Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables (#3495)

Prof. Dr. Jakob Runge

October 18, 2021

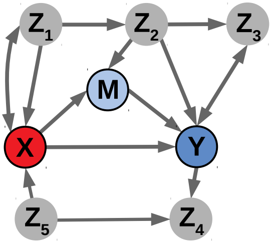
DLR Institute of Data Science and TU Berlin

Knowledge for Tomorrow

Causal inference preliminaries

Task Given a qualitative causal graph and data, estimate causal effect of X on Y [Pearl, 2009]:

$$p(Y \mid do(X = x))$$

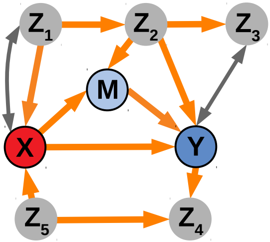


Causal inference preliminaries

Task Given a qualitative causal graph and data, estimate causal effect of X on Y [Pearl, 2009]:

$$p(Y \mid do(X = x))$$

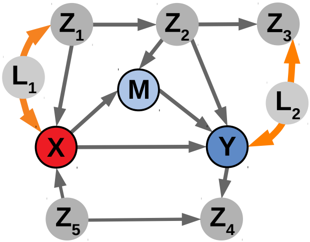
Graph type Acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ with **directed (**$\rightarrow$**)** and *bi-directed* ($\leftrightarrow$) edges representing arbitrary latent confounders



Task Given a qualitative causal graph and data, estimate causal effect of X on Y [Pearl, 2009]:

$$p(Y \mid do(X = x))$$

Graph type Acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ with *directed* ($\rightarrow$) and **bi-directed ($\leftrightarrow$)** edges representing arbitrary latent confounders

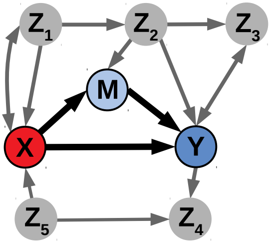


Causal inference preliminaries

Task Given a qualitative causal graph and data, estimate causal effect of X on Y [Pearl, 2009]:

$$p(Y \mid do(X = x))$$

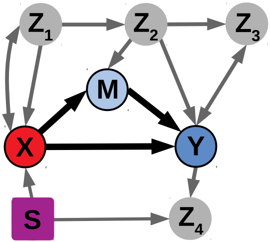
Different types of effects Here total causal effect through direct and indirect path through mediator(s) M



Causal inference preliminaries

Extended task Given a qualitative causal graph and data: Estimate *conditional causal effect* of X on Y given S

$$p(Y \mid do(X = x), S = s)$$

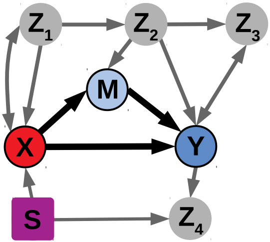


Causal inference preliminaries

Identifiability Effect is *identifiable* if it can be expressed as a function of the observational distribution $p(\mathbf{V})$ [Pearl, 2009]:

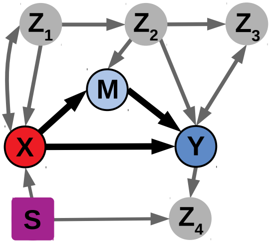
$$p(Y \mid do(X = x), S = s) = q(p(\mathbf{V}))$$

Different approaches: **Backdoor adjustment** / Frontdoor adjustment / General do-calculus



Valid backdoor adjustment sets A set $\mathbf{Z}$ for the total causal effect of X on Y is called *valid* relative to (X, Y) if the interventional distribution for setting $do(X = x)$ factorizes as:

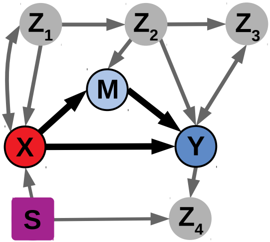
$$p(Y|do(X = x)) = q(p(\mathbf{V})) = \int_{\mathbf{Z}} p(Y|x, \mathbf{z})p(\mathbf{z})d\mathbf{z}$$



Causal inference preliminaries

Generalized backdoor criterion [Perković et al., 2018]: With $\mathbf{forb}(X, Y) = X \cup des(Y\mathbf{M})$ a set $\mathbf{Z}$ is valid if:

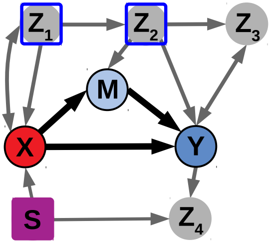
1. $\mathbf{Z} \cap \mathbf{forb} = \emptyset$, and
2. all proper non-causal paths from X to Y are blocked by $\mathbf{Z}$.



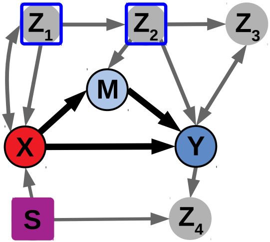
Causal inference preliminaries

Adjust-set [Perković et al., 2018] is valid if and only if a valid set exists:

$$\textbf{vancs}(X, Y, \mathbf{S}) = an(XY\mathbf{S}) \setminus \textbf{forb} \quad (1)$$



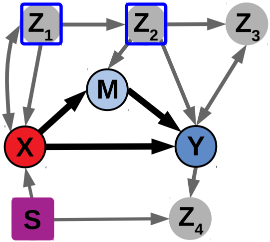
Causal inference preliminaries

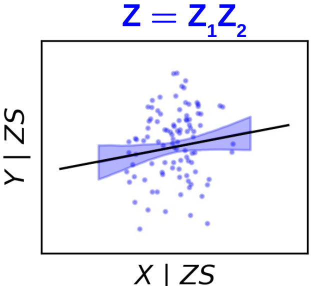


Causal inference preliminaries

(Linear) total causal effect for $x = x' + 1$ with a valid set $\mathbf{Z}$ is equal to $\beta_{YX\cdot\mathbf{ZS}}$ in

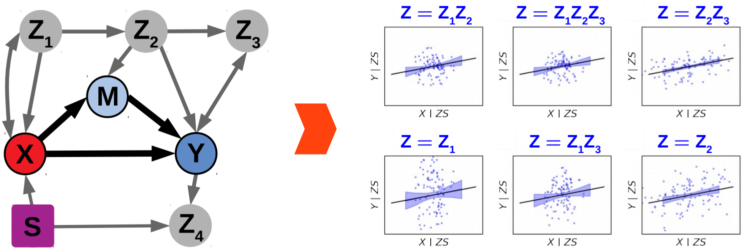
$$Y = \boxed{\beta_{YX\cdot\mathbf{ZS}}} X + \sum_i \beta_{YZ_i\cdot X\mathbf{S}} Z_i + \sum_i \beta_{YS_i\cdot X\mathbf{Z}} S_i \tag{1}$$





Problem setting

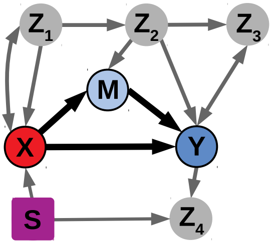
Consider all adjustment sets

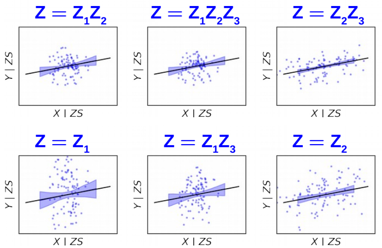


Problem setting

Consider all adjustment sets

All valid sets lead to estimates with zero bias of $\widehat{\beta}_{YX\cdot\mathbf{ZS}}$, but variance strongly differs.

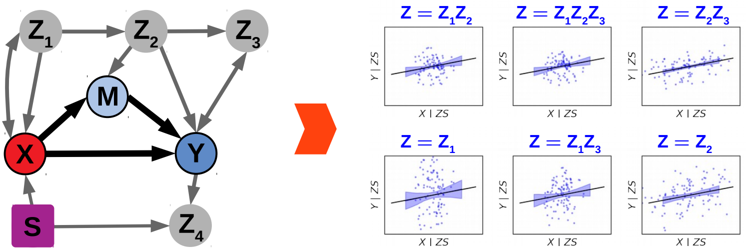




Problem setting

Open problem Find valid adjustment set that yields minimal *asymptotic* variance:

$$\mathbf{Z}_{\text{optimal}} \in \operatorname{argmin}_{\mathbf{Z} \in \mathcal{Z}} E[(\Delta_{yxx'|\mathbf{s}} - \widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{z}})^2]\,. \qquad (2)$$



Information-theoretic optimal adjustment theory

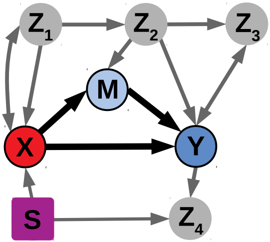
Def.: Conditional mutual information (CMI) for Shannon entropy

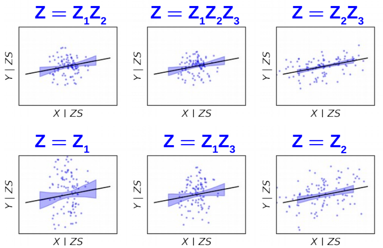
$H_{Y|X} = -\int_{x,y} p(x,y) \ln p(y|x) dxdy$

$$I_{X;Y|Z} \equiv H_{Y|Z} - H_{Y|ZX} \quad (3)$$

$$\geq 0 \quad (4)$$

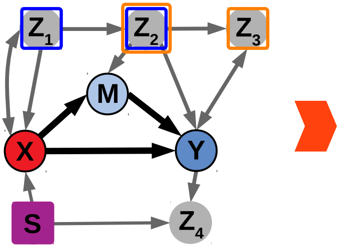
$$= 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z \quad (5)$$

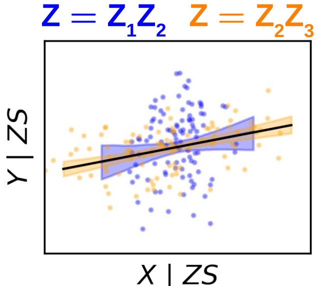




Information-theoretic optimal adjustment theory

Compare Adjust set $\mathbf{Z} = Z_1 Z_2$ vs $\mathbf{O} = Z_2 Z_3$



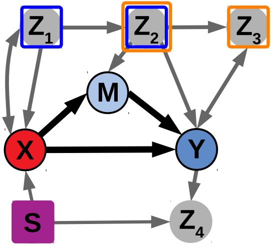


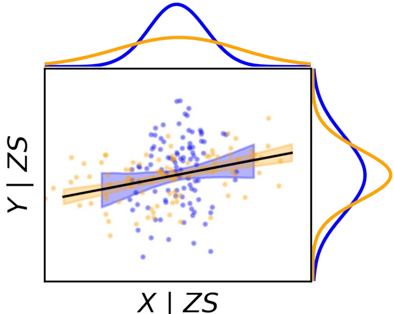
Information-theoretic optimal adjustment theory

Compare Adjust set Z $= Z_1 Z_2$ **vs O** $= Z_2 Z_3$

Two reasons for smaller estimator variance:

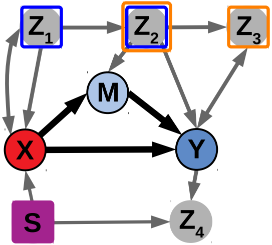
1. **Larger** residual variance of X
2. **Smaller** residual variance of Y

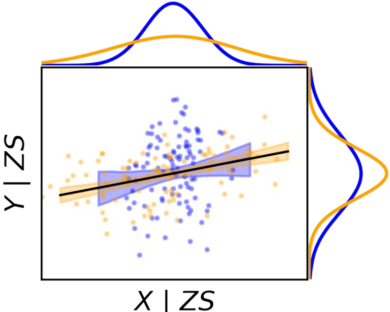




Information-theoretic optimal adjustment theory

Intuition Choose an adjustment set **Z** that maximally constrains Y and minimally constrains X



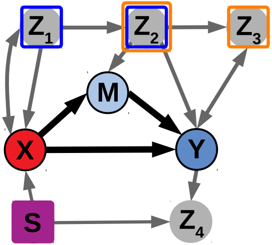


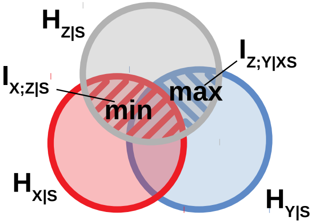
Information-theoretic optimal adjustment theory

Intuition Choose an adjustment set $\mathbf{Z}$ that maximally constrains Y and minimally constrains X

Def. 1: Adjustment information

$$J_{\mathbf{Z}} \equiv J_{XY|\mathbf{S}.\mathbf{Z}} \equiv I_{\mathbf{Z};Y|X\mathbf{S}} - I_{X;\mathbf{Z}|\mathbf{S}} \tag{3}$$



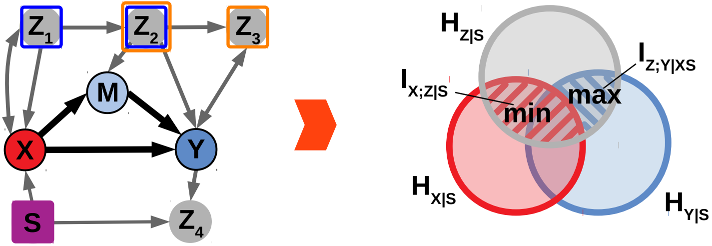


Information-theoretic optimal adjustment theory

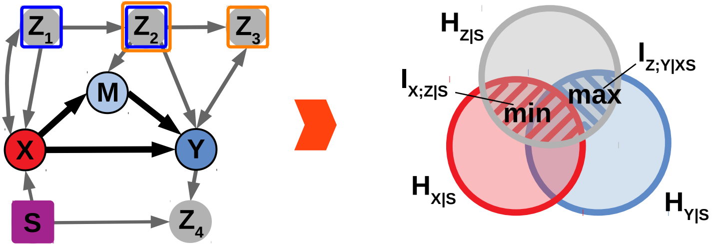
Optimality results are valid for estimators $\widehat{\Delta}_{yxx'|\mathbf{s.z}}$ that obey

$$\mathbf{Z}_{\text{optimal}} \in \operatorname{argmax}_{\mathbf{Z}\in\mathcal{Z}} J_{\mathbf{Z}} \;\Rightarrow\; Var(\widehat{\Delta}_{yxx'|\mathbf{s.z}_{\text{optimal}}}) = \min_{\mathbf{Z}\in\mathcal{Z}} Var(\widehat{\Delta}_{yxx'|\mathbf{s.z}})$$

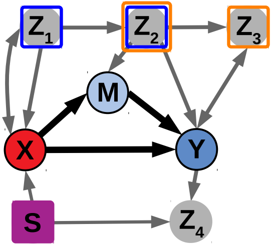
In paper theoretically shown for **OLS**, experimentally also for other estimators.

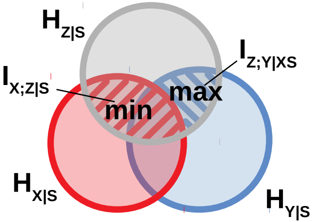


Information-theoretic optimal adjustment theory



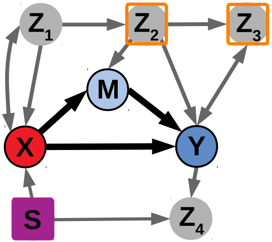
Information-theoretic optimal adjustment theory

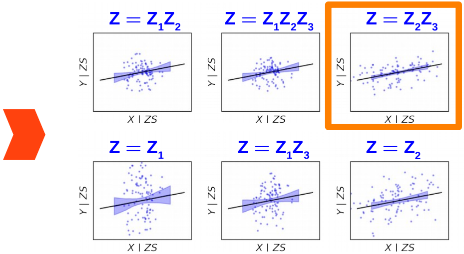




Information-theoretic optimal adjustment theory

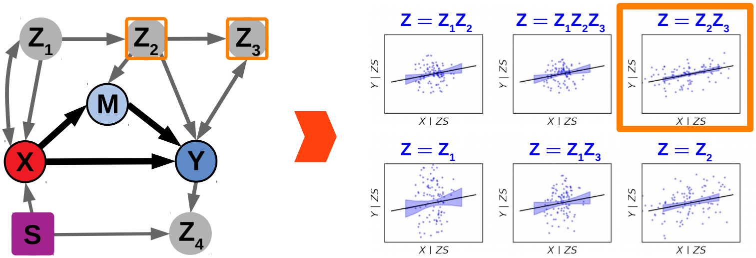
Def. 2: Graphical optimality For a tuple $(\mathcal{G}, X, Y, S)$ *graphical optimality holds* if there is a $\mathbf{Z} \in \mathcal{Z}$ s.t. for all other $\mathbf{Z}' \neq \mathbf{Z} \in \mathcal{Z}$ and *all* distributions $\mathcal{P}$ consistent with $\mathcal{G}$ we have $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$.



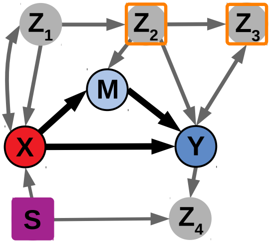


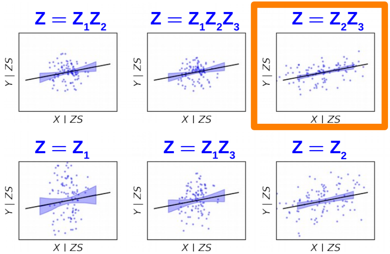
Information-theoretic optimal adjustment theory

Is there always an optimal adjustment set?

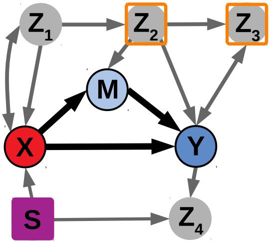


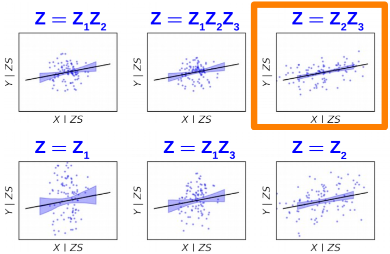
Information-theoretic optimal adjustment theory





Information-theoretic optimal adjustment theory



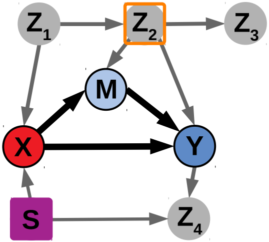


Information-theoretic optimal adjustment theory

Yes for DAGs without hidden variables

([Henckel et al., 2019, Witte et al., 2020, Rotnitzky and Smucler, 2019]):

$$\mathbf{O} = \mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}\,. \quad (3)$$

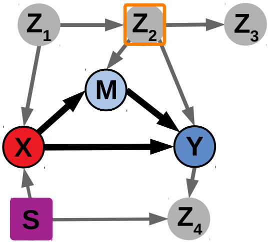


Information-theoretic optimal adjustment theory

Yes for DAGs without hidden variables

([Henckel et al., 2019, Witte et al., 2020, Rotnitzky and Smucler, 2019]):

$$\mathbf{O} = \mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}\,. \tag{3}$$

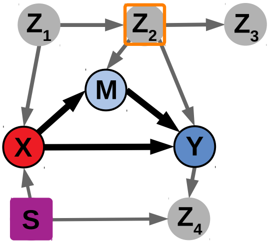


Information-theoretic optimal adjustment theory

Yes for DAGs without hidden variables

([Henckel et al., 2019, Witte et al., 2020, Rotnitzky and Smucler, 2019]):

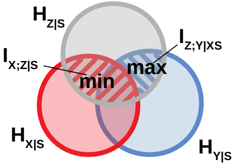
$$\mathbf{O} = \mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}\,. \quad (3)$$

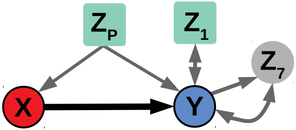


The optimal adjustment set for ADMGs with hidden variables

Intuition Constraining Y by $pa(Y\mathbf{M})$ not enough...

...add spouses since $I(Z_1; Y) > 0$ (as long as $\notin$ **forb**)...



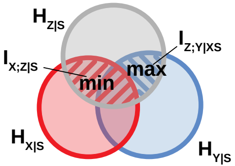


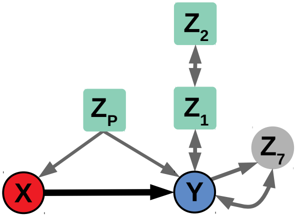
The optimal adjustment set for ADMGs with hidden variables

Intuition Constraining Y by $pa(Y\mathbf{M})$ not enough...

...add spouses since $I(Z_1;Y) > 0$ (as long as $\notin$ **forb**)...

...and spouses of spouses since $I(Z_1 Z_2;Y) = I(Z_1;Y) + \underbrace{I(Z_2;Y|Z_1)}_{\geq 0}$...





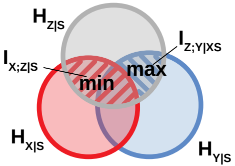
The optimal adjustment set for ADMGs with hidden variables

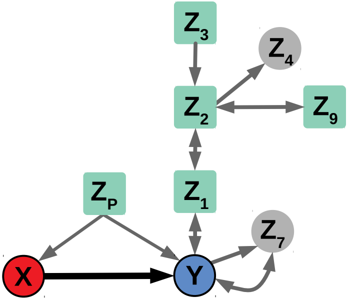
Intuition Constraining Y by $pa(Y\mathbf{M})$ not enough...

...add spouses since $I(Z_1; Y) > 0$ (as long as $\notin$ **forb**)...

...and spouses of spouses since $I(Z_1 Z_2; Y) = I(Z_1; Y) + \underbrace{I(Z_2; Y|Z_1)}_{\geq 0}$...

...until a tail is reached or the path ends...





The optimal adjustment set for ADMGs with hidden variables

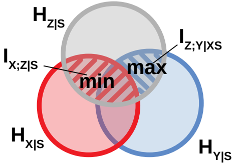
Intuition Constraining Y by $pa(Y\mathbf{M})$ not enough...

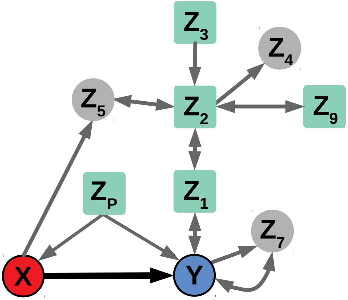
...add spouses since $I(Z_1;Y) > 0$ (as long as $\notin$ **forb**)...

...and spouses of spouses since $I(Z_1Z_2;Y) = I(Z_1;Y) + \underbrace{I(Z_2;Y|Z_1)}_{\geq 0}$...

...until a tail is reached or the path ends...

...exclude collider if $C \not\perp\!\!\!\perp X \mid \mathbf{vancs}$ (avoids non-causal paths)...





The optimal adjustment set for ADMGs with hidden variables

Intuition Constraining Y by $pa(Y\mathbf{M})$ not enough...

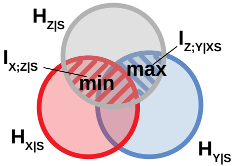
...add spouses since $I(Z_1;Y) > 0$ (as long as $\notin$ **forb**)...

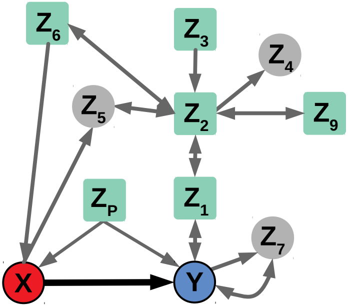
...and spouses of spouses since $I(Z_1 Z_2;Y) = I(Z_1;Y) + \underbrace{I(Z_2;Y|Z_1)}_{\geq 0}$...

...until a tail is reached or the path ends...

...exclude collider if $C \not\perp\!\!\!\perp X \mid \mathbf{vancs}$ (avoids non-causal paths)...

...except if $C \in \mathbf{vancs}$ where $\mathbf{vancs} = an(XY\mathbf{S}) \setminus \mathbf{forb}$





The optimal adjustment set for ADMGs with hidden variables

Def. O-set: $\mathbf{O}(X, Y, \mathbf{S}) = \mathbf{P} \cup \mathbf{C} \cup \mathbf{P_C}$ where

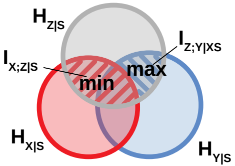
$$\mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}$$

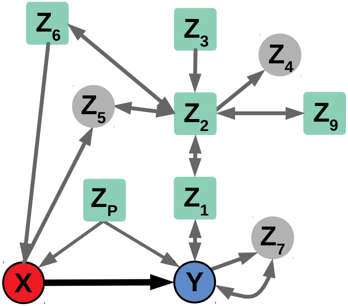
$$\mathbf{C} = \text{“valid collider paths from } W \in Y\mathbf{M}\text{”}$$

$$\mathbf{P_C} = pa(\mathbf{C})$$

where colliders $C \in \mathbf{C}$ fulfill

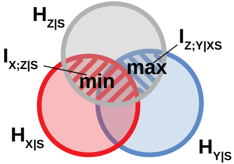
$$(1)\ C \notin \mathbf{forb}, \quad \text{and} \quad (2a)\ C \in \mathbf{vancs} \text{ or } (2b)\ C \perp\!\!\!\perp X \mid \mathbf{vancs}\,. \qquad (4)$$

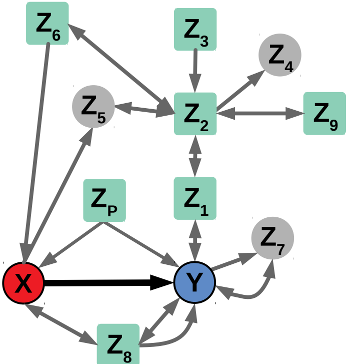




The optimal adjustment set for ADMGs with hidden variables

Theorem 1 (Validity) If and only if a valid backdoor adjustment set exists, then **O** is a valid adjustment set.

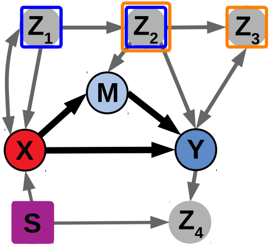




Theorem 2 (O-set vs Adjust set)

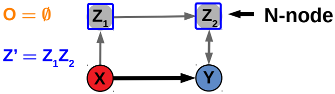
$J_{\mathbf{O}} \geq J_{\mathbf{vancs}}$ for any graph $\mathcal{G}$ (...).

$\Longrightarrow Var(\widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{o}}) \leq Var(\widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{adjust}})$



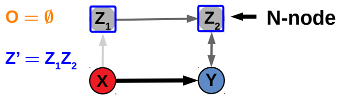
Necessary and sufficient conditions for graphical optimality

Theorem 3 If and only if (...)
(I) for *all* $N \in \mathbf{N} = sp(Y\mathbf{MC}) \setminus (\mathbf{forbOS})$ and all its collider paths i to $W \in Y\mathbf{M}$ (...) it holds that $\mathbf{O}_{\pi_i^N} = \mathbf{O}(X, Y, \mathbf{S}' = \mathbf{S}N\pi_i^N)$ is non-valid, and
(II) for *all* $E \in \mathbf{O} \setminus \mathbf{P}$ with $E \not\perp\!\!\!\perp X \mid \mathbf{SO} \setminus \{E\}$ there exists $E {\leftrightarrow} W$ or $E {*\!\!\rightarrow} C {\leftrightarrow} \cdots {\leftrightarrow} W$ where all colliders $C \in \mathbf{vancs}$,
then $\mathbf{O}$ is optimal for all probability densities consistent with $\mathcal{G}$.



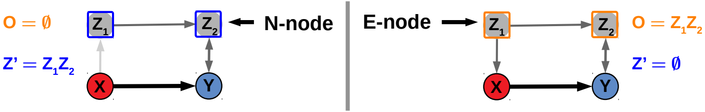
Necessary and sufficient conditions for graphical optimality

Theorem 3 If and only if (...)
(I) for *all* $N \in \mathbf{N} = sp(Y\mathbf{MC}) \setminus (\mathbf{forbOS})$ and all its collider paths i to $W \in Y\mathbf{M}$ (...) it holds that $\mathbf{O}_{\pi_i^N} = \mathbf{O}(X, Y, \mathbf{S}' = \mathbf{S}N\pi_i^N)$ is non-valid, and
(II) for *all* $E \in \mathbf{O} \setminus \mathbf{P}$ with $E \not\perp\!\!\!\perp X \mid \mathbf{SO} \setminus \{E\}$ there exists $E {\leftrightarrow} W$ or $E {*\!\!\rightarrow} C {\leftrightarrow} \cdots {\leftrightarrow} W$ where all colliders $C \in \mathbf{vancs}$,
then $\mathbf{O}$ is optimal for all probability densities consistent with $\mathcal{G}$.



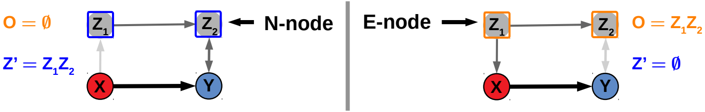
Necessary and sufficient conditions for graphical optimality

Theorem 3 If and only if (...)
(I) for *all* $N \in \mathbf{N} = sp(Y\mathbf{MC}) \setminus (\mathbf{forbOS})$ and all its collider paths i to $W \in Y\mathbf{M}$ (...) it holds that $\mathbf{O}_{\pi_i^N} = \mathbf{O}(X, Y, \mathbf{S}' = \mathbf{S}N\pi_i^N)$ is non-valid, and
(II) for *all* $E \in \mathbf{O} \setminus \mathbf{P}$ with $E \not\perp\!\!\!\perp X \mid \mathbf{SO} \setminus \{E\}$ there exists $E {\leftrightarrow} W$ or $E {*\!\!\rightarrow} C {\leftrightarrow} \cdots {\leftrightarrow} W$ where all colliders $C \in \mathbf{vancs}$,
then $\mathbf{O}$ is optimal for all probability densities consistent with $\mathcal{G}$.



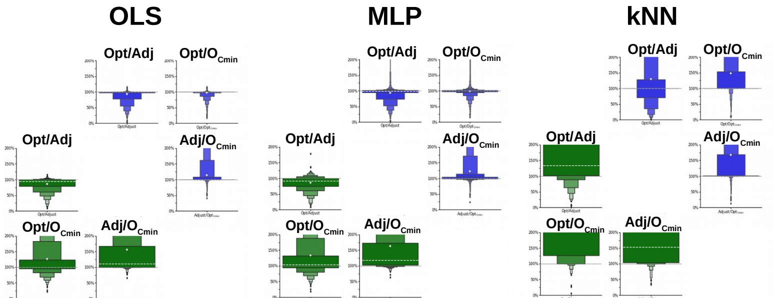
Necessary and sufficient conditions for graphical optimality

Theorem 3 If and only if (...)
(I) for *all* $N \in \mathbf{N} = sp(Y\mathbf{MC}) \setminus (\mathbf{forbOS})$ and all its collider paths i to $W \in Y\mathbf{M}$ (...) it holds that $\mathbf{O}_{\pi_i^N} = \mathbf{O}(X, Y, \mathbf{S}' = \mathbf{S}N\pi_i^N)$ is non-valid, and
(II) for *all* $E \in \mathbf{O} \setminus \mathbf{P}$ with $E \not\perp\!\!\!\perp X \mid \mathbf{SO} \setminus \{E\}$ there exists $E {\leftrightarrow} W$ or $E {*\!\!\rightarrow} C {\leftrightarrow} \cdots {\leftrightarrow} W$ where all colliders $C \in \mathbf{vancs}$,
then $\mathbf{O}$ is optimal for all probability densities consistent with $\mathcal{G}$.



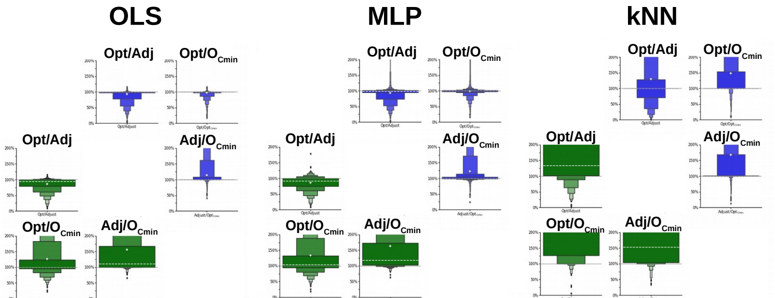
Numerical experiments in paper

- Among 12,000 randomly created configurations **95%** fulfill optimality!



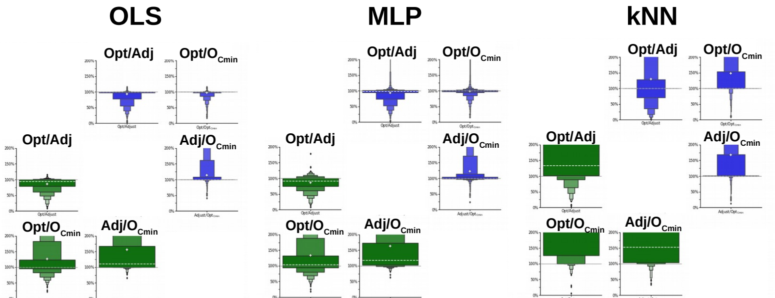
Numerical experiments in paper

- Among 12,000 randomly created configurations **95%** fulfill optimality!
- **OLS estimator:** Theoretical asymptotic results also hold for finite samples up to very small sample sizes



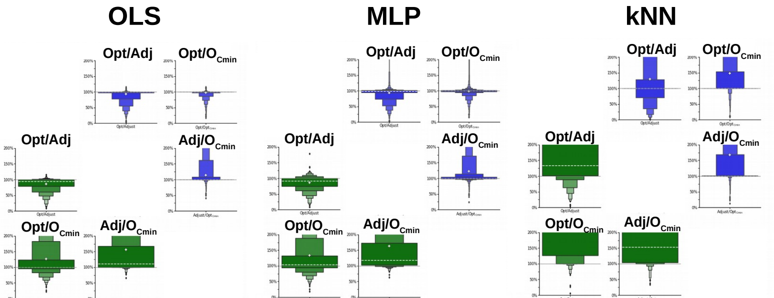
Numerical experiments in paper

- Among 12,000 randomly created configurations **95%** fulfill optimality!
- **OLS estimator:** Theoretical asymptotic results also hold for finite samples up to very small sample sizes
- **Neural net estimator:** Theory also applies to linear SCMs, but not for nonlinear SCMs



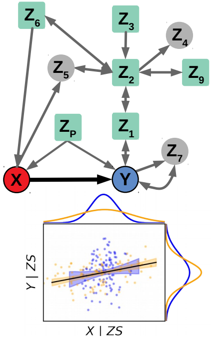
Numerical experiments in paper

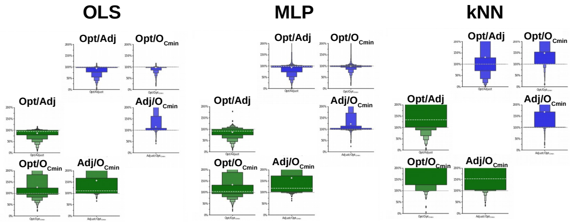
- Among 12,000 randomly created configurations **95%** fulfill optimality!
- **OLS estimator:** Theoretical asymptotic results also hold for finite samples up to very small sample sizes
- **Neural net estimator:** Theory also applies to linear SCMs, but not for nonlinear SCMs
- **kNN-estimator:** Theory not applicable, but a variant of **O**-set seems to outperform others



Summary

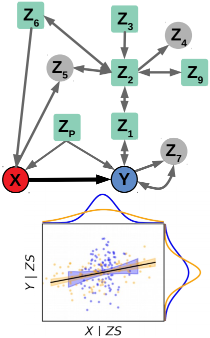
- Theorem 3 completely characterizes graphical optimality for ADMGs (and DMAGs)

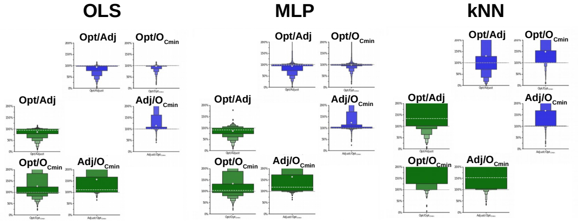




Summary

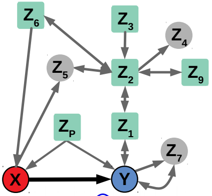
- Theorem 3 completely characterizes graphical optimality for ADMGs (and DMAGs)
- **O**-set is valid iff a valid set exists and always better than Adj-set
  $\rightarrow$ natural choice in automated causal inference

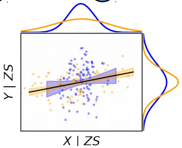


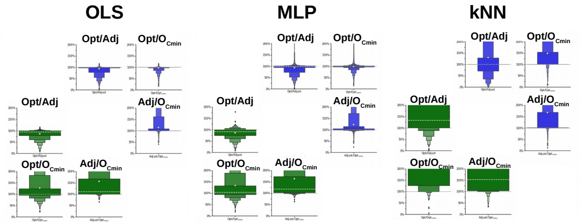


Summary

- Theorem 3 completely characterizes graphical optimality for ADMGs (and DMAGs)
- **O**-set is valid iff a valid set exists and always better than Adj-set $\rightarrow$ natural choice in automated causal inference
- Python code: `https://github.com/jakobrunge/tigramite`

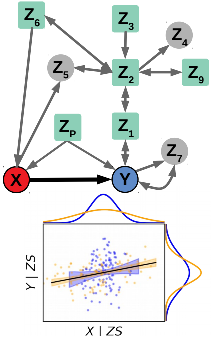


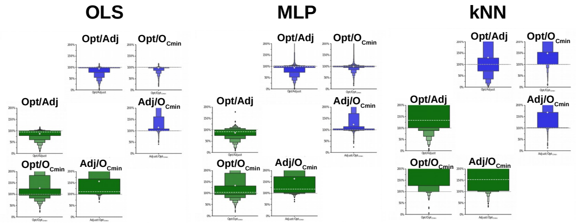




Summary

- Theorem 3 completely characterizes graphical optimality for ADMGs (and DMAGs)
- **O**-set is valid iff a valid set exists and always better than Adj-set
  $\rightarrow$ natural choice in automated causal inference
- Python code: `https://github.com/jakobrunge/tigramite`
- Open questions: Theory for non-parametric estimators, PAGs, ...





Thank you! Questions?

- *Nature Comm.* Perspective on causal discovery in time series [Runge et al., 2019a]
- Causal inference: full theory [Pearl, 2009], primer [Pearl et al., 2016], linear models [Pearl, 2013], popular science book [Pearl and Mackenzie, 2018]
- Causal discovery: general [Spirtes et al., 2000], for time series [Runge, 2018, Runge et al., 2019a]
- Restricted SCMs [Peters et al., 2017]
- PCMCI [Runge et al., 2019b] in *Science Advances*
- PCMCI$^+$ [Runge, 2020] in *UAI*
- LPCMCI [Gerhardus and Runge, 2020] in *NeurIPS*
- Optimal adjustment [Runge, 2021] in *NeurIPS*
- My software: `jakobrunge.github.io/tigramite`

Henckel, L., Perković, E., and Maathuis, M. H. (2019).
Graphical criteria for efficient total effect estimation via adjustment in causal linear models.
arXiv preprint arXiv:1907.02435.

Pearl, J. (2009).
Causality: Models, reasoning, and inference.
Cambridge University Press.

Pearl, J. (2013).
Linear models: A useful microscope for causal analysis.
J. Causal Inference, 1(1):155–170.

Pearl, J., Glymour, M., and Jewell, N. P. (2016).
Causal inference in statistics: A Primer.
John Wiley & Sons.

References ii

Pearl, J. and Mackenzie, D. (2018).
The Book of Why: The New Science of Cause and Effect.
Basic books, New York.

Perković, E., Textor, J., and Kalisch, M. (2018).
Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.
Journal of Machine Learning Research, 18:1–62.

Peters, J., Janzing, D., and Schölkopf, B. (2017).
Elements of causal inference: foundations and learning algorithms.
MIT Press, Cambridge, MA.

References iii

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

Runge, J. (2018).
Causal network reconstruction from time series: From theoretical assumptions to practical estimation.
Chaos An Interdiscip. J. Nonlinear Sci., 28(7):075310.

Runge, J. (2020).
Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets.
In Sontag, D. and Peters, J., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020, Toronto, Canada, 2019*. AUAI Press.

References iv

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019a).
Inferring causation from time series in earth system sciences.
Nature Communications, 10(1):2553.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019b).
Detecting and quantifying causal associations in large nonlinear time series datasets.
Science Advances, eaau4996(5).

References v

Spirtes, P., Glymour, C., and Scheines, R. (2000).
Causation, Prediction, and Search.
MIT Press, Boston.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020).
On efficient adjustment in causal graphs.
Journal of Machine Learning Research, 21(246):1–45.