{D,Y},$$

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

**Theorem**

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

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

where $\hat{\Lambda}^{\mathcal{G}}$ is an $|V| \times |V|$ 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?
Define the set of vertices $D := \text{An}(Y, \mathcal{G}_{V \setminus A})$. $\mathcal{G}$-regression estimator is
\[
\hat{\tau}_{AY}^\mathcal{G} := \hat{\Lambda}_{A,D}^\mathcal{G} \left[ (I - \hat{\Lambda}_{D,D}^\mathcal{G})^{-1} \right]_{D,Y},
\]
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.
Define the set of vertices $D := \text{An}(Y, \mathcal{G}_{V\setminus A})$. \textit{\mathcal{G}-regression estimator} is

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

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

\textbf{Theorem}

\textit{\mathcal{G}-regression estimator is}

1. \textbf{complete},

2. the \textbf{most efficient} estimator among all consistent, regular estimators that only depend on the \textit{first two moments} of data.

\begin{itemize}
  \item \textit{How to derive this estimator?}
  \begin{itemize}
    \item 1. Find the MLE under Gaussian errors.
    \item 2. Show that this MLE is “efficient” even when errors are non-Gaussian.
  \end{itemize}
\end{itemize}
Let “buckets” be the maximal connected components of the undirected part of $G$. Further, buckets can be topologically ordered by the directed part of $G$: $B_1 = \{S\}$, $B_2 = \{A\}$, $B_3 = \{Z, W, T\}$, $B_4 = \{Y\}$.

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

![Graph diagram with nodes labeled A, Z, W, Y, S, T, and arrows indicating connections between nodes.](image-url)
Let “buckets” be the maximal connected components of the undirected part of $G$. 

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

**Lemma: Restrictive property**

For each bucket $B_i$, vertices in $B_i$ have the same set of external parents, denoted as $\text{Pa}(B_i)$. 
Let “buckets” be the maximal connected components of the undirected part of $G$.

**Lemma:** Restrictive property

For each bucket $B_i$, vertices in $B_i$ have the same set of external parents, denoted as $\text{Pa}(B_i)$. 
Let “buckets” be the maximal connected components of the undirected part of $G$.

Further, buckets can be topologically ordered by the directed part of $G$:

\[ B_1 = \{S\}, \quad B_2 = \{A\}, \quad B_3 = \{Z, W, T\}, \quad B_4 = \{Y\}. \]
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 $\text{Pa}(B_i)$. 
The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k), B_k}^T X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$  

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

Two nice things happen under this reparametrization:

1. With $D = \text{An}(Y, G_{V \setminus A})$, $\tau_{\text{AY}}$ can be identified as $\tau_{\text{AY}} = \Lambda A$, $D [ (I - \Lambda D, D) ]^{-1} D Y$.

The bucket-wise error distribution is nuisance.

2. Under Gaussian errors, the MLE for each $\Lambda_{\text{Pa}(B_k), B_k}$, $B_k$ is just the least squares coefficients of $B_k \sim \text{Pa}(B_k)$. $G$-regression.
The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k), B_k}^T X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$  

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

Two nice things happen under this reparametrization:
The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k), B_k}^\top X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$ 

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

Two nice things happen under this reparametrization:

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

$$\tau_{AY} = \Lambda_{A,D} \left[ (I - \Lambda_{D,D})^{-1} \right]_{D,Y}. $$
The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k), B_k}^T X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$ 

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

Two nice things happen under this reparametrization:

1. With $D = \text{An}(Y, \mathcal{G}_{V \setminus A})$, $\tau_{AY}$ 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.
The SEM according to $\mathcal{D}$ can be reparametrized as a block-recursive form according to the buckets:

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k),B_k}^T X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$  

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

Two nice things happen under this reparametrization:

1. With $D = \text{An}(Y, \mathcal{G}_{V\setminus A})$, $\tau_{AY}$ 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 $\Lambda_{\text{Pa}(B_k),B_k}$ is just the least squares coefficients of $B_k \sim \text{Pa}(B_k)$.  

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

$$X_{B_1} = \varepsilon_{B_1}, \quad X_{B_k} = \Lambda_{\text{Pa}(B_k),B_k}^T X_{\text{Pa}(B_k)} + \varepsilon_{B_k}, \quad k = 2, \ldots, K.$$ 

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

Two nice things happen under this reparametrization:

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

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

   The bucket-wise error distribution is nuisance.

2. Under Gaussian errors, the MLE for each $\Lambda_{\text{Pa}(B_k),B_k}$ is just the least squares coefficients of $B_k \sim \text{Pa}(B_k)$. $\mathcal{G}$-regression.
The second property is a special case of “seemingly unrelated regression” due to the **restrictive property**.

\[(X_Z, X_W, X_T) = (\lambda_{AZ}, \lambda_{AW}, \lambda_{AT})X_A + \varepsilon_{B_3},\]

\[\varepsilon_{B_3} \sim N(0, \Omega_3), \quad (\Omega_3)_{ZT \cdot W} = 0.\]
The second property is a special case of “seemingly unrelated regression” due to the \textit{restrictive property}.

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

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

Let $\Sigma_n$ be the sample covariance. Consider the class of estimators

$$T = \left\{ \hat{\tau}(\Sigma_n) : \mathbb{R}^{V \times V}_{\text{PD}} \rightarrow \mathbb{R}^{|A|} : \hat{\tau}(\Sigma_n) \text{ is a consistent, asymptotically linear estimator of } \tau_{AY} \right\}.$$
Let $\Sigma_n$ be the sample covariance. Consider the class of estimators

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

The efficiency theory entails two parts.
Efficiency theory

Let $\Sigma_n$ be the sample covariance. Consider the class of estimators

$$\mathcal{T} = \left\{ \hat{\tau}(\Sigma_n) : \mathbb{R}_{\text{PD}}^{|V| \times |V|} \rightarrow \mathbb{R}^{|A|} : \right\}$$

$\hat{\tau}(\Sigma_n)$ is a consistent, asymptotically linear estimator of $\tau_{AY}$.

The efficiency theory entails two parts.

1. 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**

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

   associated with $\bar{G}$, where $\bar{G}$ is the saturated version of $G$. 

2. Verifying that $\hat{\tau}_G$ achieves this bound.

This generalizes a result from Drton (2018).
Efficiency theory

Let $\Sigma_n$ be the sample covariance. Consider the class of estimators

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

The efficiency theory entails two parts.

1. 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

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

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

2. This generalizes a result from Drton (2018).
Efficiency theory

Let $\Sigma_n$ be the sample covariance. Consider the class of estimators

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

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

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

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

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

1. Suppose $|A| = 1$. Rewrite $\hat{\tau} \in T$ as $\hat{\tau}(\Sigma_n) = \hat{\tau}(\hat{\Lambda}_k, G, \hat{\Lambda}_k, G_c, \hat{\Omega}_k )$, where $(\hat{\Lambda}_k, G_c) = (\hat{\Lambda}_k, \bar{G} \setminus G)$ are introduced dashed edges.

2. Consistency of $\hat{\tau}$ implies $\partial \hat{\tau} / \partial \hat{\Lambda}_k, G = \partial \tau / \partial \hat{\Lambda}_k, G (k = 2, \ldots, K)$, $\partial \hat{\tau} / \partial \hat{\Omega}_k = 0 (k = 1, \ldots, K)$, but $\partial \hat{\tau} / \partial \hat{\Lambda}_k, G_c$ is free.

3. Compute acov of $(\hat{\Lambda}_k, G)_{k = 1, \ldots, K}, (\hat{\Lambda}_k, G_c)_{k = 1, \ldots, K}$ via asymptotic linear expansions.

4. By the delta method, an upper bound can be derived from quadratic form $\text{avar}(\hat{\tau}) = \frac{1}{2} \begin{pmatrix} \partial \hat{\tau} / \partial (\hat{\Lambda}_k, G) & \partial \hat{\tau} / \partial (\hat{\Lambda}_k, G_c) \end{pmatrix} \text{acov}((\hat{\Lambda}_k, G)_{k = 1, \ldots, K}, (\hat{\Lambda}_k, G_c)_{k = 1, \ldots, K}) \begin{pmatrix} \partial \hat{\tau} / \partial (\hat{\Lambda}_k, G) & \partial \hat{\tau} / \partial (\hat{\Lambda}_k, G_c) \end{pmatrix}$.
Proof sketch

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

$$\hat{\tau}(\Sigma_n) = \hat{\tau}\left((\hat{\Lambda}_k)_k, (\hat{\Lambda}_k)_k, (\hat{\Omega}_k)_k\right),$$

where $(\hat{\Lambda}_k)_k, (\hat{\Lambda}_k)_k$ are introduced dashed edges.
1. Suppose $|A| = 1$. Rewrite $\hat{\tau} \in \mathcal{T}$ as

$$\hat{\tau}(\Sigma_n) = \hat{\tau}\left((\hat{\Lambda}_k)_k, (\hat{\Lambda}_k)_k, (\hat{\Omega}_k)_k\right),$$

where $(\hat{\Lambda}_k)_k, (\hat{\Lambda}_k)_k, (\hat{\Omega}_k)_k$ are introduced dashed edges.

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

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

but $\frac{\partial \hat{\tau}}{\partial \hat{\Lambda}_k, \mathcal{G}_c}$ is free.
1. Suppose $|A| = 1$. Rewrite $\hat{\tau} \in \mathcal{T}$ as

$$\hat{\tau}(\Sigma_n) = \hat{\tau}\left((\hat{\Lambda}_k)_k, (\hat{\Lambda}_k)_k, (\hat{\Omega}_k)_k\right),$$

where $(\hat{\Lambda}_k)_k, (\hat{\Omega}_k)_k$ are introduced dashed edges.

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

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

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 $\hat{\tau} \in T$ as
   \[
   \hat{\tau}(\Sigma_n) = \hat{\tau} \left( (\hat{\Lambda}_k)_k, (\hat{\Lambda}_k, G), (\hat{\Omega}_k)_k \right),
   \]
   where $(\hat{\Lambda}_k)_k, g_c = (\hat{\Lambda}_k)_k, \bar{\bar{g}} \setminus g$ are introduced dashed edges.

2. Consistency of $\hat{\tau}$ implies
   \[
   \frac{\partial \hat{\tau}}{\partial \hat{\Lambda}_k, g} = \frac{\partial \tau_g}{\partial \hat{\Lambda}_k, g} \quad (k = 2, \ldots, K), \quad \frac{\partial \hat{\tau}}{\partial \hat{\Omega}_k} = 0 \quad (k = 1, \ldots, K),
   \]
   but $\frac{\partial \hat{\tau}}{\partial \hat{\Lambda}_k, g_c}$ is free.

3. Compute acov of $\left( (\hat{\Lambda}_k, g)_k, (\hat{\Lambda}_k, g_c)_k \right)$ via asymptotic linear expansions.

4. By the delta method, an upper bound can be derived from quadratic form
   \[
   \text{avar}(\hat{\tau}) = \left( \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_k, g)_k} \right)^T \text{acov} \left( (\hat{\Lambda}_k, g)_k, (\hat{\Lambda}_k, g_c)_k \right) \left( \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_k, g)_k} \right) \leq \sup_{\frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_k, g_c)_k}} \left( \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_k, g)_k} \right)^T \text{acov} \left( (\hat{\Lambda}_k, g)_k, (\hat{\Lambda}_k, g_c)_k \right) \left( \frac{\partial \hat{\tau}}{\partial (\hat{\Lambda}_k, g)_k} \right). \]
Simulation results

An instance is simulated by the following steps.

1. Draw $D$ from a random graph ensemble.
2. Take $G = \text{CPDAG}(D)$.
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 $\tau A Y$ is identified from $G$.
5. Compute squared error $\|\tau A Y - \hat{\tau} A Y\|^2$.

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

An instance is simulated by the following steps.

1. Draw $D$ 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 $\mathcal{G} = \text{CPDAG}(\mathcal{D})$. 
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 $\mathcal{G} = \text{CPDAG}(\mathcal{D})$.
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 $\tau_{AY}$ is identified from $\mathcal{G}$.

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

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 $\tau_{AY}$ 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

An instance is simulated by the following steps.

1. Draw $D$ from a random graph ensemble.
2. Take $G = \text{CPDAG}(D)$.
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 $\tau_{AY}$ is identified from $G$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

We compare to the following estimators in the literature:

- adj.0: 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 $\tau_{AY}$ 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.0: 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 $\tau_{AY}$ is identified from $\mathcal{G}$.
5. Compute squared error $\|\tau_{AY} - \hat{\tau}_{AY}\|^2$.

We compare to the following estimators in the literature:

- \texttt{adj.0}: optimal adjustment estimator (Henckel, Perković, and Maathuis, 2019),
- \texttt{IDA.M}: joint-IDA estimator based on modifying Cholesky decompositions (Nandy, Maathuis, and Richardson, 2017),
- \texttt{IDA.R}: joint-IDA estimator based on recursive regressions (Nandy, Maathuis, and Richardson, 2017).
Table 1: Percentage of identified instances not estimable using contending estimators. All instances are estimable with $G$-regression.

| Estimator | $|A|$ | $|V| = 20$ | $|V| = 50$ | $|V| = 100$ |
|-----------|------|----------|----------|----------|
| adj.0     | 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%      |
Table 2: Geometric average of squared errors relative to $G$-regression, computed from estimable instances.

| | $|V| = 20$ | $|V| = 50$ | $|V| = 100$ |
|---|---|---|---|
| | $n = 100$ | $n = 1000$ | $n = 100$ | $n = 1000$ | $n = 100$ | $n = 1000$ |
| adj.o | 1.3 | 1.3 | 1.4 | 1.3 | 1.5 | 1.5 |
| | 3.4 | 4.2 | 4.7 | 4.9 | 4.2 | 4.5 |
| | 6.3 | 5.9 | 7.4 | 7.2 | 7.8 | 8.0 |
| | 9.3 | 9.3 | 12 | 14 | 12 | 12 |
| IDA.M | 20 | 19 | 61 | 48 | 103 | 108 |
| | 62 | 65 | 220 | 182 | 293 | 356 |
| | 93 | 119 | 354 | 396 | 749 | 771 |
| | 154 | 222 | 533 | 895 | 1188 | 1604 |
| IDA.R | 20 | 19 | 61 | 48 | 103 | 108 |
| | 33 | 38 | 121 | 113 | 176 | 199 |
| | 30 | 39 | 171 | 135 | 342 | 312 |
| | 48 | 50 | 187 | 214 | 405 | 432 |
Final remarks

• Details: arxiv.org/abs/2008.03481
• R package eff: github.com/richardkwo/eff

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).
Final remarks

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

- **Details**: arxiv.org/abs/2008.03481
- **R package** eff²: github.com/richardkwo/eff2
Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** eff²: 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).
Final remarks

- **Details**: arxiv.org/abs/2008.03481
- **R package** eff²: 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).


Meek’s rules

The orientation rules from Meek (1995).