# Contrasting identification criteria of average causal effects: Asymptotic variances and semiparametric estimators

**Tetiana Gorbach[1], Xavier de Luna[2], Juha Karvanen[3], Ingeborg Waernbaum[4]**

[1] *Department of Statistics, Umeå University, Sweden, tetiana.gorbach@umu.se*
[2] *Department of Statistics, Umeå University, Sweden, xavier.de.luna@umu.se*
[3] *Department of Mathematics and Statistics, University of Jyväskylä, Finland, juha.t.karvanen@jyu.fi*
[4] *Department of Statistics, Uppsala University, Sweden, ingeborg.waernbaum@statistik.uu.se*

Pre-treatment covariates are commonly used to estimate an average causal effect from observational data under the back-door identification of the effect. The effect can also be estimated using mediators only or mediators and pre-treatment covariates under the front-door or the so-called two-door identification, respectively. When several of these identification assumptions are fulfilled, the choice of the estimation strategy may be based, among others, on estimation efficiency.

In this talk we provide the semiparametric efficiency bounds for regular estimators of the average causal effect under the front-door and the two-door identification assumptions specified using the potential outcome framework. We also derive the efficiency bounds when at least two of the identification assumptions are fulfilled simultaneously. We compare the bounds and show that neither the back-door, the front-door, nor the two-door identification assumptions yield the lowest bound irrespective of the data distribution. We, however, provide sufficient conditions for the variance bound of the estimation using information on the mediators and the pre-treatment covariates to be lower (or higher) than the bound of the estimation using information on the pre-treatment covariates only. Estimators reaching the different bounds are also proposed and studied.

# Stabilizing variable selection and regression

**Niklas Pfister**

*University of Copenhagen, Denmark, np@math.ku.dk*

A common setup in many data-driven scientific fields is to observe data across several different experiments or environments. Often these experiments are explicitly constructed to ensure that parts of the data generating mechanism change – the hope being that correlation structures that are detectable across multiple different settings are more likely to be causal. In this talk, we will introduce a formal framework for this type of analysis by considering a multi-environment regression setting in which a response $Y$ is regressed on a set of predictors $X$. We will show that we can gain additional insights into the causal structure between $Y$ and $X$ by distinguishing between stable and unstable predictors (i.e., predictors which have a fixed or a changing functional dependence on the response, respectively). We apply these ideas to hypothesis generation in multiomic data.

This talk is based on joint work with Evan G. Williams, Jonas Peters, Ruedi Aebersold and Peter Bühlmann [1].

## References

[1] Pfister, N., Williams, E. G., Peters, J., Aebersold, R. and Bühlmann, P. (2021). Stabilizing variable selection and regression. *Annals of Applied Statistics*, (to appear).

# Identifying causal effects via context-specific independence relations

**Santtu Tikka**[1], **Antti Hyttinen**[2] **and Juha Karvanen**[1]

[1] *University of Jyväkylä, Finland, santtu.tikka@jyu.fi, juha.t.karvanen@jyu.fi*
[2] *University of Helsinki, Finland, antti.hyttinen@helsinki.fi*

Graphical models are an important aspect of causal inference. In a typical setting of causal effect identification, the causal model is represented by a directed acyclic graph (DAG) which completely characterizes the conditional independence properties exhibited by the available data via d-separation. However, a more general type of independence known as context-specific independence (CSI) cannot be represented via DAGs. In the simplest case, CSI means that $X$ and $Y$ are independent given $Z = 0$ but dependent given $Z = 1$. In other words, $X$ and $Y$ are independent in the context $Z = 0$. More generally, context refers to a set of variables that have been assigned to some values.

In this talk, we use a graphical model known as labeled directed acyclic graph (LDAG) for taking CSI relations into account and show that deciding causal effect non-identifiability is NP-hard in the presence of CSI relations. We present a calculus similar to standard do-calculus for causal effect identification in LDAGs and a search procedure over the rules of the calculus [1]. The search procedure has been implement a as a part of R package `dosearch` [2]. With the approach we can obtain identifying formulas that were unobtainable previously. We demonstrate that a small number of CSI-relations may be sufficient to turn a previously non-identifiable instance to identifiable.

## References

[1] S. Tikka, A. Hyttinen, J. Karvanen (2019). Identifying causal effects via context-specific independence relations, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, http://papers.nips.cc/paper/8547-identifying-causal-effects-via-context-specific-independence-relations.pdf

[2] S. Tikka, A. Hyttinen, J. Karvanen (2020). *dosearch: Causalausal Effect Identification from Multiple Incomplete Data Sources*, R package version 1.0.6, https://cran.r-project.org/package=dosearch