Falsifying causal models via nonparametric conditional independence testing

Lucas Kook[1]

[1] Institute for Statistics and Mathematics, WU Vienna, Vienna, Austria

E-mail for correspondence: lucas.kook@wu.ac.at

Abstract: Estimating causal effects from observational data requires (i) assumptions on the underlying data-generating process, such as a graphical causal model, and (ii) an identification strategy for the causal effect of interest, such as covariate adjustment. Both (i) and (ii) typically involve untestable assumptions, making it crucial to be able to criticize or falsify the resulting effect estimates. This work proposes one way to do so: Given a putative causal model and an observational dataset, we first extract testable conditional independence relations from the causal model. We then nonparametrically test those relations, potentially falsifying the causal model, while controlling Type I error. We illustrate the approach based on covariance measure tests, a family of regression-based nonparametric conditional independence tests, by falsifying two causal models of protein interactions using publicly available single cell flow cytometry data.

Keywords: Causal models; conditional independence; sensitivity analysis.

1 Introduction

Estimating causal effects from observational data is an important, yet challenging task, which requires a causal model and typically strong and untestable assumptions [Peters et al., 2017]. Scientists rely on subject matter knowledge to justify these untestable assumptions and to identify and estimate the causal effect of interest. In order for the resulting estimates to be trustworthy, it is important to check testable implications of the causal model [Su and Henckel, 2022] or to conduct sensitivity analyses [Cinelli et al., 2019]. Such testable implications oftentimes come in the form of conditional independence (CI) relations [Dawid, 1979], which (under the causal Markov condition) can be directly read off a causal graphical model [Pearl, 2009], or are explicitly stated for (or encoded in the functional form of) potential outcomes [Rubin, 1974].
In this work, we focus on testing CI relations implied by a causal model using nonparametric CI tests. Given a causal model and an observational dataset

This paper was published as a part of the proceedings of the 39th International Workshop on Statistical Modelling (IWSM), Limerick, Ireland, 13–18 July 2025. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

assumed to be generated by this causal model, we first extract testable CI relations from the causal model and then perform nonparametric tests using the observational data to find conditional *dependencies* which are inconsistent with and thereby falsify the causal model, while controlling the rate of false rejections (Type I error). A nonparametric approach is taken in order to avoid parametric assumptions which may be inconsistent with the causal model in question. We use covariance measure tests (COMETs, Section 2), a family of regression-based CI tests [Kook and Lundborg, 2024] to falsify CI relations in functional proteomics based on a publicly available flow cytometry dataset (Section 3). The validity of these tests depends on the predictive performance of those regressions, and thus relies crucially on principles from statistical modeling machine learning.

2 Covariance measure tests

We rely on nonparametric tests for the null hypothesis of conditional independence $H_0 : X \perp\!\!\!\perp Y \mid Z$, where $(X, Y, Z) \sim P$, $P \in \mathcal{P}$, and $(X, Y, Z) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y} \times \mathbb{R}^{d_Z}$. Without restricting $\mathcal{P}$, conditional independence is untestable in the sense that there exists no valid test with non-trivial power against arbitrary alternatives [Shah and Peters, 2020]. However, it is still possible to devise valid and powerful nonparametric tests for implications of H_0 at the cost of having power against restricted classes of alternatives.
In this work, we focus on two COMETs: The Generalised [GCM, Shah and Peters, 2020] and Projected Covariance Measure [PCM, Lundborg et al., 2024] tests. The GCM test targets the following implication of H_0,

$$H_0^{\mathrm{GCM}} : \mathbb{E}[\mathrm{Cov}(X, Y \mid Z)] = \mathbb{E}[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])] = 0,$$

while the PCM test, requiring $d_Y = 1$, targets the stronger null hypothesis of conditional mean independence,

$$H_0^{\mathrm{PCM}} : \mathbb{E}[Y \mid X, Z] = \mathbb{E}[Y \mid Z] \implies \mathbb{E}[\mathrm{Cov}(f(X, Z), Y \mid Z)] = 0,$$

where $f(X, Z) := \frac{\mathbb{E}[Y|X,Z] - \mathbb{E}[Y|Z]}{\mathrm{Var}(Y|X,Z)}$, and thus has power against a larger class of alternatives than the GCM test. The assumptions for both tests to be valid rely on the prediction performance of the involved conditional mean regressions (for instance, $\mathbb{E}[Y \mid Z]$ and $\mathbb{E}[X \mid Z]$ for the GCM test) and are given in [Shah and Peters, 2020, Theorem 6] and [Lundborg et al., 2024, Theorem 4], for the GCM and PCM tests respectively. A less technical exposition of COMETs can be found in Kook and Lundborg [2024].

3 Falsifying protein interaction networks

We apply the proposed falsification approach to two proposed causal graphical models of the interactions between eleven proteins: A consensus graph based on biological domain knowledge and a graph due to Sachs et al. [2005] (see Figure 1). To conduct the tests, we use a publicly available single cell flow cytometry dataset [Sachs et al., 2005] and the the `comets` [Kook and Lundborg, 2024], `ranger` [Wright and Ziegler, 2017], and `dagitty` [Textor et al., 2016] R packages, for testing, random forests, and listing CI relations, respectively. The R code to reproduce all results is available at `https://github.com/LucasKook/fcm-iwsm`.

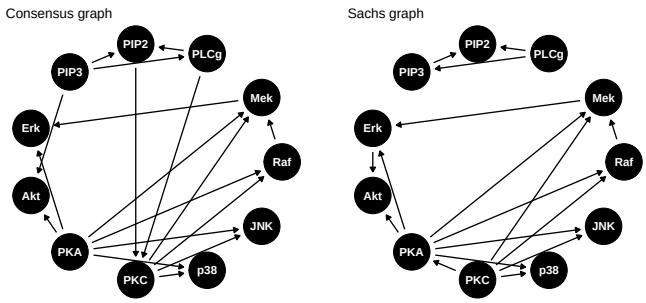

FIGURE 1. Consensus (left) and Sachs (right) graph summarizing the putative causal relations between the proteins.

Under the causal Markov condition [Peters et al., 2017, Definition 6.21], the two graphs imply conditional independence relations in the observational distribution. To limit the number of tests (and thus preserve power when applying multiple testing corrections), we enumerate at least one CI relation per missing edge in the graph with the smallest non-overlapping and non-empty conditioning sets. This results in 55 CI relations implied by the consensus graph and 22 CI relations implied by the Sachs graph. Using the observational dataset (853 observations of log-transformed concentrations for all 11 proteins), we test those CI relations using both the GCM and the PCM test with random forest regressions and adjust for multiple testing using a Holm correction separately for each graph.

For the consensus graph, both the GCM and PCM test reject the same five out of 55 CI relations at the 5% level, thereby falsifying the graph. Four of those CI relations involve the conditional independence between Akt and Erk and the following conditioning sets: {Mek, PKA}, {PKA, PKC}, {PIP2, PKA, PLCg}, and {PIP3, PKA}. The last rejected CI relation is JNK independent of p38 given PKA and PKC. In the Sachs graph, the same (and only this one) CI relation is rejected at the 5% level, again consistently by both the GCM and PCM test (see Table 1). Thus, COMETs can identify CI relations in both graphs that are inconsistent with the observed data.

Besides the observational data, Sachs et al. [2005] provide several interventional datasets in which individual proteins (Akt, PIP2, Erk, PKC, PIP3) were perturbed. Assuming that these experimental conditions correspond to perfect interventions, the graph describing the interventional distribution can be constructed from a given graph by removing all edges pointing to the intervened node [Pearl, 2009]. Under this assumption, the interventional datasets can serve as (pseudo-) replications for assessing the validity of the results obtained on the observational dataset: For all interventional datasets, the GCM (and, except for one case, also the PCM) test consistently rejects only one CI relation implied by the resulting interventional graph, namely that JNK is independent of p38 given PKA and PKC, corroborating the results obtained on the observational data (Table 1).

TABLE 1. Adjusted p-values for the conditional independence relations rejected at the 5%-level (after a Holm-adjustment separately for each graph) using the GCM or PCM test with random forest regressions. For the interventional datasets, the results are only shown for the Sachs graph.

Intervention	**Graph**	**CI Relation**	**GCM**	**PCM**
	Consensus	Akt $\perp\!\!\!\perp$ Erk | Mek, PKA	<0.0001	<0.0001
	Consensus	Akt $\perp\!\!\!\perp$ Erk | PKA, PKC	<0.0001	<0.0001
	Consensus	Akt $\perp\!\!\!\perp$ Erk | PIP2, PKA, PLCg	<0.0001	<0.0001
	Consensus	Akt $\perp\!\!\!\perp$ Erk | PIP3, PKA	<0.0001	<0.0001
	Consensus	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	<0.0001
	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	<0.0001
Akt	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	0.002
PIP2	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	<0.0001
Erk	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	0.002
PKC	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	1
PIP3	Sachs	JNK $\perp\!\!\!\perp$ p38 | PKA, PKC	<0.0001	<0.0001

4 Discussion and conclusion

Using statistical modelling to answer causal research questions is a popular and promising approach, but often relies on strong assumptions. If these assumptions are not met, the resulting causal inferences may be invalid and lead to potentially harmful decisions. We propose to falsify conjectured causal models by nonparametrically testing their implied CI relations with COMETs based on observational (or, under stronger assumptions, also interventional) data. We demonstrate the effectiveness of our approach on a well-studied protein signaling pathway and single-cell flow cytometry data, by falsifying multiple CI relations in the biological consensus graph and a single relation in the graph due to Sachs et al. [2005]. Further analyses based on the available interventional datasets underline the robustness of our conclusions that the Sachs graph still contains a CI relation that is inconsistent with the data across several experimental settings.

References

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. (2019). Sensitivity Analysis of Linear Structural Causal Models. In: *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1252 – 1261. PMLR.

Dawid, A.P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society B*, **41**(1):1 – 15.

Kook, L., and Lundborg, A.R. (2024). Algorithm-Agnostic Significance Testing in Supervised Learning With Multimodal Data. *Briefings in Bioinformatics*, **25**(6).

Lundborg, A.R., Kim, I., Shah, R.D., and Samworth, R.J. (2024). The Projected Covariance Measure for Assumption-Lean Variable Significance Testing. *The Annals of Statistics*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**(5):688 – 701.

Sachs, K., Perez, O., Pe'er, O., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, **308**(5721):523 – 529.

Shah, R.D. and Peters, J. (2020). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*, **48**(3):1514 – 1538.

Su, Z. and Henckel, L. (2022). A Robustness Test for Estimating Total Effects With Covariate Adjustment. In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1886 – 1895. PMLR.

Textor, J., van der Zander, B., Gilthorpe M.S., Lískiewicz, M., and Ellison G.T. (2016) Robust Causal Inference Using Directed Acyclic Graphs: The R Package `dagitty`. *International Journal of Epidemiology*, **45**(6):1887 – 1894.

Wright, M.N. and Ziegler, A. (2017). `ranger`: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**(1):1 – 17.