Mathematical Statistics
School of Computation, Information and Technology
Technical University of Munich

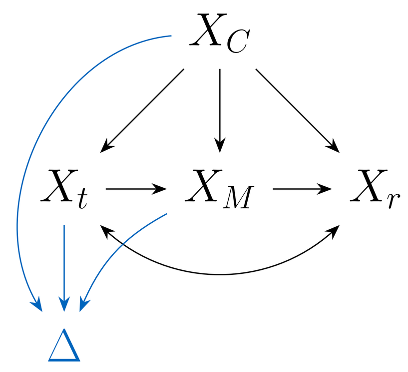
Cost-Aware Optimized Front-Door Experimental Design

Leopold Mareis[1], Mathias Drton[1,2]

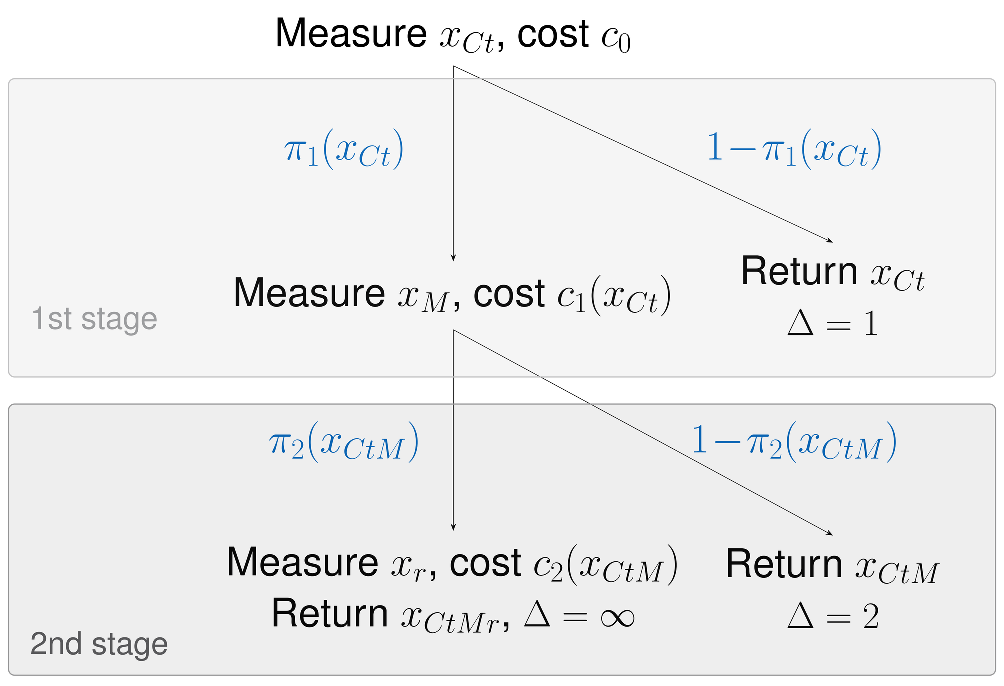
Motivation

Setting: Estimate linear causal effect ξ of treatment X_t on response X_r in a study.

- Confounder X_C, and X_t always observed
- Mediator X_M, and X_r costly to measure
- Coarsening Δ records sampled components



Coarsening Δ	Observed-Data $G_\Delta(X)$
1	X_{Ct}
2	X_{CtM}
∞	X_{CtMr}



Joint Problem: Optimize the

1) sampling design $\pi=(\pi_1,\pi_2)$
2) causal effect estimator $\hat{\xi}_n$

to minimize the asymptotic variance under the budget constraint

$$\mathbb{E}[c_0+\pi_1 c_1(X_{Ct})+\pi_1\pi_2\, c_2(X_{CtM})]=b_0.$$

Sample further components per unit proportional to its information-to-cost ratio.

Observed-Data Model $\mathcal{M}_\pi$

The full-data model $\mathcal{M}_1$ follows the linear SEM

$$\begin{pmatrix}X_C\\X_t\\X_M\\X_r\end{pmatrix}=\begin{pmatrix}0&0&0&0\\\beta_{tC}&0&0&0\\\beta_{MC}&\beta_{Mt}&0&0\\\beta_{rC}&0&\beta_{rM}&0\end{pmatrix}\begin{pmatrix}X_C\\X_t\\X_M\\X_r\end{pmatrix}+\begin{pmatrix}\varepsilon_C\\\varepsilon_t\\\varepsilon_M\\\varepsilon_r\end{pmatrix},$$

with absolutely continuous errors satisfying:

$$\mathcal{P}_\varepsilon=\mathcal{P}_{\varepsilon_C}\otimes\mathcal{P}_{\varepsilon_{t,r}}\otimes\mathcal{P}_{\varepsilon_M}$$

The observed-data model $\mathcal{M}_\pi$ is constructed as

$$\mathcal{M}_\pi=\left\{\mathcal{L}(\Delta,G_\Delta(X)):\begin{array}{c}\theta=(\beta,\varepsilon)\in\mathcal{B}\times\mathcal{E},\\X\sim\mathcal{P}_{\beta,\varepsilon}\in\mathcal{M}_1,\\\Delta|X\sim\mathcal{P}_\pi(\cdot\,|\,X)\end{array}\right\}.$$

➡ Arbitrary unobserved confounding $X_t\leftrightarrow X_r$.
➡ Causal effect of interest: $\xi=\beta_{rM}\beta_{Mt}$

Optimized Experimental Design

Regular asymptotically linear estimator:

$$\sqrt{n}(\hat{\xi}_n-\xi)\xrightarrow{d}\mathcal{N}(0,\mathrm{E}[\varphi\varphi^\top])$$

Characterized by an influence function φ.

Theorem. In $\mathcal{M}_1$, the efficient influence function for ξ is $\varphi_\xi^{F,\text{eff}}=\beta_{rM}\varphi^{F,\text{eff}}_{\beta_{Mt}}+\varphi^{F,\text{eff}}_{\beta_{rM}}\beta_{Mt}$, with

$$\varphi^{F,\text{eff}}_{\beta_{Mt}}=\varepsilon_M\varepsilon_t\,\mathrm{Var}(\varepsilon_t)^{-1}$$
$$\varphi^{F,\text{eff}}_{\beta_{rM}}=\varepsilon_r^\perp\,\varepsilon_M^\top\,\mathrm{Var}(\varepsilon_M)^{-1}$$

and $\varepsilon_r^\perp:=\varepsilon_r-\mathrm{E}[\varepsilon_r|\varepsilon_t]$. The two terms are **orthogonal**, separately targeting β_{Mt} and β_{rM}.

In $\mathcal{M}_\pi$ each EIF term is re-weighted by the inverse sampling propensity. Let $g_1(X_{Ct})$, $g_2(X_{CtM})$ denote the unit-level information scores for β_{Mt} and β_{rM}:

$$g_1(X_{C,t}):=\beta_{rM}\mathrm{Var}(\varepsilon_M)\,\beta_{rM}^\top\frac{\varepsilon_t^2}{\mathrm{Var}(\varepsilon_t)^2},$$

$$g_2(X_{C,t,M}):=\mathrm{Var}(\varepsilon_r\mid\varepsilon_t)\left(\varepsilon_M^\top\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2$$

➡ The asymptotic variance of the optimal augmented estimator $\hat{\xi}_n$ of $\varphi^{F,\text{eff}}$ is

$$\mathrm{Var}_\infty(\hat{\xi}_n;\pi)=\mathrm{E}\left[\frac{g_1}{\pi_1}\right]+\mathrm{E}\left[\frac{g_2}{\pi_1\pi_2}\right].$$

Theorem. The experimental design π^* minimizing $\mathrm{Var}_\infty(\hat{\xi}_n;\pi)$ under $g_1<\lambda c_1$ is

$$\pi_1^*=\min\left(1,\max\left(\sqrt{\frac{g_1}{\lambda c_1}},\sqrt{\frac{g_1+\mathrm{E}[g_2|\varepsilon_{C,t}]}{\lambda(c_1+\mathrm{E}[c_2|\varepsilon_{C,t}])}}\right)\right)$$

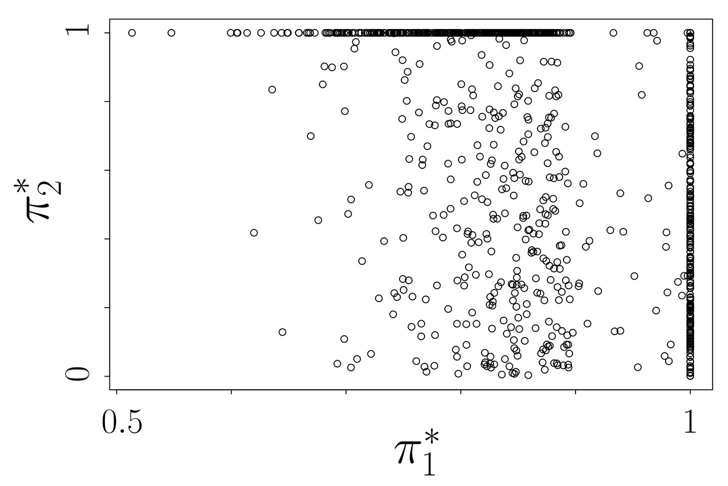
$$\pi_2^*=\min\left(1,\sqrt{\frac{g_2\,c_1}{g_1\,c_2}}\right).$$

For $g_1\geq\lambda c_1$:

$$(\pi_1^*,\pi_2^*)=\left(1,\min\left(1,\sqrt{\frac{g_2}{\lambda c_2}}\right)\right),$$

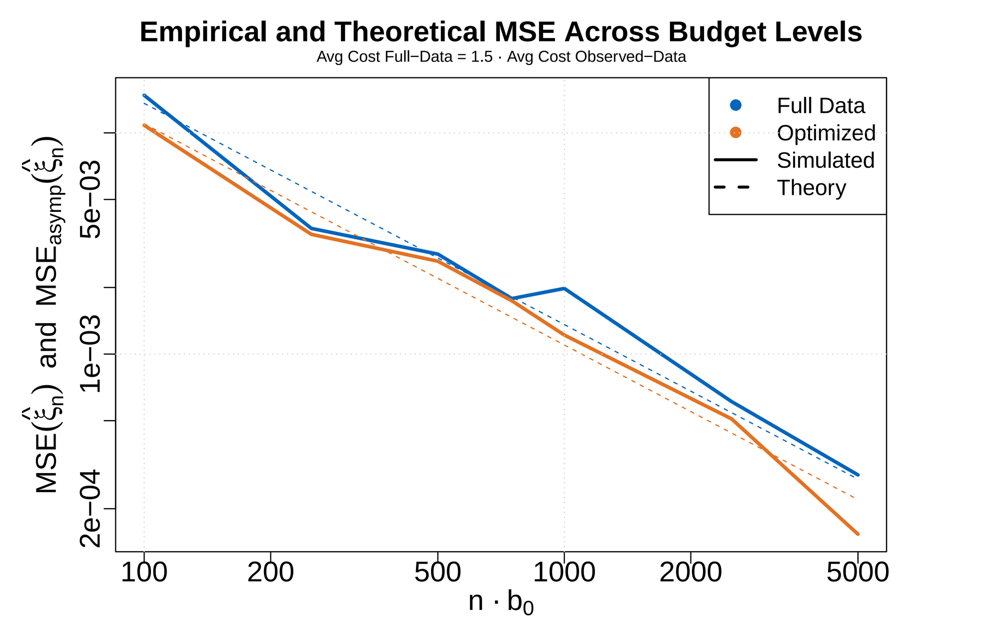
where λ satisfies the budget constraint.

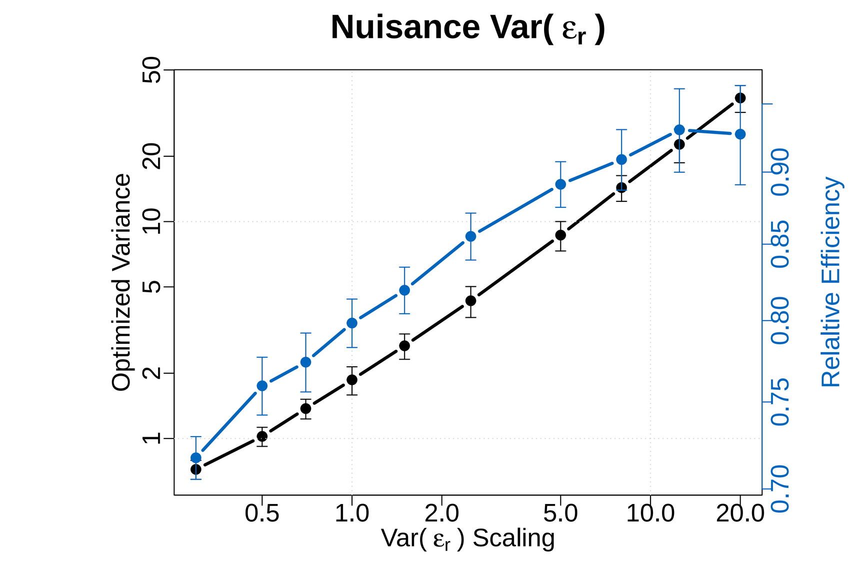
➡ π^* is driven by unit-level g/c ratios, yielding mixed propensities:



➡ Estimator: $\hat{\xi}_n=\hat{\beta}_{rM,n}\cdot\hat{\beta}_{Mt,n}$ via weighted residual LS on $\{\Delta=\infty\}$, $\{\Delta\geq2\}$, and $\{\Delta=1\}$.
➡ Optimized back-door experimental design follows as a corollary.

Simulation Study





Low noise in X_r reduces estimation uncertainty for β_{rM}, making partial-data sampling relatively more efficient compared to the full-data design.

Applications

Optimized design and estimation applied to datasets from biology, medicine, and industry. Rel. eff. $=\mathrm{Var}(\pi^*)/\mathrm{Var}(\pi\equiv1)$ at equal budget.

Dataset	n	Oversamp.	Rel. eff.
Front-door Estimation			
Sachs (I)	853	135%	76.6%
Sachs (II)	853	154%	$\mathbf{68.0\%}$
Back-door Estimation			
MIMIC	999	115%	90.0%
CausalAssembly	$3,000$	109%	94.7%

Up to 32.0% variance reduction (Sachs II). The partial-data design generates 54% more samples at the same budget level.

Conclusion

➡ Full-data EIF and optimal observed-data augmentation derived for front-door causal effects in linear SEMs.
➡ Closed-form π^* based on unit-level information-to-cost ratio. Extends to back-door.
➡ Substantial efficiency gains over full-measurement designs. Robust to mild misspecification.

Future: Non-linear models and identified graphs beyond the front-door.

arXiv

Code

L. Mareis

[1] Research Group Mathematical Statistics, Technical University of Munich, Germany.
[2] Munich Center for Machine Learning, Germany.
* Contact: leopold.mareis@tum.de

CLeaR 2026

This project has received funding by the Mathematical Research Data Initiative (MaRDI, 460135501) and the European Research Council (ERC, 883818).