Optimizing Experimental Design for Causal Effect Estimation with Partial Measurements

Fraunhofer IKS

Fraunhofer-Institut für Kognitive Systeme IKS

Leopold Mareis[1,2]

[1] Fraunhofer IKS [2] Technical University of Munich

Background

Estimating **causal effects** between two variables is essential for understanding a random system, and these finite-sample estimates are uncertain. We study the problem of **gathering additional data** to increase the accuracy of causal effect estimates. Inspired by applications, we **allow for requesting non-complete, and thus cheaper, observations**. Depending on the domain, cheaper can also mean faster or more ethical. Our approach leverages asymptotic variances in a Gaussian Graphical Model [6] to form decisions on subsequent data requests. Although this approach can be applied to arbitrary DAGs, we focus on the well-studied **instrumental variable approach** [5]. If the requirements are satisfied, this setup allows for causal effects estimation without the necessity of randomization. To make informed decisions, our theory provides information on the **interplay between available budget, p-values, and power computations**. This is especially important in safety-critical domains where data sampling is restricted and needs to be approved and assessed.

Instrumental Variable Model

We consider the causal model given by the structural equation model

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ 0 & \lambda_{23} & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \varepsilon,\ \varepsilon \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{11} & 0 & 0 \\ 0 & \omega_{22} & \omega_{23} \\ 0 & \omega_{23} & \omega_{33} \end{pmatrix} \right), \tag{1}$$

$$\begin{tikzcd} X_1 \arrow[r, "\lambda_{12}"'] & X_2 \arrow[r, "\lambda_{23}"'] \arrow[r, leftrightarrow, dashed, bend left, "\omega_{23}"] & X_3 \end{tikzcd}$$

with $\lambda_{12} \neq 0, \lambda_{23}, \omega_{23} \in \mathbb{R}$ and $\omega_{11}, \omega_{22}, \omega_{33} > 0$ satisfying $\omega_{22}\omega_{33} > \omega_{23}^2$. The random vector ($X_1$, X_2, X_3) represents (**Instrument**, **Treatment**, **Outcome**) and ω_{23} models the hidden relationship between X_2 and X_3. If $\lambda_{12} \neq 0$, the unbiased estimator for the causal effect $\hat\lambda_{23}$ is $\hat\sigma_{13}/\hat\sigma_{12}$ [1] as

$$\mathrm{Var}(X) = \begin{pmatrix} \omega_{11} & \omega_{11}\lambda_{12} & \omega_{11}\lambda_{12}\lambda_{23} \\ \cdot & \omega_{22} + \omega_{11}\lambda_{12}^2 & \omega_{23} + \lambda_{23}\sigma_{22} \\ \cdot & \cdot & \omega_{33} + 2\omega_{23}\lambda_{23} + \lambda_{23}^2\sigma_{22} \end{pmatrix} =: \Sigma. \tag{2}$$

Partial Measurements

Assume that N initial samples of X_{123} are collected. Further, fix budget $b > 0$, cost(X_{123} sample) $= c_1$, and cost(X_{12} sample) $= c_2 < c_1$. Let m_1 denote the number of additional X_{123} samples and m_2 the number of additional X_{12} samples. Let $\hat\sigma_{jk;1}$ and $\hat\sigma_{jk;2}$ be the covariances on $(X_{123}^{(i)})_{i\in[n+m_1]}$, respectively $(X_{12}^{(j)})_{j\in[m_2]}$. The **partial measurement estimator** is

$$\hat\lambda_{23} = \frac{\hat\sigma_{13;1}}{\frac{1}{1+\gamma}\hat\sigma_{12;1} + \frac{\gamma}{1+\gamma}\hat\sigma_{12;2}},$$

where $\gamma = m_2/(N+m_1)$. The goal is to **spend the available budget in an optimal way** to reduce the asymptotic ($m_1 \to \infty$) variance v of $\sqrt{N+m_1}(\hat\lambda_{23} - \lambda_{23})$:

$$\min_{(m_1,m_2)\in\mathbb{N}^2} \frac{v(\gamma = \frac{m_2}{N+m_1})}{N+m_1}, \text{ s.t. } c_1 m_1 + c_2 m_2 \leq b. \tag{3}$$

Theoretical Results

Theorem: The asymptotics $\sqrt{N+m_1}(\hat\lambda_{23} - \lambda_{23}) \xrightarrow{d} N(0, v(\gamma))$ hold with the variance

$$\begin{aligned} v(\gamma) &= \frac{1}{\sigma_{12}^2}\left(\frac{\sigma_{13}^2\sigma_{11}\sigma_{22}}{(1+\gamma)\sigma_{12}^2} - 2\frac{\sigma_{13}\sigma_{11}\sigma_{23}}{(1+\gamma)\sigma_{12}} + \sigma_{11}\sigma_{33} + 2\sigma_{13}^2 \right) \\ &= \frac{1}{\omega_{11}\lambda_{12}^2}\left(\frac{\gamma}{(1+\gamma)}\lambda_{23}^2\omega_{22} + \frac{(2+3\gamma)}{(1+\gamma)}\lambda_{23}^2\omega_{11}\lambda_{12}^2 + \frac{2\gamma}{(1+\gamma)}\lambda_{23}\omega_{23} + \omega_{33} \right). \end{aligned} \tag{4}$$

Theorem: The solution to Problem (3) over $\mathbb{R}_{\geq 0}^2$ is given by $(m_1^*, m_2^*) = (b/c_1, 0)$ if

$$\chi_1 := (c_1 - c_2)(\sigma_{11}\sigma_{13}^2\sigma_{22} - 2\,\sigma_{11}\sigma_{12}\sigma_{13}\sigma_{23}) \leq 2\,\sigma_{12}^2\sigma_{13}^2 c_2 + \sigma_{11}\sigma_{12}^2 c_2\sigma_{33} =: \chi_2.$$

Otherwise, a positive amount of X_{12} samples are requested and the solution reads

$$(m_1^*, m_2^*) = \left(\frac{b - \xi c_2 N}{c_1 + \xi c_2}, \frac{\xi(b + c_1 N)}{c_1 + \xi c_2}\right), \text{ where } \xi = \min\left(-1 + \frac{\sqrt{\chi_1\chi_2}}{\chi_2}, \frac{b}{c_2 N}\right).$$

- Additional X_{12} samples are only beneficial if $\lambda_{23}\omega_{23} < 0$.
- Weakening the instrument λ_{12}^2 always weakens the condition $\chi_1 \leq \chi_2$.

Applications

Aggressive driving is a risk for drivers and positively affects the number of accidents. In this dataset, $10,932$ drivers were monitored while driving. Environmental variables such as lighting, temperature, humidity, and car-specific variables were recorded. The causal DAG is obtained by the PC-Algorithm [4]. This DAG suggests an instrumental variable setup with $(X_1, X_2, X_3) =$ (lighting, driving style, speed). Therefore, the question of interest is whether the driving style influences the car's average speed. Here, the direct effect is -5.6km/h average speed per unit increase of aggressive driving style. Hence, actions to relax the driving style should be taken to reduce the average speed, preventing unnecessary accidents. When applying our *Adaptive Sampling Method* (3), the competing estimates have the following variances: Optimized variance: 1.012. Baseline variance: 1.274. Theoretical optimized variance: 7.329. Theoretical baseline variance: 7.526. Our proposed optimized method achieves a 21% smaller variance over the baseline procedure. For $N = 500$, $b = 300$, and $c = (1, 0.3)$, the optimized allocation is $\lfloor m^* \rfloor = (211, 293)$.

While our method also advised gathering more partial data on the **MIMIC-IV ICU-dataset** [3], the covariance matrix of the **Assembly Line dataset** [2] did not satisfy the requirements to further sample from X_{12}.

Limitations and Conclusion

- L The optimization problem is solved over the reals instead of the integers.
- L The theoretical results have not yet covered the identifiable additive-noise instrumental variable model.
- C **Sample size , significance level and power calculations** are provided for the test $H_0 : \lambda \leq 0$.
- C The optimization problem can be **adapted to various causal DAGs** with identified causal effects. In this case, the solutions must be approximated numerically.
- C The **Gaussianity assumption can be relaxed** towards, e.g., asymptotic normality of the MLE.
- C The solution of the optimization problem can **give structural insights**, e.g., indicate aligned confounding and causal effect signs.

References

• Drton, M. et al. (2018). Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77, pages 35–87. Mathematical Society of Japan Tokyo.• Göbler, K., Windisch, T., Drton, M., Pychynski, T., Roth, M., and Sonntag, S. (2024). causalassembly: Generating realistic production data for benchmarking causal discovery. In *Causal Learning and Reasoning*, pages 609–642. PMLR.• Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2020). Mimic-iv. *PhysioNet*, pages 49–55.• Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3).• Kilbertus, N., Kusner, M. J., and Silva, R. (2020). A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33:20108–20119.• Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Contact:
leopold.mareis@iks.fraunhofer.de
www.iks.fraunhofer.de
LinkedIn:

The full paper is accepted at @PCIC2024 and with be published in the proceedings.
This work has been funded by the Federal Ministry of Education and Research (BMBF, 01IS22087Q), TM, and the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Fraunhofer Institute for Cognitive Systems.