Causal Inference in Finite Samples: The Potential of Invalid Adjustments

Nadja Rutsch[1], Sara Magliacane[2], Stéphanie van der Pas[1]

[1] Vrije Universiteit Amsterdam
[2] Universiteit van Amsterdam

April 10, 2025

Funded by the European Union

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe program under Grant agreement No. 101074082. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Introduction

Research | Open access | Published: 14 November 2024

F. prausnitzii potentially modulates the association between citrus intake and depression

Chatpol Samuthpongtorn, Allison A. Chan, Wenjie Ma, Fenglei Wang, Long H. Nguyen, Dong D. Wang, Olivia I. Okereke, Curtis Huttenhower, Andrew T. Chan & Raaj S. Mehta

Microbiome **12**, Article number: 237 (2024) | Cite this article

Results

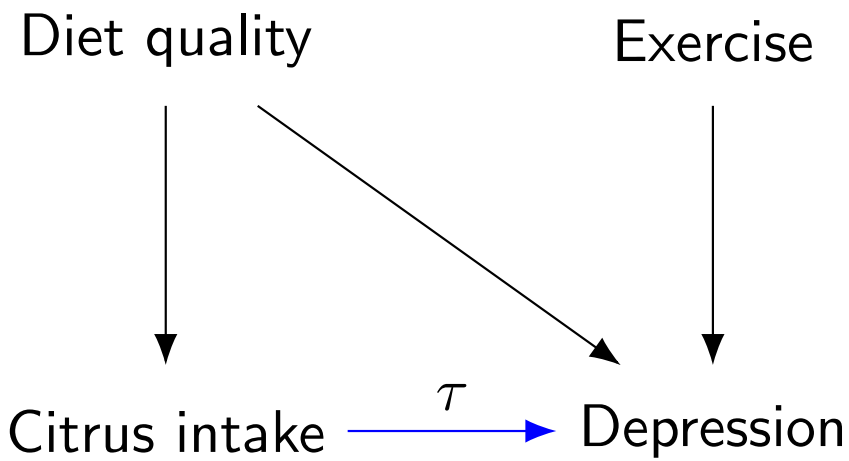
Citrus consumption is prospectively associated with decreased risk of depression

From 2003 through 2017, we identified 2173 cases of depression among 32,427 women free of self-reported physician/clinician-diagnosed depression and regular use of antidepressants at baseline. Over 222,923 person-years of follow-up [...].

Discussion

We acknowledge that our study has several limitations. First, being observational, we cannot directly infer causal relationships, and the possibility of residual confounders remains. Nevertheless, we controlled for numerous variables including age, BMI, exercise, and diet quality in our statistical analysis and found that antidepressant use did not significantly influence the results. [...]

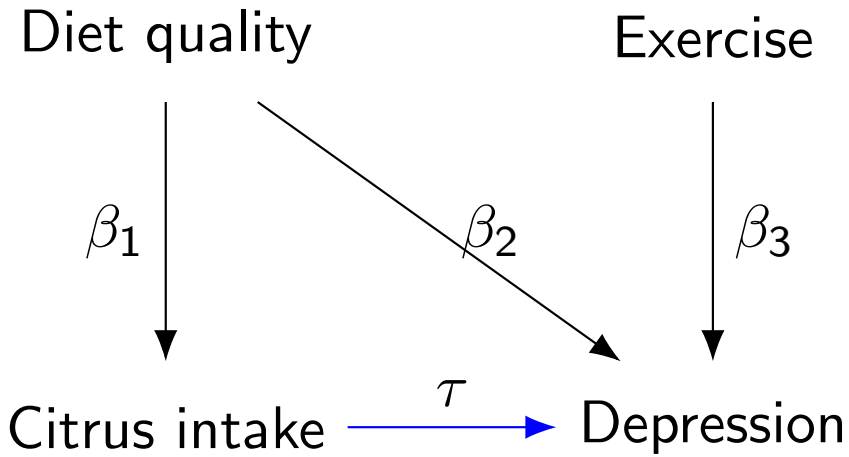
Causal graph



Confounding path: Citrus intake $\leftarrow$ Diet quality $\rightarrow$ Depression

To estimate the **causal effect** τ of Citrus intake on Depression, we need to break the confounding path.

Treatment effect estimation

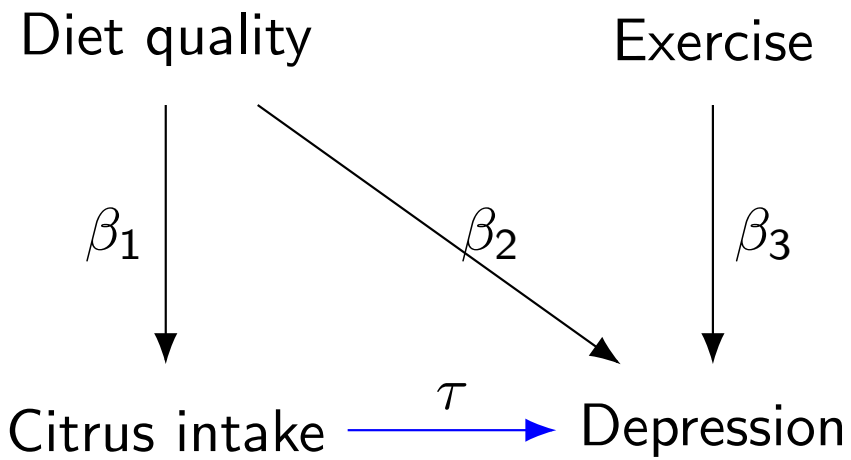


Causal linear model (ground truth):

$$\text{Citrus intake} = \beta_1 \cdot \text{Diet quality} + \epsilon_1$$

$$\text{Depression} = \tau \cdot \text{Citrus intake} + \beta_2 \cdot \text{Diet quality} + \beta_3 \cdot \text{Exercise} + \epsilon_2$$

Treatment effect estimation



Causal linear model (estimation):

$$\text{Depression} = \hat{\tau} \cdot \text{Citrus intake} + \hat{\beta}_2 \cdot \text{Diet quality} + \hat{\beta}_3 \cdot \text{Exercise} + \hat{\epsilon}_2$$

- Adjustment set: $\{\text{Diet quality}, \text{Exercise}\}$
- Including Diet quality gives an **unbiased** estimate of τ

But: what about the **variance**?

Setting

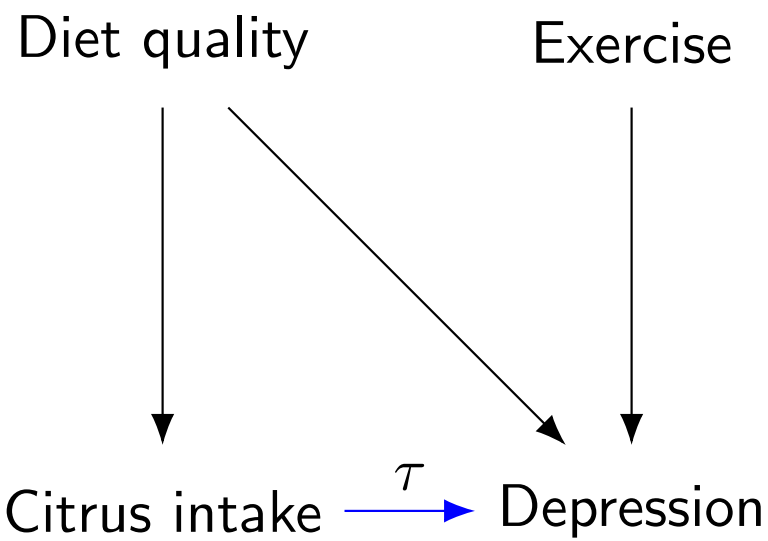
- causal model is linear Gaussian
- OLS estimator
- only pre-treatment variables
- baseline: optimal valid adjustment set $\mathbf{O}$ [Henckel et al., 2022]

Asymptotically optimal adjustment set

Henckel et al. (2022) defined the asymptotically optimal adjustment set $\mathbf{O}$:

$$\mathbf{O} \equiv \mathrm{Pa}(\mathbf{M} \cup Y) \setminus (\mathbf{M} \cup A)$$

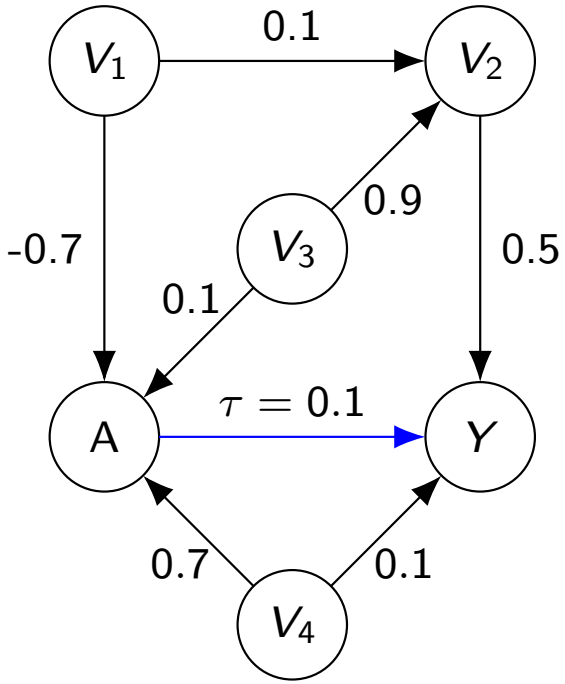
Pa(.) : parents $\quad$ $\mathbf{M}$: mediators $\quad$ A : treatment $\quad$ Y : outcome



$$\mathbf{O} = \{\text{Diet quality}, \text{Exercise}\}$$

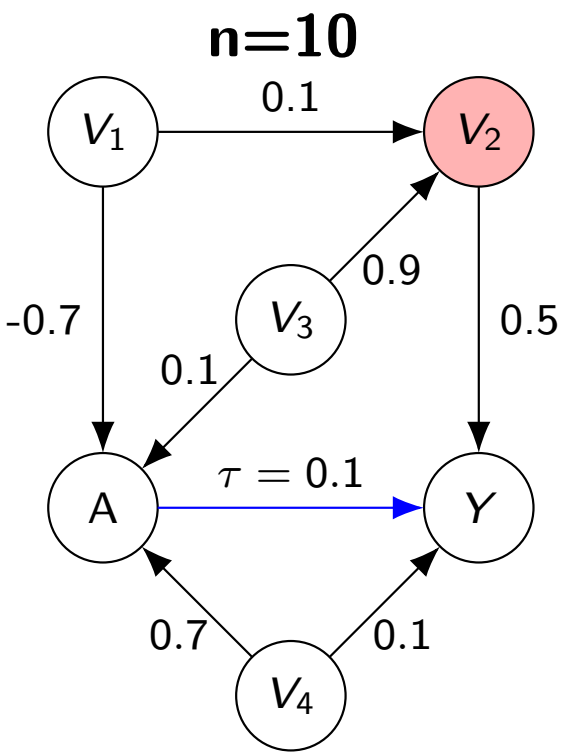
Example

Question: Which covariates should we include?

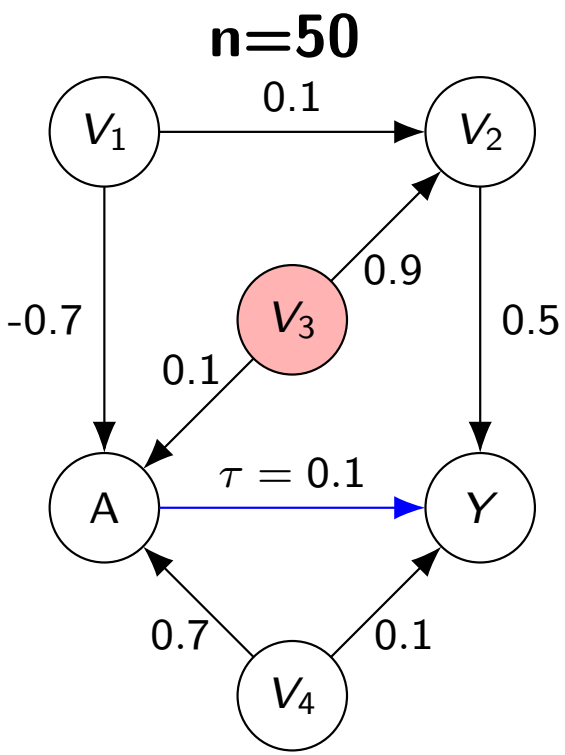


Answer: It depends on the sample size!

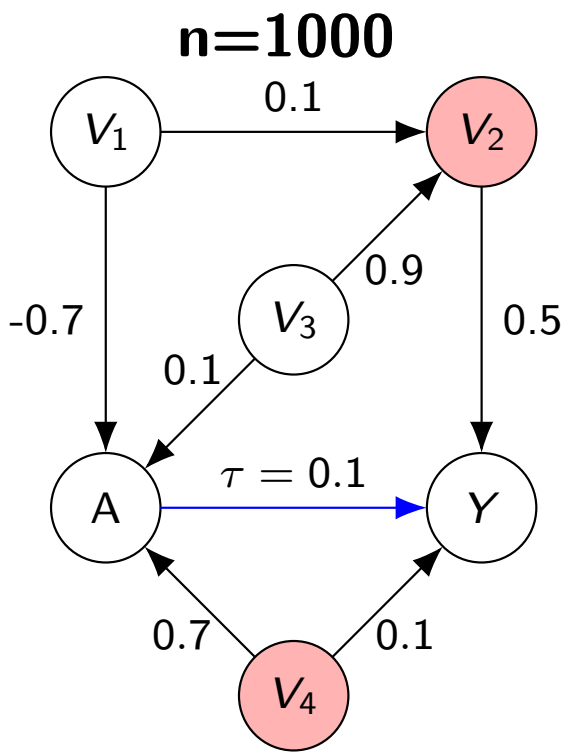
Finite sample-optimal adjustment set



Set	MSE
O	0.286
$\{V_2\}$	0.128



Set	MSE
O	0.031
$\{V_3\}$	0.019



Set	MSE
O	0.001
$= \{V_2, V_4\}$	

MSE-optimal adjustment set

Our definition:

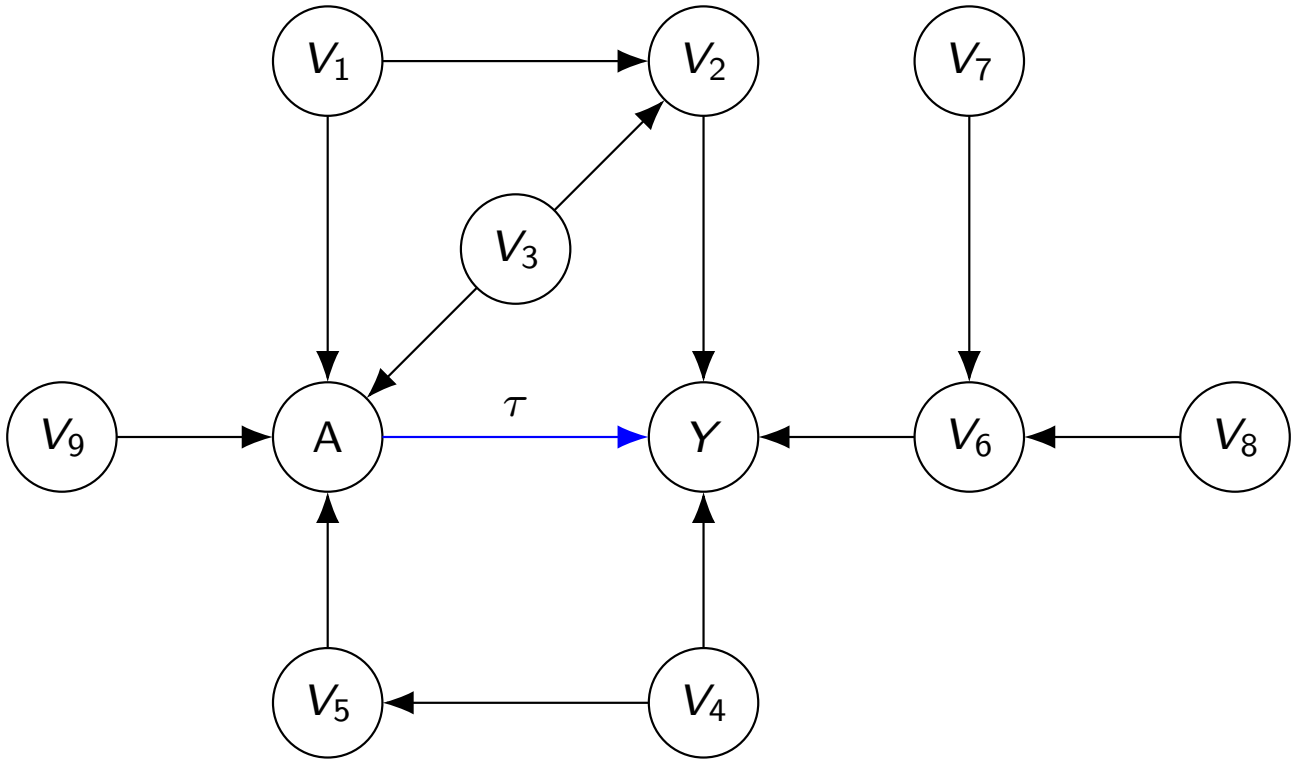
$$O_n(\mathcal{M}, \hat{\tau}_K) = \underset{K \subseteq \mathcal{V} \setminus \{A,Y\}}{\operatorname{argmin}} E_{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n \sim \mathcal{M}}\{(\hat{\tau}_K - \tau)^2\}$$

- $\mathcal{M}$: causal model
- K: adjustment set
- $\hat{\tau}_K$: estimator adjusting for K

How can we find the MSE-optimal adjustment set?

Problem: large search space!

Example



9 covariates $\rightarrow 2^9 = 512$ possible adjustment sets

Can we exclude some of these sets based on the causal graph?

$\rightarrow$ Yes, can be reduced to 18 adjustment sets

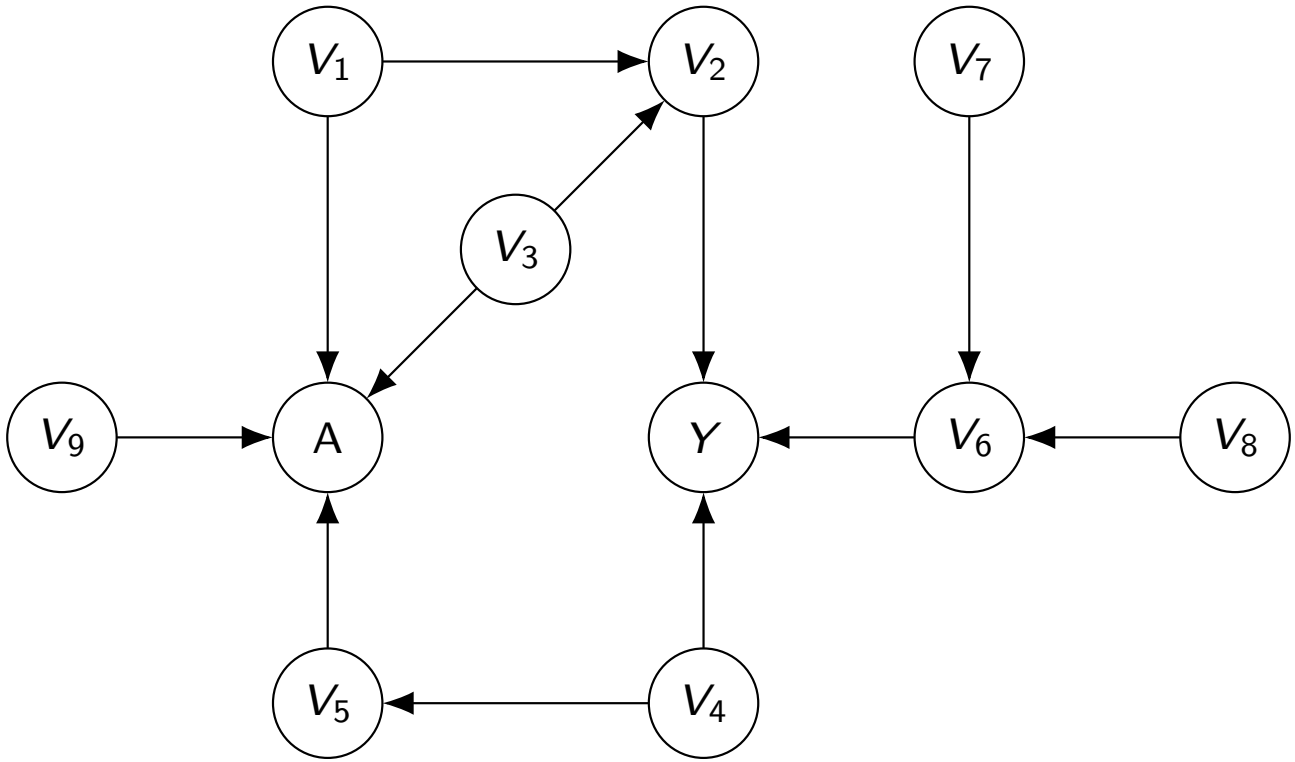
Graphical criteria

1. Exclusion of suboptimal variables
2. Forbidden combinations
3. Suboptimal valid adjustment sets

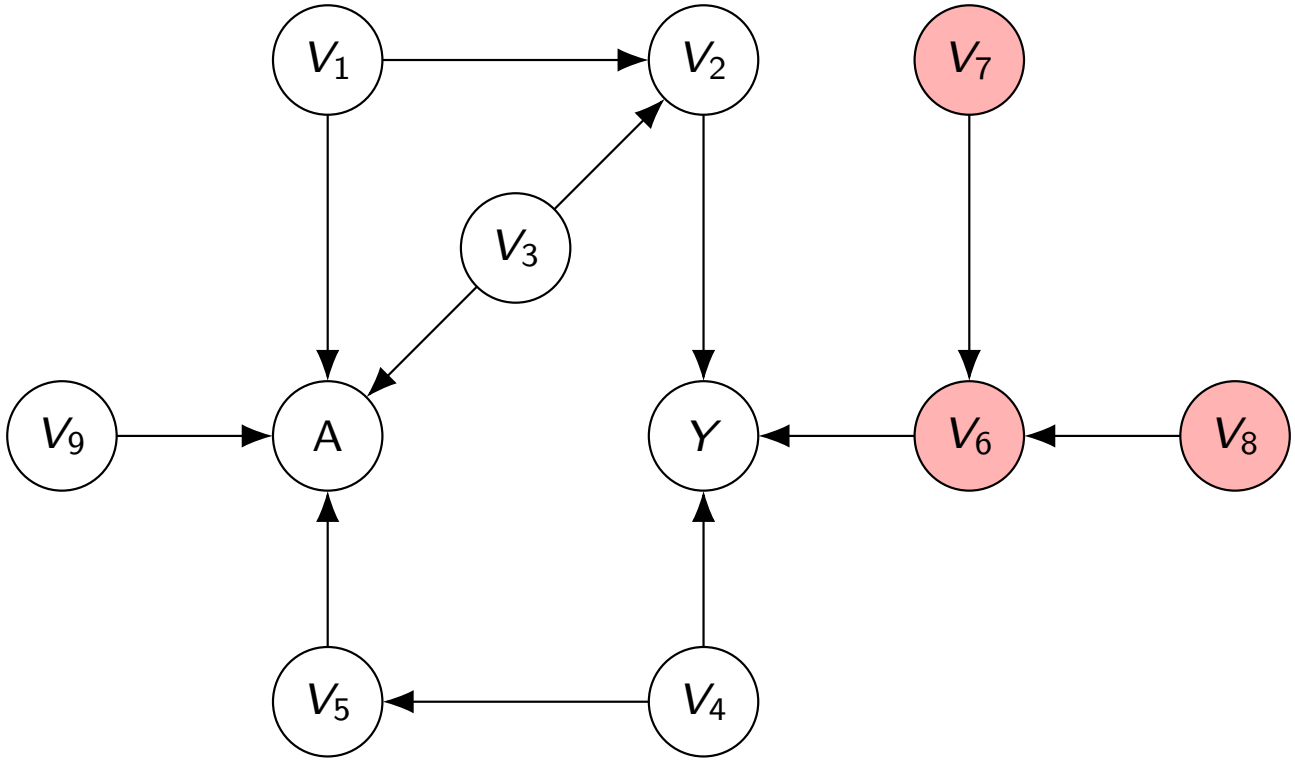
Partitioning of covariates

We define a partitioning of all variables $\mathcal{V} \setminus \{A, Y\}$ in $\mathcal{G}$ on the graph $\mathcal{G}'$:

$$\mathcal{G}' = \mathcal{G} \setminus (A \to Y)$$



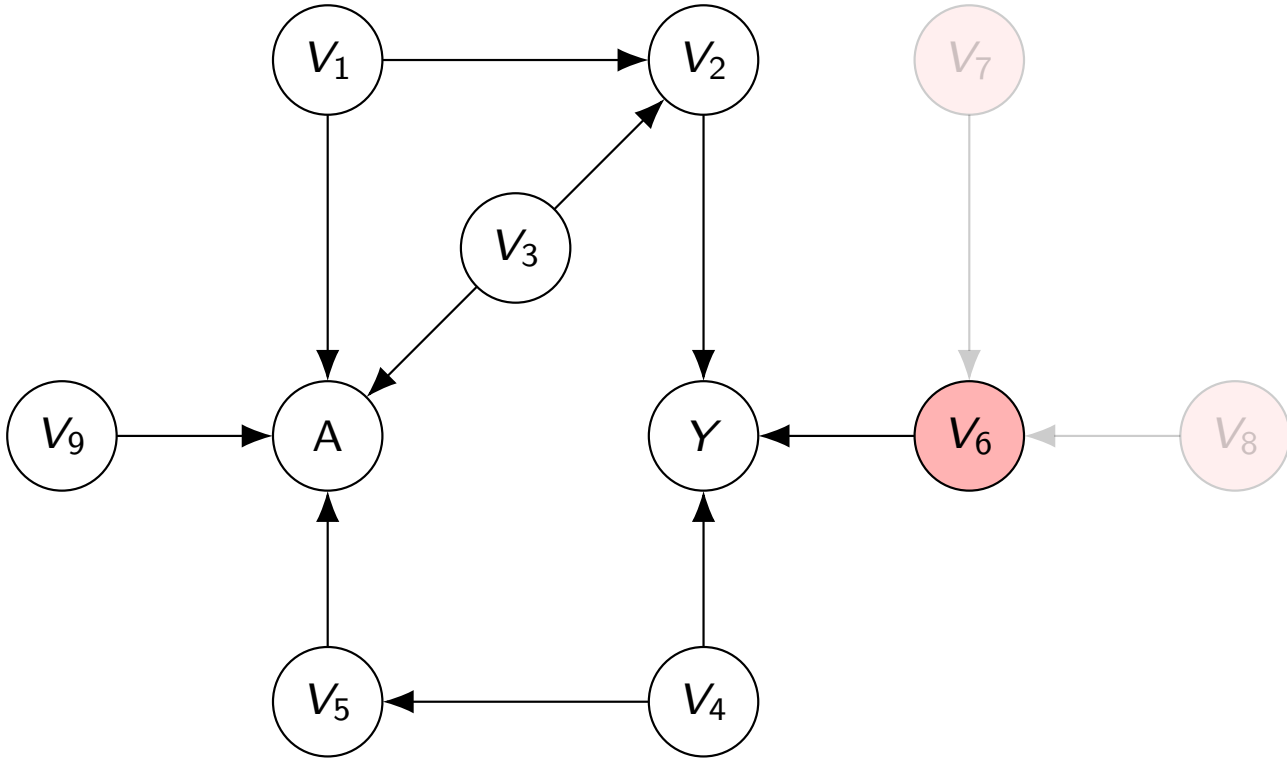
Precision Variables



Precision variables are d-connected to Y and d-separated from A.

$\rightarrow$ Affect only variance, not bias.

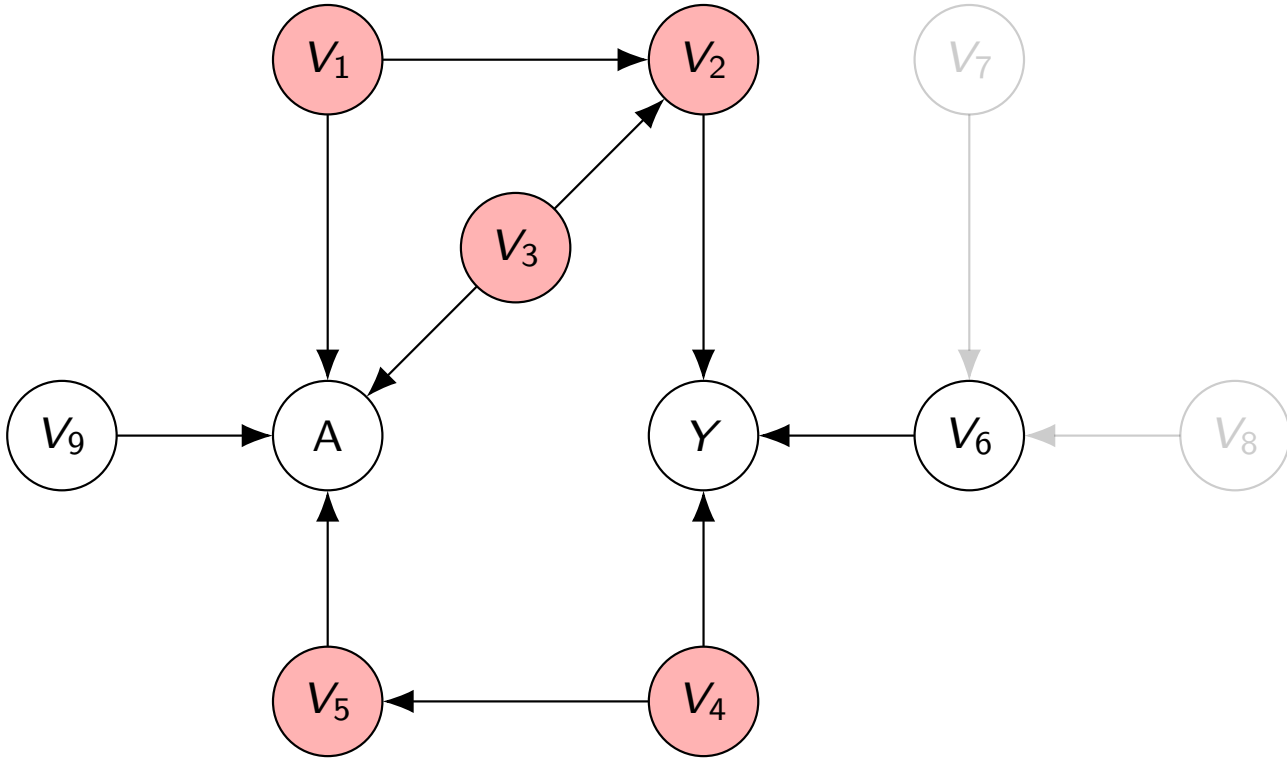
Suboptimal Precision Variables



V_7 and V_8 are d-separated from Y given V_6.

$\rightarrow$ Always better to adjust for V_6 instead.

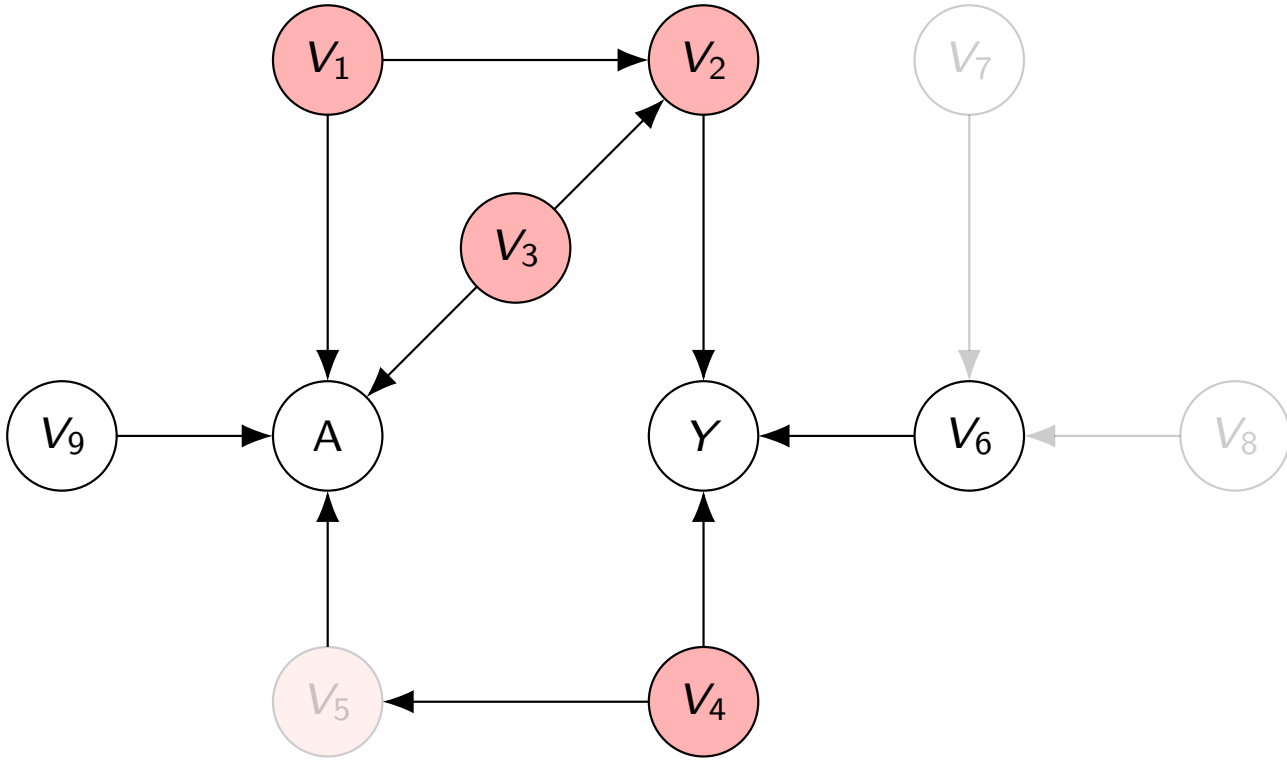
Extended Confounding Variables



Extended confounding variables are d-connected to A and Y.

$\rightarrow$ Affect both variance and bias.

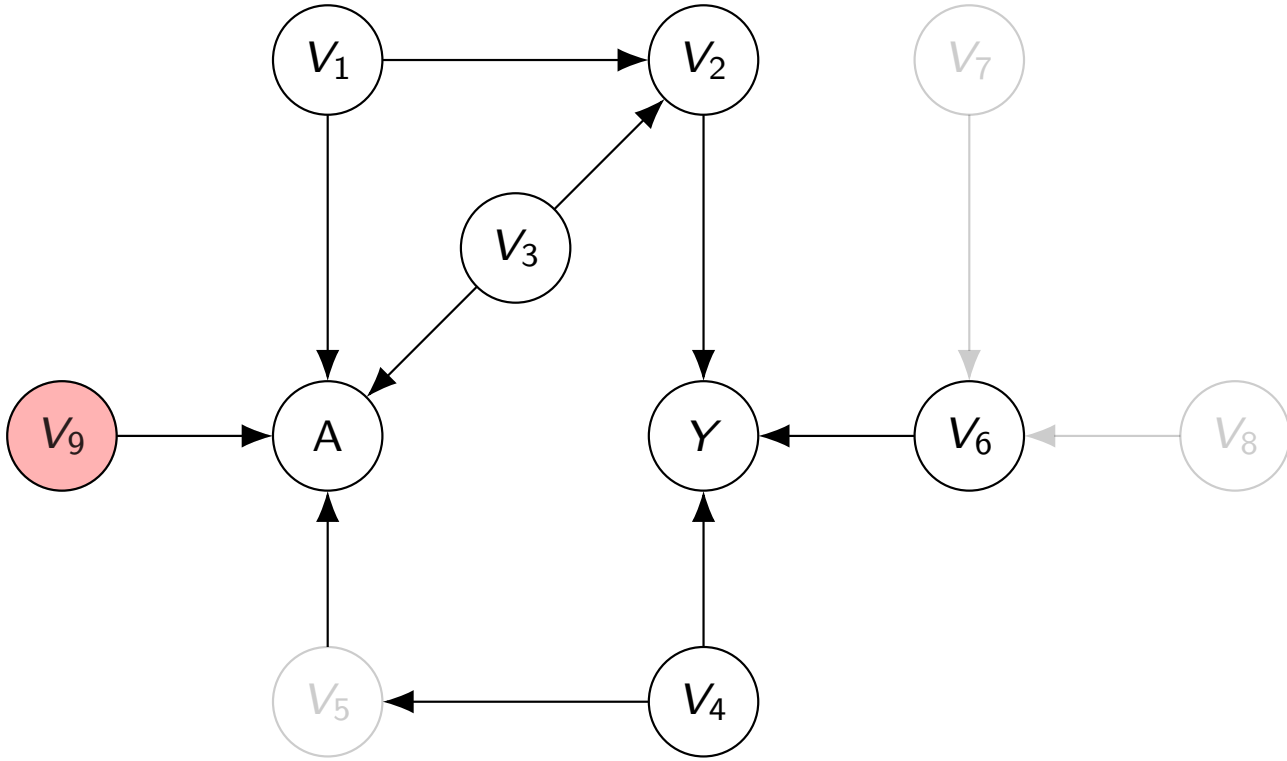
Suboptimal Confounding Variables



V_5 is d-separated from Y given V_4 **and** V_4 is d-separated from A given V_5.

$\rightarrow$ Always better to adjust for V_4 instead of V_5.

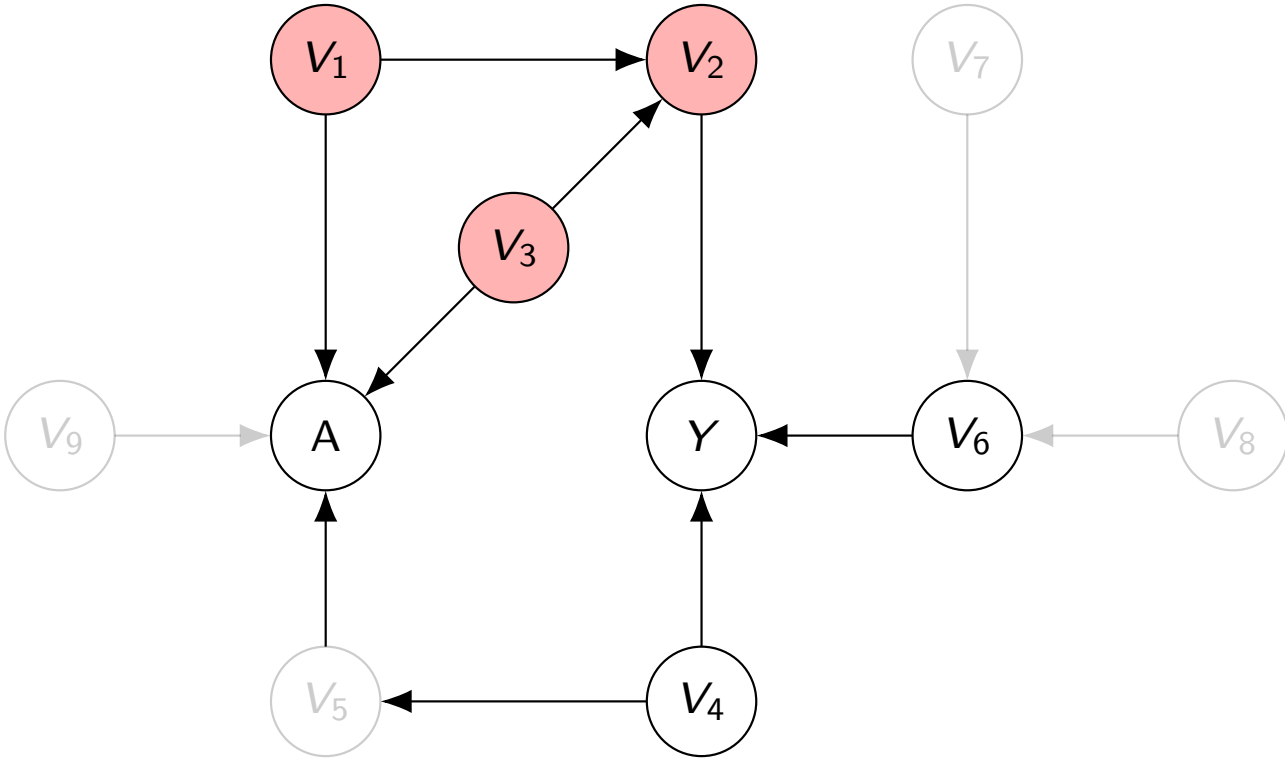
Irrelevant Variables



Irrelevant variables are d-separated from Y.

$\rightarrow$ Always increase MSE.

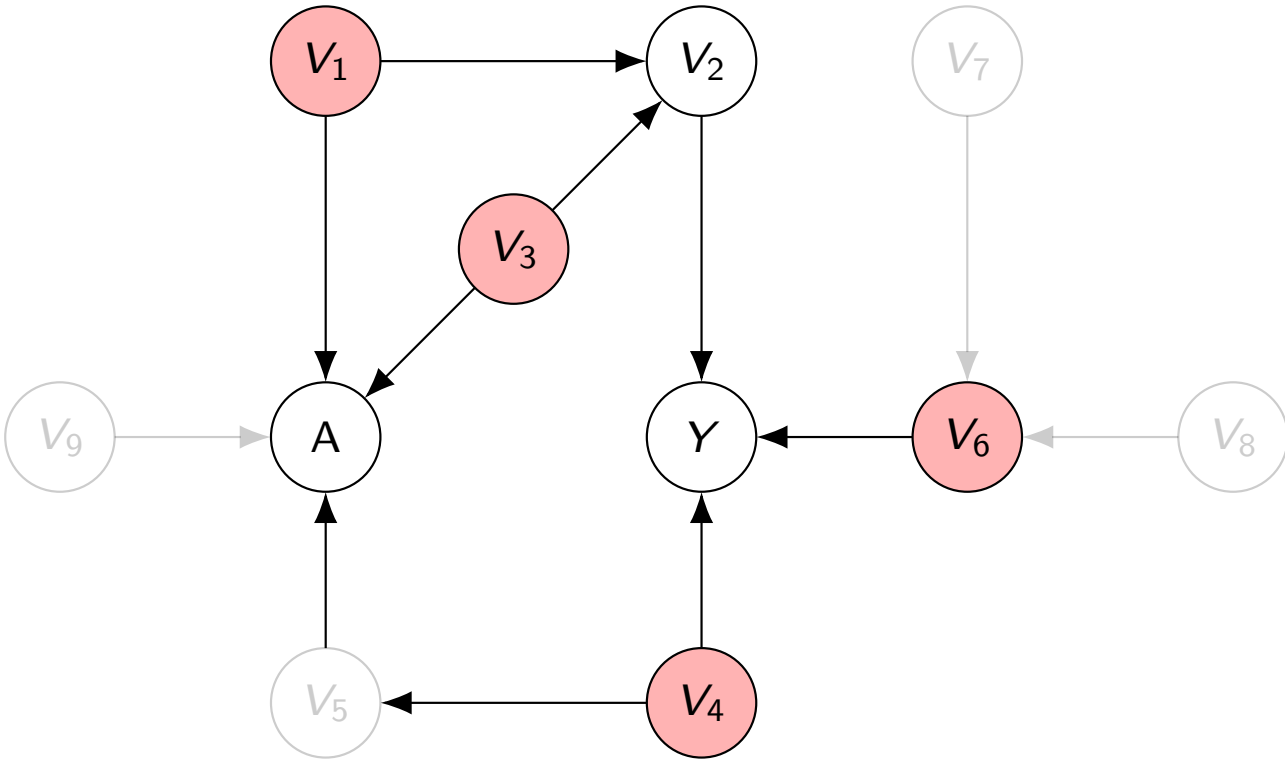
Forbidden Combinations



V_1 and V_3 are d-separated from Y given V_2.

$\rightarrow$ Always better to include $\{V_2\}$ instead of $\{V_1, V_2\}$ or $\{V_3, V_2\}$.

Suboptimal valid adjustment sets



The valid adjustment sets $\{V_1, V_3, V_4, V_6\}$ and $\{V_1, V_3, V_4\}$ are at least as large as $\mathbf{O} = \{V_2, V_4, V_6\}$.

$\rightarrow$ Always better to use $\mathbf{O}$ instead.

Algorithm

How can we find the MSE-optimal adjustment set?

Input: data, causal graph

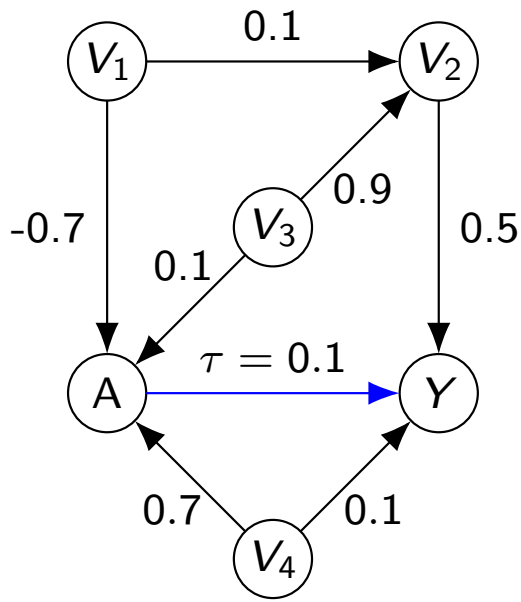
1. **Prune** adjustment sets with graphical criteria
2. For each remaining adjustment set:
   - Estimate the **variance**
   - Estimate the **bias**
3. Select the adjustment set with the **lowest estimated MSE** out of all remaining adjustment sets.

Output: Estimated MSE-optimal adjustment set $\hat{O}_n$

Experiments

Setup:

- Simulated data based on linear Gaussian causal model
- Compare MSE of the **estimated MSE-optimal adjustment set** and the **asymptotically optimal adjustment set** [Henckel et al., 2022]
- Vary sample sizes to observe the impact on MSE.



Results

Sample Size	MSE using O	MSE using $\hat{O}_n$
10	0.2789	**0.2517**
50	0.0317	**0.0293**
100	0.0154	**0.0146**
500	**0.0029**	0.0032
1000	**0.0015**	0.0016

Table: Comparison of MSE for different adjustment sets

- $\hat{O}_n$ outperforms O in small sample sizes.
- As sample size increases, O becomes optimal.

Set with lowest estimated variance

Bias estimation is challenging in small samples.

What happens if we choose the adjustment set with the lowest estimated variance?

Sample Size	MSE	MSE using O	MSE using $\hat{O}_n$
10	**0.1525**	0.2789	0.2517
50	**0.0206**	0.0317	0.0293
100	**0.0125**	0.0154	0.0146
500	0.0063	**0.0029**	0.0032
1000	0.0056	**0.0015**	0.0016

Table: Average MSE for different sample sizes, 10,000 random seeds

Conclusion

- Introduced method to find **MSE-optimal adjustment sets** in finite samples.
- Graphical criteria help reduce the **search space**.
- Experiments demonstrate the potential of **invalid adjustments** in finite samples.

Future work:

- Nonparametric estimators
- Non-linear, non-Gaussian causal models
- Post-treatment variables

References I


Henckel, L., Perković, E., and Maathuis, M. H. (2022).
Graphical Criteria for Efficient Total Effect Estimation Via Adjustment in Causal Linear Models.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(2):579–599.