Causal Effect Estimation from Observational and Interventional Data Through Matrix Weighted Linear Estimators

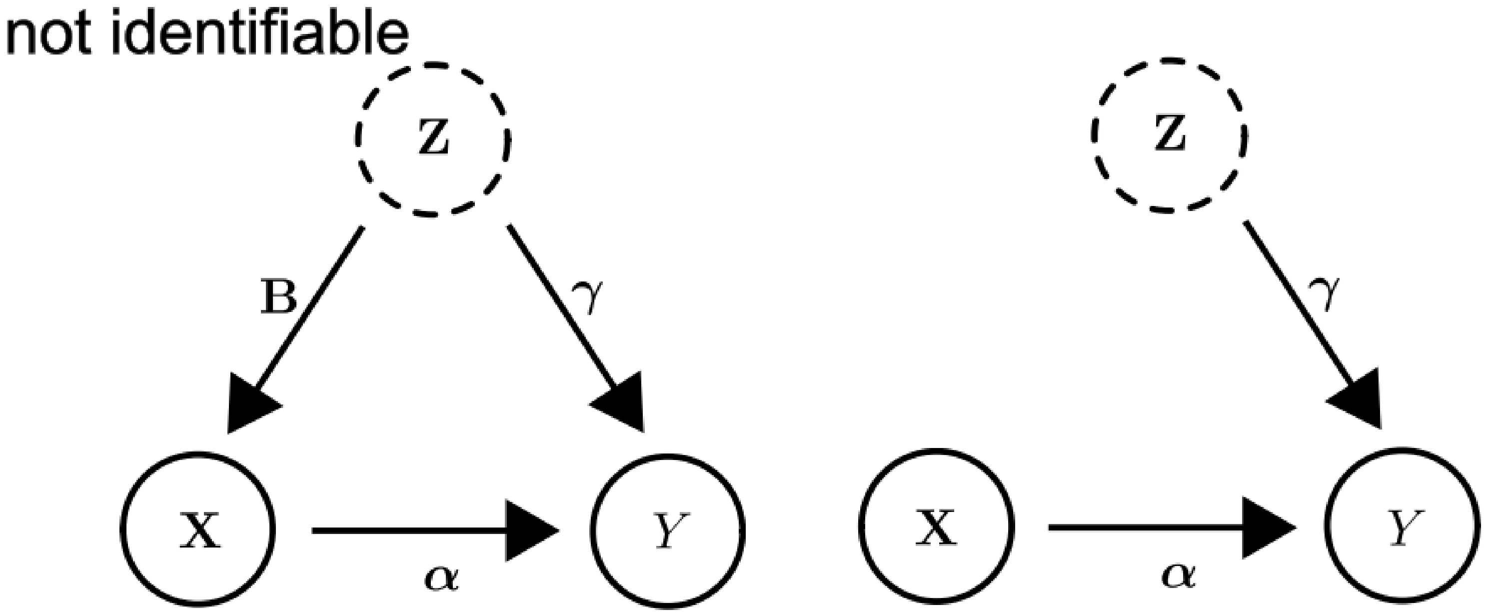
Klaus-Rudolf Kladny,Julius von Kügelgen,Bernhard Schölkopf,Michael Muehlebach

39th Conference on Uncertainty in Artificial Intelligence (UAI2023)

Presented by Raj Mohanty

CS276 Class Project

Setting & Motivation



(a) observational (b) interventional

- More expensive to collect
- Unbiased
- High variance (less sample size)

- Objective: Estimate **α** and minimize mean squared error (MSE)

Structural Equation Model (SEM)

$$\mathbf{Z} \leftarrow \mathbf{N_Z}, \qquad \mathbf{N_Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{N_Z}}, \boldsymbol{\Sigma}_{\mathbf{N_Z}}) \tag{1}$$

$$\mathbf{X} \leftarrow \mathbf{B}\mathbf{Z} + \mathbf{N_X}, \qquad \mathbf{N_X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{N_X}}, \boldsymbol{\Sigma}_{\mathbf{N_X}}) \tag{2}$$

Observational

$$Y \leftarrow \mathbf{Z}^\top \boldsymbol{\gamma} + \mathbf{X}^\top \boldsymbol{\alpha} + N_Y, \qquad N_Y \sim \mathcal{N}(\mu_{N_Y}, \sigma^2_{N_Y}) \tag{3}$$

$\mathbf{B} \in \mathbb{R}^{p \times d}$, $\boldsymbol{\gamma} \in \mathbb{R}^d$, $\boldsymbol{\alpha} \in \mathbb{R}^p$, and $(\mathbf{N_Z}, \mathbf{N_X}, N_Y)$ mutually independent exogenous noise variables.

$$\mathbf{X} \leftarrow \widetilde{\mathbf{N}}_\mathbf{X}, \qquad \widetilde{\mathbf{N}}_\mathbf{X} \sim \mathbb{P}_{\widetilde{\mathbf{N}}_\mathbf{X}},$$

Interventional

Data

X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	Y
1.00592726	1.2787921151	0.443962735	0.1705931897	-2.850389408	0.397202040	-0.514577067	-0.636647485	-4.00066820	-0.248848560	1.54932463	10.91980856
0.96963508	0.4436062900	-0.459429788	-0.6621852569	2.423316022	0.588088002	1.896189713	-0.261339119	-1.78389097	0.365062204	0.73290129	4.42119563
-0.14952796	1.1733352771	1.039452590	-2.3642812727	1.436724641	0.730779426	1.958336295	-1.118176230	0.82301035	-0.650660724	1.16844667	5.25947781
0.21977736	-1.3452224632	0.869056491	0.2600761076	0.365042388	0.308059736	-0.522131348	0.995933149	-0.46432349	-0.801367789	-1.14322925	-21.68270604
-0.50224671	0.0467913623	-0.158891164	-2.9652332227	-1.608405006	0.793533504	0.615774537	0.035995469	-1.27869539	0.044263137	1.20834069	6.02485721
0.30308422	2.5350282384	-0.527153628	-0.9131070537	0.902630897	1.177967785	-0.360796617	-2.561334260	0.12519685	-1.196351774	1.45872136	-2.78384788
1.07912025	1.9917275122	1.550950839	-1.9569204047	-0.721567113	-0.213309007	0.853201534	-0.730229571	-0.22903474	0.513373030	2.13800972	11.53902907
-0.31549905	-0.2645641463	-0.011056906	0.1802381813	2.909312401	-1.112350186	-0.652627569	-1.491590484	3.46765643	1.751346929	-0.38104736	-20.18052366
-0.65163493	-0.2771581482	-2.296690986	-0.6407696892	3.275055390	0.297086096	0.779710503	0.458982041	1.79215183	-0.815272106	0.30920951	-10.68456287
-2.23253046	0.1381552972	-1.595194322	-0.2564935303	2.753748584	-2.625471748	-1.325933352	-1.120547488	1.44109907	1.344225793	1.35401086	-33.19822244
-2.30956190	-3.5209891006	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.45217767	-16.45860815
-0.45658241	-0.8807510710	-1.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]648277	-1.37908516	-20.49765524
-0.52577459	0.0517549048	0.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]971033	-1.09667948	10.34897959
0.17462354	-0.4619581733	1.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]876228	0.21095895	23.37870508
2.08736974	0.7531031829	-0.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]072393	0.45436902	7.84018789
-0.79102995	-0.5407004093	-0.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]946212	-0.16310612	-5.14623216
1.14869965	1.4572758654	1.406627814	2.3126291504	-1.772125091	-0.289090777	0.382485063	-0.534258785	-1.22987057	-1.643358103	-0.47422267	-2.06790774
-0.59152628	0.0691140206	-2.884587773	0.1279244626	2.844617234	-0.882101667	0.798246392	-1.275567804	2.42137505	1.419113751	0.64100688	-13.18999913
-0.52235156	1.5061235489	0.414092310	-0.5951153613	-0.183686340	0.683922832	0.682663960	-0.470038139	-0.65327780	-0.679488653	-0.71007030	16.44762653
1.91304206	-0.8208789049	1.404268775	2.4792551695	-3.647763884	0.326659726	0.574262351	-1.064505467	-2.74732140	-0.924587781	2.03721127	38.54644230
-0.24714332	0.9357142699	0.958405730	-0.8382359468	-2.298081199	-0.405131217	-0.368413596	-0.506248213	-1.06148982	-0.176993628	0.76188419	4.48102995
0.42234120	0.5088636509	-1.147625195	-1.0827433786	2.467230433	0.323783241	-0.748094768	0.656103195	0.50010538	0.046236318	0.55767528	-5.53494571
0.38788706	-0.7086168759	0.351481035	-0.3316012742	1.692319139	-0.428779489	-1.237217142	0.376001650	2.02562916	0.922832707	-0.86168892	-21.27305522
1.25632051	0.9674743341	1.445208872	1.0964117245	-2.583249847	1.465278855	1.450211037	-2.125630474	-2.79203232	0.548682034	1.78185893	22.45360422
0.93722798	-0.0398427533	0.476051913	2.2002390803	-0.982557314	-1.912159638	-0.616822332	0.001688483	-0.67823759	0.314983569	1.09388915	0.24653020

$$(\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{obs}}, \quad i = 1, \dots, n,$$

$$(\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{int}}, \quad i = n+1, \dots, n+m,$$

Proposed Methods

$$\boldsymbol{\alpha} = \nabla_{\mathbf{x}} \mathbb{E}[Y | \mathrm{do}(\mathbf{X} \leftarrow \mathbf{x})]$$

True value

$$\mathbb{P}_{\text{obs}}(Y | \mathbf{X} = \mathbf{x}) \neq \mathbb{P}(Y | \mathrm{do}(\mathbf{X} \leftarrow \mathbf{x})) = \mathbb{P}_{\text{int}}(Y | \mathbf{X} = \mathbf{x})$$

$$\mathbb{E}_{\text{obs}}[Y | \mathbf{X} = \mathbf{x}] = (\boldsymbol{\alpha} + \boldsymbol{\Delta})^\top \mathbf{x}$$

$$\boldsymbol{\Delta} = (\boldsymbol{\Sigma}_{\mathbf{N_X}} + \mathbf{B} \boldsymbol{\Sigma}_{\mathbf{N_Z}} \mathbf{B}^\top)^{-1} \mathbf{B} \boldsymbol{\Sigma}_{\mathbf{N_Z}} \boldsymbol{\gamma}$$

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n} := (\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}})^{-1} \mathbf{X}_{\mathrm{O}}^\top \mathbf{y}_{\mathrm{O}},$$
$$\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m} := (\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}})^{-1} \mathbf{X}_{\mathrm{I}}^\top \mathbf{y}_{\mathrm{I}}.$$

$$\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}) = (\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}})^{-1} \sigma_{Y|\mathbf{X}}^2,$$
$$\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}) = (\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}})^{-1} \sigma_{Y|\mathrm{do}(\mathbf{X})}^2$$

Matrix Weighted Linear Estimator

$$\widehat{\alpha}_{\mathbf{W}}^{m} := \mathbf{W}\widehat{\alpha}_{\mathrm{I}}^{m} + (\mathbf{I}_p - \mathbf{W})\widehat{\alpha}_{\mathrm{O}}^{n}.$$

W is a weight matrix

$$\mathbf{W} = \mathbf{I}_p.$$

Purely interventional

$$\mathbf{W}_*^m = \left(\mathrm{Cov}(\widehat{\alpha}_{\mathrm{O}}^n) + \mathbf{\Delta}\mathbf{\Delta}^\top\right) \left(\mathrm{Cov}(\widehat{\alpha}_{\mathrm{I}}^m) + \mathrm{Cov}(\widehat{\alpha}_{\mathrm{O}}^n) + \mathbf{\Delta}\mathbf{\Delta}^\top\right)^{-1}$$

Theoretically Optimal

$$\widehat{\mathrm{Cov}}(\widehat{\alpha}_{\mathrm{I}}^m) = (\mathbf{X}_{\mathrm{I}}^\top \mathbf{X}_{\mathrm{I}})^{-1}\widehat{\sigma}_{Y|\mathrm{do}(\mathbf{X})}^2,$$
$$\widehat{\mathrm{Cov}}(\widehat{\alpha}_{\mathrm{O}}^n) = (\mathbf{X}_{\mathrm{O}}^\top \mathbf{X}_{\mathrm{O}})^{-1}\widehat{\sigma}_{Y|\mathbf{X}}^2,$$

$$\widehat{\sigma}_{Y|\mathrm{do}(\mathbf{X})}^2 = \frac{1}{m-1}\|\mathbf{y}_{\mathrm{I}} - \mathbf{X}_{\mathrm{I}}\widehat{\alpha}_{\mathrm{I}}^m\|_2^2$$
$$\widehat{\sigma}_{Y|\mathbf{X}}^2 = \frac{1}{n-1}\|\mathbf{y}_{\mathrm{O}} - \mathbf{X}_{\mathrm{O}}\widehat{\alpha}_{\mathrm{O}}^n\|_2^2$$

Practical estimators

Practical estimators

$$\widehat{\boldsymbol{\Delta}}_m = \widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n - \widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m$$

$$\widehat{\mathbf{W}}_*^m = \left(\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n) + \widehat{\boldsymbol{\Delta}}_m \widehat{\boldsymbol{\Delta}}_m^\top + \epsilon \mathbf{I}_p\right)$$
$$\left(\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m) + \widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n) + \widehat{\boldsymbol{\Delta}}_m \widehat{\boldsymbol{\Delta}}_m^\top + \epsilon \mathbf{I}_p\right)^{-1}$$

$$\lim_{m \to \infty} \mathrm{MSE}\left(\widehat{\alpha}^m_{\widehat{\mathbf{W}}^m}\right) = 0$$

Asymptotically unbiased (infinite sample limit)

- The bias and variance of this estimate has been shown to vanish as m->infinity even with large amount of biased observational data. Its still biased for finite sample.

Minimize estimator variance

$$\widehat{\boldsymbol{\Delta}}_m = \widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n - \widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m$$

Has high variance

$$\mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\Delta}}_m)) = \mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m)) + \mathrm{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n))$$

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n &\leftarrow \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \left\{ \|\mathbf{y}_{\mathrm{O}} - \mathbf{X}_{\mathrm{O}}\boldsymbol{\alpha}\|_2^2 \right\} \\
\mathbf{r} &\leftarrow \mathbf{y}_{\mathrm{I}} - \mathbf{X}_{\mathrm{I}}\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n \\
\widehat{\boldsymbol{\Delta}}_m &\leftarrow \arg\min_{\boldsymbol{\Delta}\in\mathbb{R}^p} \left\{ \|\mathbf{r} + \mathbf{X}_{\mathrm{I}}\boldsymbol{\Delta}\|_2^2 \right\}.
\end{aligned}$$

Lasso and Ridge estimates (minimizes variance)

$$\widehat{\mathbf{\Delta}}_m^{\ell 2} \leftarrow \arg\min_{\mathbf{\Delta} \in \mathbb{R}^p} \left\{ \|\mathbf{r} + \mathbf{X}_{\mathrm{I}} \mathbf{\Delta}\|_2^2 + \lambda_{\ell 2} \|\mathbf{\Delta}\|_2^2 \right\}$$

Ridge regression to reduce variance

Cross validation

$$\widehat{\mathbf{\Delta}}_m^{\ell 1} \leftarrow \arg\min_{\mathbf{\Delta} \in \mathbb{R}^p} \left\{ \|\mathbf{r} + \mathbf{X}_{\mathrm{I}} \mathbf{\Delta}\|_2^2 + \lambda_{\ell 1} \|\mathbf{\Delta}\|_1 \right\}$$

Lasso regression to reduce variance

$$\lim_{m \to \infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m_{\ell 2}} \right) = 0$$

$$\widehat{w}^m_{\mathrm{rm}} := \max\left\{ 1 - \frac{\mathrm{Tr}\left(\widehat{\mathbf{Cov}}\left(\widehat{\alpha}^m_{\mathrm{I}} \right) \right)}{\|\widehat{\alpha}^m_{\mathrm{I}} - \widehat{\alpha}^n_{\mathrm{O}}\|_2^2}, 0 \right\}$$

Rosenman et al (2020) for comparison

Data Pooling (special case)

$$
\begin{aligned}
\widehat{\alpha}_{\mathrm{P}}^{m} &:= (\mathbf{X}_{\mathrm{P}}^{\top}\mathbf{X}_{\mathrm{P}})^{-1}\mathbf{X}_{\mathrm{P}}^{\top}\mathbf{y}_{\mathrm{P}} \\
&= (\mathbf{X}_{\mathrm{O}}^{\top}\mathbf{X}_{\mathrm{O}} + \mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}})^{-1}(\mathbf{X}_{\mathrm{O}}^{\top}\mathbf{y}_{\mathrm{O}} + \mathbf{X}_{\mathrm{I}}^{\top}\mathbf{y}_{\mathrm{I}}) \\
&= \mathbf{W}_{\mathrm{P}}^{m}\widehat{\alpha}_{\mathrm{I}}^{m} + (\mathbf{I} - \mathbf{W}_{\mathrm{P}}^{m})\widehat{\alpha}_{\mathrm{O}}^{n},
\end{aligned}
$$

where

$$
\mathbf{W}_{\mathrm{P}}^{m} := (\mathbf{X}_{\mathrm{O}}^{\top}\mathbf{X}_{\mathrm{O}} + \mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}})^{-1}\mathbf{X}_{\mathrm{I}}^{\top}\mathbf{X}_{\mathrm{I}}.
$$

$$
\lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\alpha}_{\mathrm{P}}^{m}\right) > 0.
$$

Undesirable Asymptotically biased.

Ridge Regression (Special case)

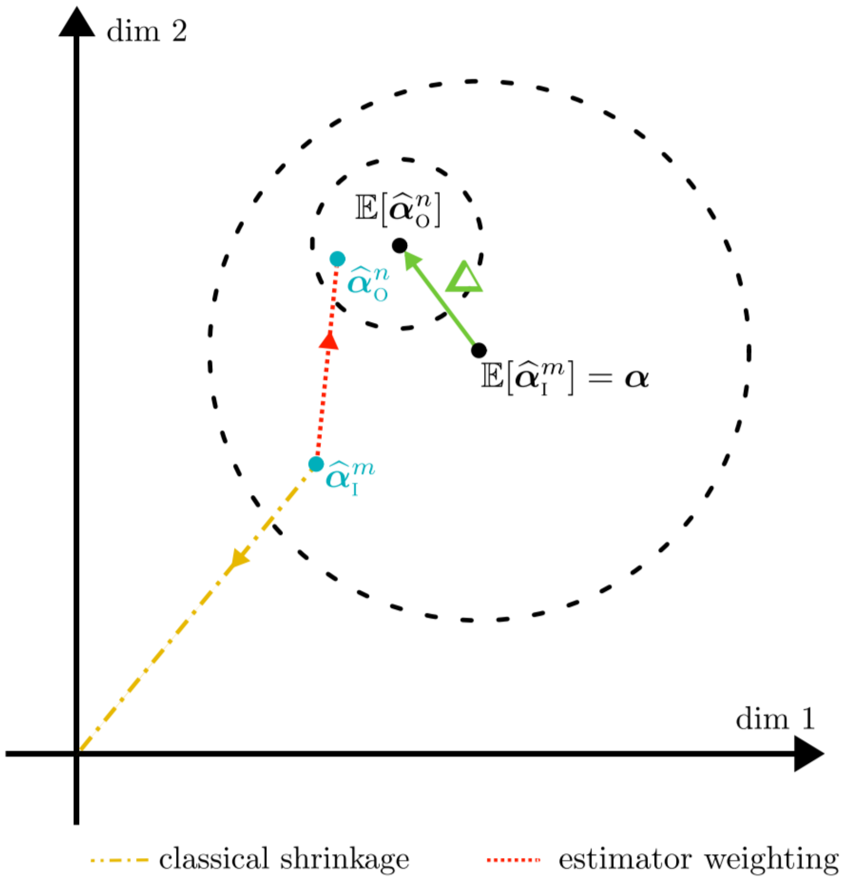
$$\widehat{\boldsymbol{\alpha}}^m_{\text{ridge}} = (\mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_{\text{I}}^\top \mathbf{y}_{\text{I}}$$
$$= \widehat{\mathbf{W}}^m_{\text{ridge}} \widehat{\boldsymbol{\alpha}}^m_{\text{I}} + (\mathbf{I}_p - \widehat{\mathbf{W}}^m_{\text{ridge}})\mathbf{0},$$

where

$$\widehat{\mathbf{W}}^m_{\text{ridge}} := (\mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}}.$$

no observational data and $\widehat{\boldsymbol{\alpha}}^n_{\text{O}} = \mathbf{0}$.

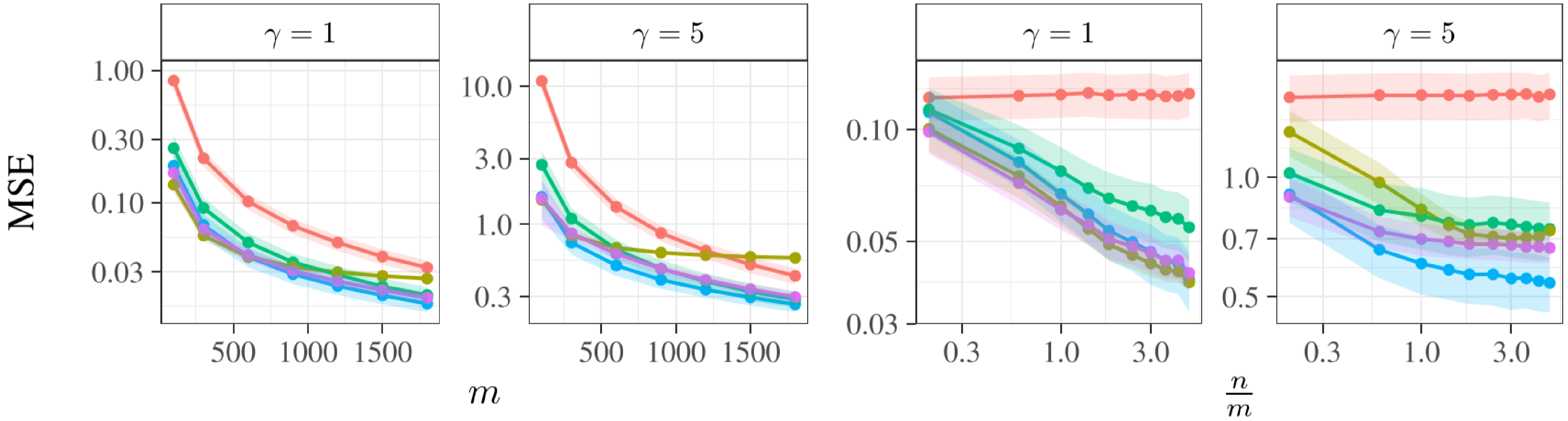
$\lim_{m\to\infty} \text{MSE}(\widehat{\boldsymbol{\alpha}}^m_{\text{ridge}}) = 0$



Experiment setup

General Setup. In all experiments, we use $p = 30$ treatments, a one-dimensional ($d = 1$) confounder Z, and unit/isotropic (co)variances: $\sigma^2_{N_Y} = \sigma^2_{N_Z} = 1$, $\mathbf{\Sigma}_{\mathbf{N_X}} = \mathbf{I}_p$. We sample $\tilde{\mathbf{N}}_\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{Cov}(\mathbf{X}_o))$, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_p)$, and choose b and γ depending on the settings described below. Unless otherwise specified, we then draw $m = 300$ interventional and $n = 600$ observational examples from $\mathbb{P}_{\text{int}}$ and $\mathbb{P}_{\text{obs}}$, respectively, and compute estimates of $\boldsymbol{\alpha}$ using the different weighting approaches. We repeat this procedure 1000 times and report the resulting mean and standard deviation of the mean squared error.

Results



n=3m

- Data pooling works well for gamma =1
- $\mathbf{W}^m_{\mathrm{I}}$ does not work well in small sample cases because of high variance

Conclusion

- A method to estimate treatment effects by utilizing both interventional and observational data.
- Minimize bias and variance
- Minimize MSE
- Future work
  - Beyond Linear Regression
  - Binary or categorical treatments
  - Binary or categorical outcome