MASARYK UNIVERSITY

Faculty of Economics and Administration

Essays in econometrics of model uncertainty

Habilitation Thesis

Lukáš Lafférs

Brno 2023

MUNI
ECON

Annotation

Author: Lukáš Lafférs

Title: Essays in econometrics of model uncertainty

Year: 2023

This habilitation thesis deals with the econometrics of model uncertainty, and it consists of four academic papers on this topic. The introductory chapter motivates two different research approaches to dealing with model uncertainty, where the four papers in this collection make a contribution. Each research approach is presented in a separate chapter, starting with an introduction that explains the basic setup, then summarizing the papers' main results and contributions to the literature and concluding with a discussion of limitations and suggestions for future research.

Acknowledgements

I wish to thank my mentors Katarína Pavlíčková, Radoslav Harman, Marian Grendár, Gernot Doppelhofer and Alexei Onatski, who supported me throughout different stages of my studies and career. I am grateful for my fantastic co-authors and curious students, who have helped me learn and expand my knowledge in many ways. I also thank my family and friends for supporting me throughout.

Contents

Introduction

Econometric models are tools that help to merge assumptions and data into conclusions. Economic theory, institutional rules or expert knowledge - all these pieces of information help to construct an *econometric model*, which consists of set of restrictions on observable variables. These restrictions take the form of mathematical statements, can be combined with various statistical techniques in order to analyze data. Unlike in some natural sciences, such as physics, where models describe behaviour of inanimate objects, economic models deal with people whose behaviour is much less predictable and subsequently, there is rarely a strong consensus how an econometric model should look like. This is why model uncertainty has always been at the forefront of econometric research.

There are different strategies for coping with model uncertainty. This collection of essays contributes to two particular streams of literature. Specifically, how to deal with model uncertainty in a high-dimensional setup by using machine learning tools within the treatment effects literature, and how to conduct sensitivity analysis of identifying assumptions in a systematic way using incomplete models.

While the contributions of the papers summarized here are mostly methodological and theoretical, they can be applied to a wide range of economic applications, especially in labor economics, health economics and policy evaluation. Importantly, each paper in this collection showcases the usefulness of the proposed method through an empirically relevant application.

Model uncertainty and machine learning

The first line of thought attempts to choose the best model among the set of (potentially many) competing alternatives. This approach has given rise to a *model selection literature* that explicitly balances model complexity and fit to the data, where, among several models that fit the data similarly, the less complex model is preferred. Examples include Akkaike information criterion or Bayesian information criterion (e.g. Claeskens, Hjort, et al. (2008) for an overview). It has been recognized that if the same dataset is used to simultaneously choose the model and estimate its parameters, the resulting model may be overly optimistic (in terms of fit) and its properties may be poor (Leeb and Pötscher (2005)). In the recent years, more and more data became available which makes the number of possible models very large. This gave rise to high-dimensional methods capable of handling situations where the number of variables may be larger than the sample size (see e.g. Bühlmann and Van De Geer (2011) for an overview). Recent advances in this stream of literature have adopted high-dimensional methods to improve the estimation of a low-dimensional object of interest, such as an average treatment effect of a policy change (Chernozhukov et al. (2018)).

The first two essays in this collection contribute to this stream of literature. The contribution of these papers is that they extend the existing results of Chernozhukov et al. (2018) to two new frameworks, namely mediation analysis and dynamic treatment effects. These two frameworks were not considered in the original Chernozhukov et al. (2018) and the presented two papers therefore constitute a distinct value-added to the literature. Mediation analysis and dynamic treatment effects cover a wide range of useful economic applications and thanks to these papers, high-dimensional data can be utilized which further broaden their empirical appeal.

Farbmacher, Huber, Lafférs, Langen, and Spindler (2022)[1] studies the problem of causal mediation analysis in a high-dimensional setup. The paper makes use of the Double Machine Learning framework (DML) of Chernozhukov et al. (2018) to guide the choice of observed confounders, therefore the choice of the model, in a data-driven way. It proves that the estimators of direct and indirect (mediated) effects based on DML possess good statistical properties, which is also supported via a simulation study. Empirical illustration based on National Longitudinal Survey of Youth data suggests that health insurance coverage effect on general health does not appear to be mediated via more frequent routine checkups.

Bodory, Huber, and Lafférs (2022)[2] considers estimation of dynamic treatment effects in a situation with high-dimensional set of covariates. Instead of effect of a single treatment, it considers effects of a sequence of treatments, where past treatments may influence the receipt of future treatments. The paper justifies the use of DML framework of Chernozhukov et al. (2018) and therefore chooses a set of relevant confounders in a data-driven way using machine learning algorithms. Dynamic treatment effects for specific subgroups is also considered. The performance of estimators are demonstrated using a simulation study and it is illustrated on an evaluation of training sequences provided by the Job Corps programme on employment.

Model uncertainty and incomplete models

A completely different approach to deal with model uncertainty is to make use of least information possible and create a model that is not completely specified. This leads to situations where the object of interest in not point identified but only partially identified (e.g. Tamer (2010) for a review). Even if a sample size goes to infinity, the parameter of interest is still only bounded - there is an interval of values for the parameter and all of them are compatible with model assumptions. There is a tension between the amount of information that is assumed and how much can be learnt, which was summarized by Charles Manski (Manski (2003), p.1):

"Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained."

Empirical examples may include interval estimates for different menus of assumptions. The width of these estimated intervals represents a *model uncertainty* that is different from the statistical uncertainty. Another scope of usefulness of these methods is related to *sensitivity analysis*. It is useful to explore how important different assumptions are by relaxing them and considering the width of the identified set. The last two essays in this collection deal with sensitivity analysis by extending its application to two important classes of problems in econometric practice, namely mediation analysis and sample selection models.

Huber and Lafférs (2022)[3] deals with causal mediation analysis, exploring sensitivity to identifying assumptions that are commonly violated in empirical applications. These include treatment and mediator

[1] Farbmacher, H., Huber, M., Lafférs, L., Langen, H., & Spindler, M. (2022). Causal mediation analysis with double machine learning. The Econometrics Journal, 25(2), 277-300. doi; 10.1093/ectj/utac003. Full-text is available at https://academic.oup.com/ectj/article/25/2/277/6517682.

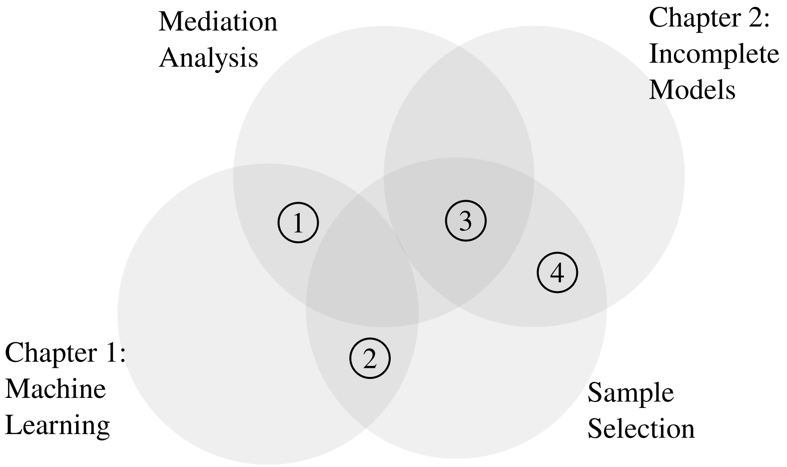
[2] Bodory, H., Huber, M., & Lafférs, L. (2022). Evaluating (weighted) dynamic treatment effects by double machine learning. The Econometrics Journal, 25(3), 628-648. doi; 10.1093/ectj/utac018. Full-text is available at https://academic.oup.com/ectj/article-abstract/25/3/628/6604379.

[3] Huber, M., & Lafférs, L. (2022). Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition. Econometric Reviews, 41(10), 1141-1163. doi; 10.1080/07474938.2022.2127077. Full-text is available at https://www.tandfonline.com/doi/full/10.1080/07474938.2022.2127077.

exogeneity and outcome attrition, even after controlling for observable variables. This paper proposes a method for bounding direct and indirect effects under relaxations of these assumptions. The relaxation parameters are interpretable and varying them we can set the amount of misspecification we wish to consider. The paper provides some suggestions how these can be set in practice, so that the comparisons are meaningful. The method is applied to gender wage gap decomposition using National Longitudinal Survey of Youth 1979 data, where the overall effect is decomposed into direct and indirect (mediated) effect. The sensitivity analysis suggests that the results tends to be sensitive to both outcome attrition and to violations of mediator exogeneity.

Lafférs and Nedela (2017)[4] provides a computational tool to conduct a sensitivity analysis to the method of estimating the average treatment effects under sample selection problem presented in David S Lee. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76 (3): 1071–1102. The empirical applications include for instance a job market program where even under the treatment randomization some people choose not to work and ignoring this selectivity leads to biased estimators. Lafférs and Nedela (2017) controls the departure from the identifying assumptions in Lee (2009) via relaxation parameters that are easy to interpret. One of them is that the proportion of individuals for whom the assignment to the program would have negative effect on their employment (monotonicity assumption), is not set to zero, but bounded from above by some number, say 1%. The method shows that the results presented in Lee (2009) are sensitive even to mild deviations from monotonicity assumption and treatment exogeneity assumption.

The four papers summarized in this collection contribute to the different research areas. Figure 1 schematically depicts the placement of the different papers within the literature.



① Farbmacher, Huber, Lafférs, Langen, and Spindler (2022)
② Bodory, Huber, and Lafférs (2022)
③ Huber and Lafférs (2022)
④ Lafférs and Nedela (2017)

Figure 1: The placement of the different papers within the literature.

[4] Lafférs, L., & Nedela Jr, R. (2017). Sensitivity of the bounds on the ATE in the presence of sample selection. Economics Letters, 158, 84-87. doi; 10.1016/j.econlet.2017.06.039 . Full-text is available at https://www.sciencedirect.com/science/article/abs/pii/S0165176517302665.

Authors' contributions (qualitative and quantitative) are listed in Chapter 4. Complete papers and appendices are provided in Appendix A.[5]

[5] Due to copyright restrictions, the public version of the habilitation thesis does not include the full-texts, links to these papers are in footnotes 1-4.

Chapter 1

Model uncertainty and machine learning

In recent decades machine learning (ML) algorithms have revolutionized many fields of science. The powerful prediction capabilities have facilitated progress in classification, computer vision, pattern recognition, business intelligence, and many other fields. Economics is no exception. While most of ML advances were about prediction, the main interest in economics is about understanding underlying mechanisms, so it is an effect of a certain variable that is often of interest. Thus, estimation is at the forefront of econometric research.

In this chapter, the focus is placed on ML based solutions of dealing with model uncertainty in the context of treatment effects estimation, which is a large but by far not exhaustive part of the econometric literature at the intersection of ML methods and economics. Notable advances has been recently made in the treatment effect heterogeneity (Wager and Athey (2018)) and many other subfields. The usefulness of ML methods in economics was recognized earlier (Varian (2014)) and for a more recent reviews, we refer to Mullainathan and Spiess (2017), Athey and Imbens (2019) or Athey (2019).

Consider the following example. A job-seeker went through a training/course to improve her skill-set. Suppose we wish to estimate a causal effect to find out whether the intervention actually worked, whether it improved her employment chances or raised her wage. We typically have a rich set of information about job-seekers that are managed via employment offices often collected via questionnaires and possibly linked with registry data. The number of potential predictors can be very large and the choice of the model, e.g. which of the predictors are relevant, is challenging. It may be based on institutional knowledge or expert opinion but it is inherently subjective to some extent. An important question is whether the prediction performance of ML based estimators can help to provide unbiased estimators in a high-dimensional setup. Another reason why handling high-dimensional model is useful, is the fact that most models are based on non-experimental data and rely on identifying assumptions. In the context of the job-seeker, treatment assignment is not a result of a randomization but it is a choice and therefore naive comparisons would lead to biased estimates of the true causal effects. We have to rely on selection-on-observables assumption and that controlling for a rich set of characteristics (and e.g. socioeconomic variables and employment history) the people who went through the training are comparable to those that did not. In other words that we control for all or at least most of possible confounders. The more information we have, the more plausible this assumption and, consequently, our conclusions are.

In the rest of this introduction the main ideas of Double Machine Learning (DML) framework (Chernozhukov et al. (2018)) for estimation in high-dimensional setup will be presented. The two presented essays build heavily on these results. This part will present DML framework on a very simple example - estimation of an average treatment effect (ATE) under selection-on-observables assumption (also known as the conditional independence assumption). This particular example is selected because it closely resembles the two setups considered in this chapter.

Double Machine Learning - an illustrative example

Consider a scenario in which we have an outcome of interest Y, a binary treatment variable D, and a high-dimensional vector of covariates X. We use potential outcome notation and therefore the observed outcome is a combination of the two potential outcomes: $Y = Y(1)D + Y(0)(1 - D)$. Suppose that are interested in the average treatment effect of the binary treatment $\Delta = E[Y(1) - Y(0)] = \theta_1 - \theta_0$, where $\theta_d = E[Y(d)]$. Furthermore assume that the vector of covariates X is rich enough so that it makes the treatment D as good as random in terms of it's relationship to potential outcomes (Assumption 1). Also assume that we do not lack comparison units, so that common support assumption holds (Assumption 2).

(1) Conditional independence of D: $\{Y(1), Y(0)\} \perp D \mid X$,

(2) Common support: $\Pr(D = d|X = x) > 0$.

Figure 1.1 shows Direct Acyclic Graph (DAG) (Judea Pearl (2009)) that represents the causal structure of estimation of average treatment effect of a binary treatment in this setup.

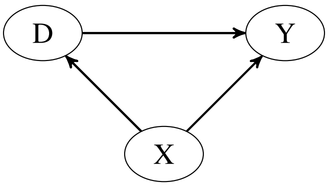


Figure 1.1: The causal structure under the conditional independence assumption of D.

In terms of the identification, the situation is simple. Controlling for X is sufficient for recovering the true potential outcome and subsequently the average treatment effect $\Delta = E[Y(1)] - E[Y(0)] = \theta_1 - \theta_0$. We may base our estimation on either estimating an outcome model $E[Y|D = d, X]$ or a propensity score $\Pr(D = d|X)$.

Under Assumption (1) we have

$$E[Y(d)] = E[E[Y|D = d, X]],$$

and

$$E[Y(d)] = E\left[\frac{Y \cdot I(D = d)}{\Pr(D = d|X)}\right].$$

As long as we can estimate conditional expectation $E[Y|D = d, X]$ or $\Pr(D = d|X)$ then we have two consistent estimators for mean potential outcomes θ_d:

$$\hat{\theta}_d^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y_i|D_i = d, X_i],$$

or

$$\hat{\theta}_d^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot I(D_i = d)}{\hat{\Pr}(D_i = d|X_i)}.$$

ML estimators are biased

The problem is that in the high-dimensional situation, when there are many covariates in X, the estimation may not be feasible if there are too many covariates relative to the sample size. ML estimators, such as lasso, random forest or neural nets, have the possibility to deal with high-dimensional data. The problem is that these estimators are in general biased and their speed of convergence is typically slow. Because they aim to minimize the prediction error, they trade of some bias for the reduction in variance. This property is well known as a bias-variance trade-off (see e.g. James, Witten, Hastie, and Tibshirani (2021)). We refer to this as to *regularization bias*. Neither $\hat{\theta}_d^{(1)}$ nor $\hat{\theta}_d^{(2)}$ will be root-n consistent, under typical assumptions, if they are estimated via ML algorithms.

Another strategy is to assume a linear model for $E[Y|D,X]$ and then interpret the coefficient next to D as a treatment effect. The interpretation of the coefficient to represent a true causal effect hinges upon very restrictive assumptions on treatment heterogeneity (Goldsmith-Pinkham, Hull, and Kolesár (2022)).[1] But even if such model would be correct, the coefficient of D would still be biased if ML estimators are used, that is again due to the bias-variance trade-off.

This example illustrates that even in such a simple situation, the estimation is in general challenging. The main problem is the bias of the ML estimators.

Removing regularization bias

Chernozhukov et al. (2018) built on early results of Neyman (1959), James M Robins and Rotnitzky (1995), Newey (1994) and developed a methodology that can overcome the problem of bias of ML estimators in the context of low-dimensional parameter estimation. The key idea is to construct a moment-condition that is locally-insensitive to some of the bias that comes from the ML estimators, they call this property *Neyman-orthogonality*.

Coming back to the illustrative example: consider two functions for $D = 0, 1$: propensity score function $p_d(X) \equiv \Pr(D = d|X)$ and outcome model $\mu(d, X) \equiv E[Y|D = d, X]$, that are not directly of interest, but they will serve as *nuisance* functions and they can be estimated using ML algorithms. The *target* parameter of interest is the mean potential outcome θ_d. We will now combine the two previous approaches and make use of *both* the propensity score $\Pr(D = d|X)$ and the outcome model $\mu(d, X)$. Consider the following moment function $\psi_d \equiv \psi(\underbrace{Y, D, X}_{\text{data}}; \underbrace{\theta_d}_{\text{target}}, \underbrace{\mu, p_d}_{\text{nuisance}})$ for estimating $\theta_d = E[Y(d)]$, where $\Delta = \theta_1 - \theta_0$ is our object of interest

$$\psi_d = \frac{I\{D = d\} \cdot [Y - \mu(d, X)]}{p(X)} + \mu(d, X) - \theta_d.$$

This moment function satisfies $E[\psi_d] = 0$. It is also doubly-robust (James M Robins and Rotnitzky (1995)), so if *either* the propensity score $\Pr(D = d|X)$ *or* the outcome model $\mu(d, X)$ is correct, then $\theta_d = E[Y(d)]$ holds. A key property of this moment function ψ_d is that of *Neyman-orthogonality*, so that in the vincinity of the true nuisance functions μ and p_d and the true target parameter θ_d, the moment function does not change.[2]

Chernozhukov et al. (2018) showed that the asymptotic behavior of $\hat{\theta}_d$ can be studied using the following expression:

$$\sqrt{n}(\hat{\theta}_d - \theta_d) = \underbrace{a^*}_{\text{Approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

[1] Goldsmith-Pinkham, Hull, and Kolesár (2022) provide an detailed discussion of this problem, which they call a *contamination bias* that mainly arises in the case of multiple valued discrete treatment.

[2] Formally

$$\frac{\partial}{\partial r} E[\psi(Y, D, X; \theta_d, \eta_0 + r(\eta - \eta_0))]\Big|_{r=0} = 0,$$

where $\eta = (\mu, p_d)$ and subscript 0 denotes the true value.

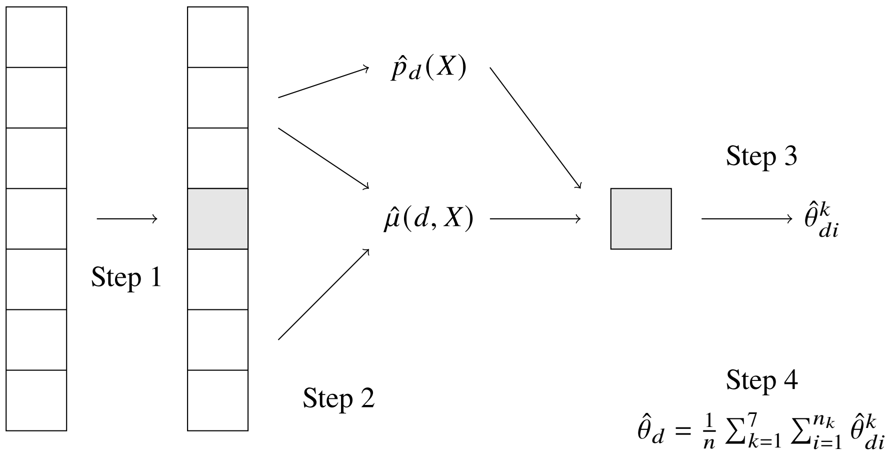

$\hat{p}_d(X)$
Step 3
Step 1
$\hat{\mu}(d, X)$
$\hat{\theta}^k_{di}$
Step 2
Step 4
$\hat{\theta}_d = \frac{1}{n}\sum_{k=1}^{7}\sum_{i=1}^{n_k}\hat{\theta}^k_{di}$

Figure 1.2: Visualization of 7-fold cross fitting procedure. Dataset is split randomly into 7 parts. For each split, only one fold (approximately 1/7 of the whole dataset) is used for the estimation of target parameter, while the rest of the dataset is used for estimation of nuisance parameters, that are used as plug-in estimators. Eventually, estimator $\hat{\theta}_d$ is calculated as an average across these 7 folds.

If the $\hat{\theta}_d$ is based on a Neyman-orthogonal moment function, then the second term vanishes under mild conditions placed on the quality of the nuisance parameter estimators.[3] This is because in this asymptotic expansion $\sqrt{n}(\hat{\theta}_d - \theta_d)$, there is a product of estimation errors of both $\hat{p}_d - p_d$ and $\hat{\mu} - \mu$, so even the ML estimators or $\hat{p}_d$ and $\hat{\mu}$ converge slowly, their product converge faster than square root of n, which is what makes the regularization term b^* go to zero. This is key for achieving the root-n consistency of the estimator.

Removing overfitting bias

The last term c^* arises due to the fact that the same dataset is used for the estimation of nuisance parameters μ, p_d and also the target parameter θ_d. An easy solution is to split the data sample randomly in half. Use one half to construct estimators $\hat{\mu}, \hat{p}_d$ and the other half to estimate $\hat{\theta}_d$ using $\hat{\mu}, \hat{p}_d$ as an input. A more efficient variant is called a *cross-fitting* procedure which is based on switching the roles of the data sample for estimating the nuisance functions and the target parameter and then taking the average over the effect estimates in either part. This procedure is depicted in Figure 1.2.[4]

Overview

Double Machine Learning framework of Chernozhukov et al. (2018) provides a general framework to estimate lower dimensional parameters without a bias in a high-dimensional setup, thus (at least partially) addressing the problem of econometric model specification. Using cross-fitting, if the moment function used for estimation is Neyman-orthogonal we have a square-root n consistent estimator that is asymptotically normally distributed, thus making statistical inference straightforward. This is especially suitable for treatment effects estimation.

The following essays leverage the usefulness of the DML framework and adapt it to two specific situations. The first one is mediation analysis and the second one is dynamic treatment effects.

[3] More precisely, they are satisfied if the ML estimators converge at rate $o(n^{-1/4})$, which is not a strong requirement and is satisfied for many commonly used ML estimators under specific conditions, such as approximate sparsity of η_0 for lasso, well-approximability of η_0 with trees for random forest, see Chernozhukov et al. (2018) for more examples.

[4] E.g. see section 3 in Chernozhukov et al. (2018) for a formal definition.

1.1 Causal mediation analysis with double machine learning (Farbmacher, Huber, Laffers, Langen, and Spindler 2022)

In many situations, we not only wish to estimate a treatment effect but also discover and quantify different causal mechanisms. Mediation analysis attempts to disentangle the various channels through which a treatment operates. There may be a direct effect of the treatment but also an indirect effect, that is the effect that operates through a variable called a mediator. As an example we may be interested whether the effect of education on health operate through employment or through health behaviour. If the mediator is endogenous, such effect decomposition is not possible without additional assumptions, even if the treatment is fully randomized (Rosenbaum (1984), Robins and Greenland (1992)). In order to identify the causal effect that operates via mediator we need to control for potential confounders of the mediator.

Early literature is build on linear and highly-restrictive linear models Cochran (1957), Judd and Kenny (1981). Later on semi-parametric and non-parametric models were considered while identification mostly relied on selection-on-observables assumption Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and Laan (2006), VanderWeele (2009). Within economics the empirical examples include but are not limited to e.g. Flores and Flores-Lagunes (2009), Heckman, Pinto, and Savelyev (2013), Keele, Tingley, and Yamamoto (2015), Conti, Heckman, and Pinto (2016), Huber (2015) or Huber, Lechner, and Mellace (2017). These papers are all based on selection-on-observables assumption, thus it requires selection of relevant confounders based on institutional knowledge and other theoretical considerations. This necessarily brings some degree of ambiguity to the modelling. The present paper addresses the model uncertainty that is implicit in this covariate choice by employing a data-driven selection from the set of high-dimensional vector of potential confounders.

Contribution of this paper is that it proves that the efficient score functions for direct and indirect effects of Tchetgen Tchetgen and Shpitser (2012) are Neyman-orthogonal and therefore DML framework of Chernozhukov et al. (2018) can be applied. This approach requires the estimation of density of a mediator, which is not feasible or problematic if mediator is a vector of multiple variables. In such case, we provide an alternative moment condition that is also Neyman-orthogonal, yet it avoids conditional mediator density estimation.

We provide a simulation study that explores the finite sample behaviour of the suggested estimators and show that it performs well in our simulation design with approximate sparsity.

As an empirical illustration we study health effects of health insurance coverage. We decompose the total effect into a direct effect and the indirect effect that operates via more regular check-ups. Our results suggest that this is not an important channel of the total effect.

In the following subsections we will show the setup, present the main results and then conclude.

1.1.1 Setup

We adopt the potential outcomes framework with Y as the observed outcome, M as the observed mediator, and D as the treatment. The potential outcome is a function of both the treatment and the mediator. For the observed outcome and mediator it holds that $Y = D \cdot Y(1, M(1)) + (1-D) \cdot Y(0, M(0))$ and $M = D \cdot M(1) + (1-D) \cdot M(0)$, thus we either observe $Y = Y(1, M(1))$ and $M = M(1)$ if $D = 1$ or $Y = Y(0, M(0))$ and $M = M(0)$ if $D = 0$.

The natural direct effect $\theta(d)$ stands for the effect in situation when the mediator is being fixed while it takes its natural value:

$$\theta(d) \quad = \quad E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}.$$

The (average) indirect effect, $\delta(d)$, equals the difference in mean potential outcomes when switching the potential mediator values while keeping the treatment fixed to block the direct effect.

$$\delta(d) \quad = \quad E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}.$$

In some case, another effect may be of interest. The controlled direct effect refers to the effect that arises when the value of the mediator is fixed at a certain level for the entire population (see Pearl (2001) for further discussion).

$$\gamma(m) = E[Y(1, m) - Y(0, m], \quad m \in \mathcal{M}.$$

The (total) average treatment effect $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$ can be decomposed into the direct and indirect effect:

$$\begin{aligned} \Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \end{aligned} \tag{1.1}$$

The following assumptions are made in order to identify the effects of interest. The first assumption states that conditioning on X makes the treatment as good as randomly assigned in terms on its influence on both the outcome and the mediator.

Assumption A1 *(conditional independence of the treatment)*
$\{Y(d', m), M(d)\} \perp D | X = x$ *for all* $d', d \in \{0, 1\}$ *and* x *in the support of* X,
where '$\perp$*' denotes statistical independence.*

The second assumption requires the mediator to be conditionally independent of the potential outcomes given the treatment and the covariates.

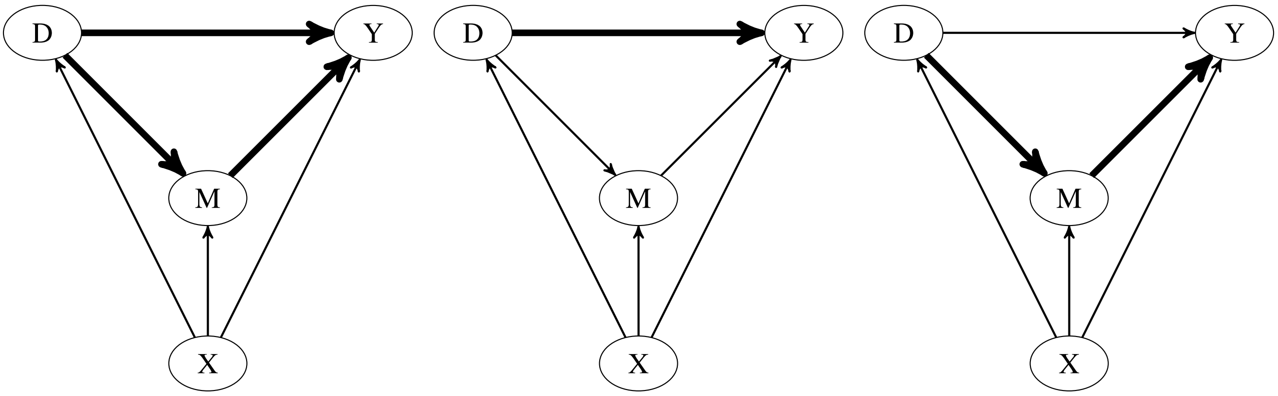
Assumption A2 *(conditional independence of the mediator)*
newline $Y(d', m) \perp M | D = d, X = x$ *for all* $d', d \in \{0, 1\}$ *and* m, x *in the support of* M, X.

Assumption A2 states that there are no confounders jointly affecting the mediator and the outcome conditional on D and X. If X is a vector of pre-treatment variables (in order to avoid "bad controls" problem Cinelli, Forney, and Pearl (2020)), this means that there are no post-treatment confounders of the mediator-outcome relation. The validity of such a restriction needs depends on the empirical context and needs to be scrutinized. It is typically less plausible if the time window between the measurement of the treatment and the mediator is large, as many potential confounders may vary in time.

The third assumption assumes common support on the conditional treatment probability across different treatment states.

Assumption CS *(common support):*
$\Pr(D = d | M = m, X = x) > 0$ *for all* $d \in \{0, 1\}$ *and* m, x *in the support of* M, X.

The directed acyclic graphs that represent the causal structure and the different effects are visualized in Figure 1.3. It is worth noting that the plausibility of these assumptions depends heavily on the richness of covariate vector X which makes the high-dimensional setup particularly appealing.



(a) Total effect Δ. **(b)** The natural direct effect $\theta(d)$. **(c)** The indirect effect, $\delta(d)$.

Figure 1.3: Directed acyclic graph under conditional exogeneity given pre-treatment covariates X.

1.1.2 Main results

We introduce additional notation, we use μ for the outcome model, p_d for the treatment propensity, which is a function of either X or M and X, and ω to represent a conditional outcome model. These nuisance functions are estimated via ML estimators and then plugged into the moment functions using the cross-fitting algorithm.

$$\begin{aligned}
\mu(D, M, X) &= E[Y|D, M, X], \\
p_d(X) &= \Pr(D = d|X), \\
p_d(M, X) &= \Pr(D = d|M, X), \\
\omega(1-d, X) &= E\Big[\mu(d, M, X)|D = 1-d, X\Big],
\end{aligned}$$

also let $f(m|D, X)$ be the conditional density of M given D and X (if M is discrete, this is $f(m|D, X) = \Pr(M = m|D, X)$ and integrals need to be replaced by sums).

In order to estimate direct effects $\theta(d)$, indirect effects $\delta(d)$, and controlled direct effects $\gamma(m)$ we need to estimate $E[Y(d, M(d))]$, $E[Y(d, M(1-d))]$ and $E[Y(d, m)]$. These moment conditions are used in the estimation using the DML framework:

$$\begin{aligned}
E[Y(d, M(d))] &= E\left[\frac{I\{D = d\} \cdot [Y - \mu(d, X)]}{p_d(X)} + \mu(d, X)\right] \\
E[Y(d, M(1-d))] &= E\Bigg[\frac{I\{D = d\} \cdot f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} \cdot [Y - \mu(d, M, X)] \\
&\quad + \frac{I\{D = 1-d\}}{1 - p_d(X)} \cdot \Big[\mu(d, M, X) - \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm\Big] \\
&\quad + \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm\Bigg]
\end{aligned}$$

$$
\begin{aligned}
E[Y(d,M(1-d))] = E\Bigg[&\frac{I\{D=d\}(1-p_d(M,X))}{p_d(M,X)\cdot(1-p_d(X))}\cdot[Y-\mu(d,M,X)] + \frac{I\{D=1-d\}}{1-p_d(X)} \\
&\cdot\Big[\mu(d,M,X) - \frac{1}{1-p_d(X)}\cdot E\Big[\mu(d,M,X)\cdot(1-p_d(M,X))\Big|X\Big]\Big] \\
&+ E\Big[\mu(d,M,X)\cdot\frac{1-p_d(M,X)}{1-p_d(X)}\Big|X\Big]\Bigg] \\
E[Y(d,m)] = E\Bigg[&\frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu(d,m,X)]}{f(m|d,X)\cdot p_d(m,X)} + \mu(d,m,X)\Bigg]
\end{aligned}
$$

There are two formulations for $E[Y(d,M(1-d))]$, the first one is the efficient score function of Tchetgen Tchetgen and Shpitser (2012), we prefer the second one as it avoids the conditional mediator density estimation.

The following statement provides a high-level summary of the results in Farbmacher, Huber, Lafférs, Langen, and Spindler (2022).[5]

Under identifying assumptions A1, A2, CS, along with various regularity conditions, the K-fold cross-fitting algorithm based on moment conditions above provides estimators of $E[Y(d,M(d))]$, $E[Y(d,M(1-d))]$ and $E[Y(d,m)]$, and subsequently those of $\theta(d)$, $\delta(d)$ and $\gamma(m)$, that are root-n consistent and asymptotically normal.

The proofs are based on the DML framework, and they require showing that the moment conditions are Neyman-orthogonal and that various regularity conditions hold. Most of these are mild technical assumptions with the exception on those placed on the quality of the ML estimators. Specifically, the ML estimators of the nuisance functions μ, p_d need to converge at the rate $o(n^{-1/4})$, which is a weaker rate than the usual parametric rate $n^{-1/2}$. Convergence rates of various ML estimators are derived in the literature based on more primitive assumptions, see Chernozhukov et al. (2018) for examples.

We also investigate a finite sample behaviour based on the following simple data generating process.

$$
\begin{aligned}
Y &= 0.5D + 0.5M + 0.5DM + X'\beta + U, \\
M &= I\{0.5D + X'\beta + V > 0\}, \quad D = I\{X'\beta + W > 0\}, \\
X &\sim N(0,\Sigma), \quad U,V,W \sim N(0,1) \text{ independently of each other and } X.
\end{aligned}
$$

where $\Sigma_{ij} = 0.5^{|i-j|}$.

In the simulation, the sample size was set to two different values $n = 1000, 4000$ with 1000 simulations for the data generating process. We chose two different amount of confounding, $\beta = 0.3/i^2$ or $\beta = 0.5/i^2$ for $i = 1, \ldots, 200$. Absolute bias was smaller than 0.01 in all our specifications and quadrupling the sample size cuts the root mean square error roughly to half, which is compatible with the square-root consistency of our main results.

1.1.3 Empirical application

We explored the method on an empirical application based on National Longitudinal Survey of Youth 1997 data in the United States (Bureau of Labor Statistics, U.S. Department of Labor (2001)). The question of interest was if the effect of health insurance coverage (D) on the general health (Y) is mediated via a regular

[5] More concretely Lemma 3.3, Lemma 3.4, Theorem 4.1 and Theorem 4.2.

check-ups (M), see Maciosek, Coffield, Flottemesch, Edwards, and Solberg (2010) for a review. Health insurance coverage is a binary variable based on the 2006 interview, mediator is also a binary variable based on the 2017 interview, where individuals were asked if they had gone for the routine checkup in 2016. The outcome is self reported health - an ordinal variable with levels 'excellent', 'very good', 'good', 'fair', and 'poor'. A wide range of control variables X are based on 2005 or earlier interviews, so that they cannot be influenced by the treatment. It includes demographics characteristics, socio-economic background, education, training, household characteristics, marital status, fertility, received monetary transfers, attitudes, expectations, physical and mental health, nutrition, physical activity. Altogether we have 755 control variables with 593 dummy variables, out of which 251 dummy variables encoded missing information to deal with non-response. The sample size was 7,489. We used post-lasso regression, with three-fold cross fitting and with trimming the propensity scores above 2%. The results suggest that the health care coverage improves general health, so that the direct effect on health is positive and statistically significant. The indirect effect is small and not significant suggesting that the more regular health checkups is not an important channel of this effect.

1.1.4 Conclusions

This paper extended Double Machine Learning framework into the mediation analysis, so that it is possible to explore different causal pathways of treatment effects in a high-dimensional setup. It avoids ad-hoc based model specification and relies on data-driven covariate choice. This performance of the new method is supported by a simulation study and illustrated on an example from health economics.

1.2 Evaluating (weighted) dynamic treatment effects by double machine learning (Bodory, Huber, and Lafférs 2022)

In the introductory section we considered estimation of the effect of a *single* treatment. In many cases, researchers are interested in the effect of *sequences* of treatments. After a job training applicant may attend a language course or after a surgery patient goes through a rehabilitation procedure. The treatment is in many empirically relevant cases a result of a choice (thus non-random) and naive comparisons are not informative of the true causal effects. In practice, we often rely on selection-on-observables assumptions, thus that the available information make the treatment as good as random. In the case with multiple treatments that happen in sequence, we may assume that in each period the treatment assignment is unconfounded given the available information at the time, that is, including the treatments in previous time periods - such assumption is called sequential conditional independence assumption. The validity of such assumption heavily depends on the richness of the information that the covariates span. In the situation with many covariates, researchers face the challenge how to pick the relevant variables. This is typically done using institutional knowledge or expert opinion, so there is always some amount of ambiguity involved. This paper addresses this concern. It provides a data-driven way of choosing the relevant covariates by combining the semi-parametrically efficient estimation (using efficient score function of Robins, Rotnitzky, and Zhao (1994), J. M. Robins and Rotnitzky (1995)) of dynamic treatment effects with the Double Machine Learning (DML) of Chernozhukov et al. (2018). We formally prove that our approach fits within the DML framework.

In the early literature on dynamic treatment effects Robins (1986) proposed a dynamic causal framework called a g-computation for recursively modeling outcomes under the sequential conditional independence assumption initially implemented by parametric maximum likelihood estimation. Robins (1998) suggested a computationally less expensive alternative representing outcomes in specific treatment states as functions of time-constant covariates only. Given the time-varying confounding, these models need to reweighted by the inverse of the dynamic treatment propensity scores (Robins, Greenland, and Hu (1999) or Robins, Hernan, and Brumback (2000)). More recently, Lechner (2009) considered inverse probability weighting (IPW) by the dynamic treatment propensity scores alone, while Lechner and Miquel (2010) apply propensity score matching and Blackwell and Strezhnev (2020) direct matching on the covariates.

The two papers closest to ours are Lewis and Syrgkanis (2020) and Viviano and Bradic (2021). Lewis and Syrgkanis (2020) provide a DML estimation that can be applied to continuous treatments too, but rely on a more restrictive (partially linear) outcome model. Viviano and Bradic (2021) also provides a method that can be combined with ML estimators, but it is based on covariate balancing (Athey, Imbens, and Wager (2018)) instead of inverse propensity score weighting (as it is done in this work).

Following subsections show the setup, present the main results (including simulation study and empirical application) and conclusions.

1.2.1 Setup

Let D_t and Y_t be the treatment and outcome in period $T = t$, respectively. For instance, D_1 and D_2 represent the treatments in the first and second periods, respectively, and can take on values $d_1, d_2 \in 0, 1, ..., Q$, where 0 indicates no treatment, and $1, ..., Q$ represent the various treatment options. Moreover, Y_2 represents the outcome of interest in the second period after the treatment sequence D_1 and D_2 has been applied. We use the potential outcome framework (Rubin (1974)): for a specific treatment sequence $\underline{d}_2 \equiv (d_1, d_2)$ with $d_1, d_2 \in 0, 1, ..., Q$, let $\underline{D}_2 \equiv (D_1, D_2)$, and $Y_2(\underline{d}_2)$ denotes the potential outcome that would be observed if the treatments were set to that sequence $\underline{d}_2$. We also consider the identification for a specific subgroup of interest ($S = 1$), for instance the treated populations in the first period.

The objects of interests are

$$\begin{aligned}\Delta(\underline{d}_2, \underline{d}_2^*) &= E[Y_2(\underline{d}_2)] - E[Y(\underline{d}_2^*)],\\ \Delta(\underline{d}_2, \underline{d}_2^*|S = 1) &= E[Y_2(\underline{d}_2)|S = 1] - E[Y_2(\underline{d}_2^*)|S = 1].\end{aligned}$$

Identification is based on sequential conditional independence assumptions and on the common support assumption. Also, the selection indicator S and the potential outcome are independent conditional on pre-treatment covariates X_0.

Assumption B1 *(conditional independence of the first treatment)*
$Y_2(\underline{d}_2) \perp D_1 | X_0$, *for* $\underline{d}_2 \in \{0, 1, ..., Q\}^2$
where '$\perp$' denotes statistical independence.

Assumption B2 *(conditional independence of the second treatment)*
$Y_2(\underline{d}_2) \perp D_2 | D_1, X_0, X_1$, *for* $\underline{d}_2 \in \{0, 1, ..., Q\}^2$.

Assumption B3 *(common support)*
$\Pr(D_1 = d_1|X_0) > 0$, $\Pr(D_2 = d_2|D_1, \underline{X}_1) > 0$ *for* $d_1, d_2 \in \{0, 1..., Q\}$.

Assumption B4 *(conditional independence of the subgroup indicator)*
$S \perp Y_2(\underline{d}_2)|X_0$, *for* $\underline{d}_2 \in \{0, 1, ..., Q\}^2$.

The causal structure in Figure 1.4 encodes the information in the identifying assumptions.

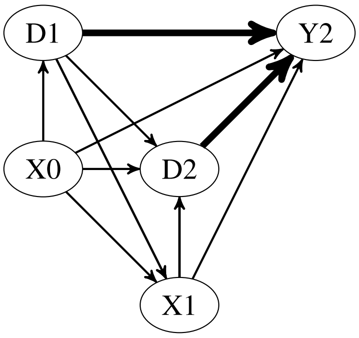


Figure 1.4: Directed acyclic graph under conditional independence of the first treatment given pre-treatment covariates X_0 and under conditional independence of the the second treatment given covariates X_0 and X_1.

1.2.2 Main results

We introduce further notation for the nuisance functions:

$$\begin{aligned}\mu^{Y_2}(\underline{D}_2, \underline{X}_1) &= E[Y_2|\underline{D}_2, X_0, X_1],\\ p^{d_1}(X_0) &= \Pr(D_1 = d_1|X_0),\\ p^{d_2}(D_1, \underline{X}_1) &= \Pr(D_2 = d_2|D_1, \underline{X}_1),\\ \nu^{Y_2}(\underline{D}_2, X_0) &= \int_{x_1 \in X_1} E[Y_2|\underline{D}_2, X_0, X_1 = x_1] dF_{X_1 = x_1|D_1, X_0},\\ g(X_0) &= \Pr(S = 1|X_0).\end{aligned}$$

These are estimated via machine learning estimators and plugged into the following moment conditions that are shown to satisfy Neyman-orthogonal property and thus are locally insensitive to the mild deviations from their true values.

$$\begin{aligned}
E[Y_2(\underline{d}_2)] &= E\Bigg[\frac{I\{D_1 = d_1\} \cdot I\{D_2 = d_2\} \cdot [Y_2 - \mu^{Y_2}(\underline{d}_2, \underline{X}_1)]}{p^{d_1}(X_0) \cdot p^{d_2}(d_1, \underline{X}_1)} \\
&+ \frac{I\{D_1 = d_1\} \cdot [\mu^{Y_2}(\underline{d}_2, \underline{X}_1) - \nu^{Y_2}(\underline{d}_2, X_0)]}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2, X_0)\Bigg]. \\
E[Y_2(\underline{d}_2)|S = 1] &= E\Bigg[\frac{g(X_0)}{\Pr(S = 1)} \cdot \frac{I\{D_1 = d_1\} \cdot I\{D_2 = d_2\} \cdot [Y_2 - \mu^{Y_2}(\underline{d}_2, \underline{X}_1)]}{p^{d_1}(X_0) \cdot p^{d_2}(d_1, \underline{X}_1)} \\
&+ \frac{g(X_0)}{\Pr(S = 1)} \cdot \frac{I\{D_1 = d_1\} \cdot [\mu^{Y_2}(\underline{d}_2, \underline{X}_1) - \nu^{Y_2}(\underline{d}_2, X_0)]}{p^{d_1}(X_0)} \\
&+ \frac{S}{\Pr(S = 1)} \cdot \nu^{Y_2}(\underline{d}_2, X_0)\Bigg].
\end{aligned}$$

Theorem 1 and Theorem 2 from Bodory, Huber, and Lafférs (2022) states that:

> Under identifying assumptions B1-B4, various regularity conditions, the K-fold cross-fitting algorithm based on moment conditions above provides estimators of $E[Y_2(\underline{d}_2)]$ and $E[Y_2(\underline{d}_2)|S = 1]$, and subsequently those of $\Delta(\underline{d}_2, \underline{d}_2^*)$ and $\Delta(\underline{d}_2, \underline{d}_2^*|S = 1)$, are root-n consistent and asymptotically normal.

The paper provides a simulation study with the following design:

$$\begin{aligned}
Y_2 &= D_1 + D_2 + X_0'\beta_{X_0} + X_1'\beta_{X_1} + U, \\
D_1 &= I\{X_0'\beta_{X_0} + V > 0\}, \\
D_2 &= I\{0.3D_1 + X_0'\beta_{X_0} + X_1'\beta_{X_1} + W > 0\}, \\
X_0 &\sim N(0, \Sigma_0), \quad X_1 \sim N(0, \Sigma_1), \\
U, V, W &\sim N(0, 1), \text{ independently of each other.}
\end{aligned}$$

In this design, coefficient vectors β_{X_0} and β_{X_0} determine the degree of confounding. The ith element in the coefficient vectors β_{X_0} and β_{X_1} was set to $0.4/i^4$ for $i = 1, ..., p$, and this quadratic decay is compatible with approximate sparsity condition suitable for the use of the lasso estimator. Two sample sizes of 2,500 and 10,000 were used, running 1,000 simulations for the smaller and 250 simulations for the larger sample sizes. The number of covariates p in X_1 and X_0, respectively, to 50, 100, or 500. Σ_0 and Σ_1 are defined based on setting the covariance of the ith and jth covariate in X_0 or X_1 to $0.5^{|i-j|}$. Based on this specification, the degree of confounding is substantial (Bodory, Huber, and Lafférs (2022), Table 1) and could therefore reasonably mimic empirical applications. We used 3-fold cross-fitting with lasso estimator using the SuperLearner package Van der Laan, Polley, and Hubbard (2007). Table 1.1 shows good performance from the simulations for the subpopulation of treated ($S = I(D_1 = 1)$) in the first period.

covariates	sample size	true effect	absolute bias	standard deviation	average SE	RMSE	coverage in %
		ATE on selected: $\hat{\Delta}(\underline{d}_2, \underline{d}_2^*, S = 1)$					
50	2,500	2	0.027	0.076	0.087	0.081	96.5
50	10,000	2	0.006	0.037	0.043	0.038	95.6
100	2,500	2	0.042	0.079	0.087	0.089	94
100	10,000	2	0.011	0.037	0.043	0.039	96.4
500	2,500	2	0.064	0.075	0.088	0.099	91.5
500	10,000	2	0.019	0.038	0.043	0.043	95.2

Table 1.1: Simulation results based on $\beta_{X_0} = \beta_{X_1} = 0.4/i^4$. Notes: SE and RMSE denote the standard error and the root mean squared error, respectively. Coverage is based on 95% confidence intervals.

1.2.3 Empirical application

The proposed method is applied to evaluate the impact of different sequences of job trainings provided by the U.S. Job Corps program on employment probability. The sample used consisted of 11,313 individuals who completed interviews four years after randomization (6,828 assigned in Job corps, 4,485 randomized out). There are four different treatments: 0 - no instruction (being randomized out), 1 - no instruction (being randomized in), 2 - academic education, 3 - vocational training. After processing there are 909 variables in X_0 and 1,427 variables X_1, most of them dummy variables, that consists of rich set of information on socio-economic characteristics, labor market history, education, trainings, job search activities health, crime, and how one learnt about the existence of Job Corps. Random forest was used for the estimation of nuisance functions. Average treatment effects are estimated in the subsample with first treatment entering one of the treatment sequences compared. We estimated that a sequence of two vocational trainings provides higher employment in comparison to sequences consisting of academic trainings or no trainings (Table 1.2), in the range of 5 to 10 percentage points.

$\underline{d}_2$	$\underline{d}_2^*$	$\hat{E}[Y_2(\underline{d}_2^*)\|S = 1]$	$\hat{\Delta}(\underline{d}_2, \underline{d}_2^*, S = 1)$	SE	p-value	observations	trimmed
33	22	0.76	0.1	0.06	0.11	3783	507
33	21	0.82	0.05	0.03	0.07	3783	43
33	11	0.81	0.08	0.03	0.02	2346	22

Table 1.2: Effect estimates with a trimming threshold of 0.01. Notes: $\underline{d}_2$ and $\underline{d}_2^*$ indicate the treatment sequences under treatment and non-treatment, respectively. $\hat{E}[Y_2(\underline{d}_2^*)|S = 1]$ denotes the mean potential outcome under non-treatment conditional on $S = 1$, where S is an indicator for the first treatment corresponding to either the first treatment in $\underline{d}_2$ or $\underline{d}_2^*$. $\hat{\Delta}(\underline{d}_2, \underline{d}_2^*, S = 1)$ provides the ATE estimate, SE is the standard error.

To assess the validity of these findings, a placebo test was conducted by comparing the effect of two sequences, 00 and 11, where neither sequence involved participation in any training programs. As expected, the estimated effect was very close to zero with a p-value of 0.92.

1.2.4 Conclusions

This paper showed how Double Machine Learning framework may be used to study dynamic treatment effects, that is how to estimate a causal effects of different sequences of treatments based on selection-on-observables assumption. This alleviates the difficulty of the covariate choice in the case with high-dimensional data. The paper demonstrated that the estimators are asymptotically normal and root-n consistent under specific regularity conditions. The simulation study examines the finite sample properties of the estimators and the methodology is applied to the U.S. Job Corps study.

1.3 Limitations and future research avenues

In the mediation analysis framework we assumed that the machine learners are "good enough" in the sense that they achieve the desired rate of convergence. One of them is nested conditional mean: $\omega(1-d,X) = E\Big[\mu(d,M,X)|D=1-d,X]$ and may require a special focus.[6] This is a plug-in estimator and we use additional data-splitting to estimate μ and ω on different subsamples *within* the cross-fitting algorithm, so that the estimation errors in both stages of the estimation are independent by design. It might be interesting to provide some more primitive conditions under which the nested estimator of ω would achieve the desired rate of convergence for the DML. This would most likely have to be specific to a particular ML method used. Such conditions would make the paper more self-contained and could be of independent interest. This problem may constitute a separate research project on its own, as this area appears to be underresearched.

While the method can choose the set of relevant variables to control for in a data-driven way, it is not an automatic tool to estimate effects and still require careful thought in the sense that the (potentially large) list of possible covariates cannot be chosen arbitrarily. The causal structure represented by the Directed Acyclic Graph (DAG) in Figure 1.3 still has to hold true. For instance we have to make sure that we do not include any "bad controls" (Cinelli, Forney, and Pearl (2020)), e.g. variables that would be influenced by the outcome itself and would introduce spurious associations. This is why we chose variables measured *before* the treatment itself.

[6] Similar situation emerged in the dynamic treatment effects framework for ν^{Y_2}.

Chapter 2

Model uncertainty and incomplete models

Consider an econometric model where the parameter of interest is not identified. As a simple example, consider identifying treatment effects under non-random treatment assignment. In such case, without additional assumptions, it is not possible to determine mean potential outcomes as outcomes are only observed in one of the treatment states. Using potential outcomes notation, let Y be an observed outcome which is one of the potential outcomes: $Y = Y(1)D + Y(0)(1 - D)$, where D is a binary treatment. Then quantities $E[Y(1)]$ and $E[Y(0)]$ are not identified as they depend on unobserved quantities $E[Y(1)|D = 0]$ and $E[Y(0)|D = 1]$:

$$\begin{aligned} E[Y(1)] &= E[Y(1)|D=1]\Pr(D=1) + E[Y(1)|D=0]\Pr(D=0) \\ &= E[Y(1)|D=1]\Pr(D=1) + \underbrace{E[Y(1)|D=0]}_{\text{unobserved}}\Pr(D=0), \\ E[Y(0)] &= E[Y(0)|D=1]\Pr(D=1) + E[Y(0)|D=0]\Pr(D=0) \\ &= \underbrace{E[Y(0)|D=1]}_{\text{unobserved}}\Pr(D=1) + E[Y|D=0]\Pr(D=0). \end{aligned}$$

Depending on the context, we may be willing to make assumptions about $E[Y(1)|D = 0]$ and $E[Y(0)|D = 1]$. For instance, if the treatment was random, we would have $E[Y(1)] = E[Y|D = 1]$ and $E[Y(0)] = E[Y|D = 0]$ and therefore the average treatment effect $\Delta = E[Y(1)] - E[Y(0)]$ would be identified. On the other hand, if the only information was that the outcome Y has a finite support, $Y \in [y_{\min}, y_{\max}]$, we would only know that

$$\begin{aligned} E[Y|D=1]\Pr(D=1) + y_{\min}\Pr(D=0) \leq &\quad E[Y(1)] &\leq E[Y|D=1]\Pr(D=1) + y_{\max}\Pr(D=0), \\ y_{\min}\Pr(D=1) + E[Y|D=0]\Pr(D=0) \leq &\quad E[Y(0)] &\leq y_{\max}\Pr(D=1) + E[Y|D=0]\Pr(D=0), \end{aligned}$$

which would translate to *bounds* for $\Delta \in [\Delta_{min}, \Delta_{max}]$. This is called an *identified set*, see Manski (2003) for a book-length presentation and Tamer (2010) for a review on partial identification literature.

The width of the bounds represents the lack of information, in other words, the degree of uncertainty. We could consider different menus of assumptions and explore how the width of the bounds vary, thus better understanding the sources of identification. This may sharpen the discussion as we could focus more on the assumptions that are important.

Suppose we have a point-identified econometric method and we wish to conduct a sensitivity analysis to different identifying assumption. Say, that the selection-on-observables only holds approximately and we wish to explore if a small deviation from this assumption could lead to a very different results. Consider two different scenarios depicted on Figure 2.1. In the first one, a relaxation of an identifying assumption leads to a small

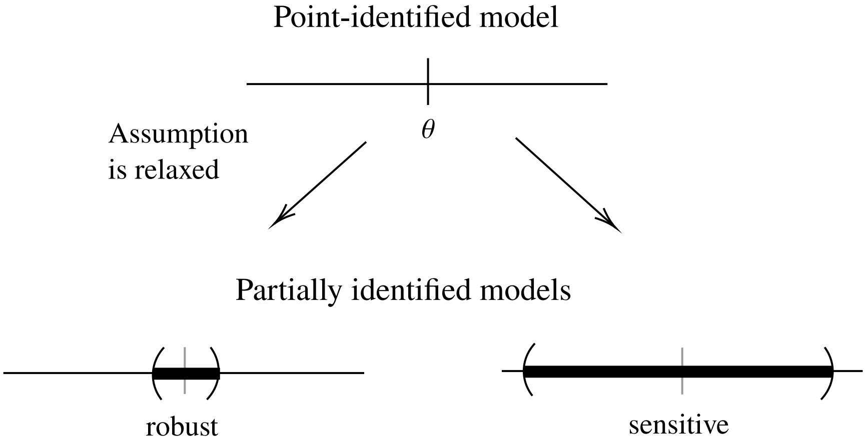


Figure 2.1: A relaxation of an identifying assumption may lead to a narrow identified set (left), indicating a robust result or, to the contrary, to a wide identified set (right).

identified set, thus we may conclude that the effect is robust to violation of this assumption. In the other example, we get a very large identified set, highlighting the importance of this particular assumption.

The following two essays present methods how bounds can be calculated and therefore offer tools to conduct sensitivity analyses. The first paper studies mediation analysis and estimators based on inverse propensity score weighting. The second paper considers the setup of estimating average treatment effects under sample selection problem, where we do not observe outcome variable for non-random proportion of the population.

2.1 Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition (Huber and Laffers 2022)

Similar to Farbmacher, Huber, Lafférs, Langen, and Spindler (2022) presented in Section 1.1, this paper also studies mediation analysis, where the objective is to decompose a total effect into a direct effect and to an effect that is channeled through additional variable(s), called a mediator(s). The motivation, review of the literature and different approaches are presented there. This paper, however, has a different objective. While in Farbmacher, Huber, Lafférs, Langen, and Spindler (2022) the focus was on the model uncertainty that was connected to the variable choice, here we address sensitivity analysis of the effects to the *identifying assumptions*. It these assumptions are relaxed, the direct and indirect effects are no longer identified, and we only get bounds on the effects. It is, however, challenging to come up with relaxations that are easy-to-interpret and, at the same time, it is technically feasible to estimate the bounds. This is the contribution of the paper. It provides a computationally feasible method that calculates bounds under relaxed assumptions and thus allows to conduct sensitivity analysis.

We now add a few notes on the literature on sensitivity analysis within the mediation analysis, which is the stream of literature where this paper contributes. Imai, Keele, and Yamamoto (2010) shows how to conduct sensitivity analysis within highly restrictive linear models, with relaxation parameters as correlation of unobserved terms in the mediation and outcome equations. Tchetgen Tchetgen and Shpitser (2011) proposed a semi-parametric framework that uses a "selection bias function" which relates confounders of the mediator-outcome to the treatment and allows multidimensional unobserved confounders. These ideas were further developed in Vansteelandt and VanderWeele (2012). The paper that is closest to ours in terms of the strategy is Hong, Qin, and Yang (2018). They consider weighting estimators and their core idea is, that if some important confounders are omitted, then this renders the weights, that are actually used in the analysis, to be incorrect. We follow similar line of reasoning, but we use a different measure to limit the errors in weights due to confounding.

In contrast to their method we do not provide analytical formulas but only a computational method, but our method has the advantage that it allows for *simultaneous relaxations* of different assumptions. This allows us to better understand the non-robustness of the results to violations of the various identification assumptions.

The following subsections present the setup and then motivate and explain step-by-step how the computational method was constructed. The methodology is then applied on gender wage gap decomposition on National Longitudinal Survey of Youth 1979 dataset in the United States.

2.1.1 Setup

The setup is very similar to that in Section 1.1.1, where we introduced mediation analysis. There is an additional layer of complication - the outcome is not observed for the whole sample, which we refer to as a *sample selection problem*. The identification or direct and indirect effects is possible based on the assumptions of conditional independence of the treatment (A1), conditional independence of the mediator (A2), common support assumptions (CS)[1] (all stated and discussed in Section 1.1.1) and a new assumption that postulates that the information in D, X, M is rich enough to capture all the possible confounders of selection S and outcome Y:

Assumption A3 *(conditional independence of selection):*
$Y \perp S | D = d, M = m, X = x$ *for all* $d \in \{0, 1\}$ *and* m, x *in the support of* M, X.

Theorem 1 of Huber and Solovyeva (2020a) states that Assumptions A1-A3 allow identifying the mean potential outcomes by

$$\begin{aligned}
E[Y(1, M(1))] &= E\left[Y \cdot D \cdot S \cdot \frac{1}{\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)}\right], \qquad (2.1)\\
E[Y(0, M(0))] &= E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{1-\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)}\right],\\
E[Y(1, M(0))] &= E\left[Y \cdot D \cdot S \cdot \frac{1}{1-\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)} \cdot \left(\frac{1}{\Pr(D=1|M,X)} - 1\right)\right],\\
E[Y(0, M(1))] &= E\left[Y \cdot (1-D) \cdot S \cdot \frac{1}{\Pr(D=1|X)} \cdot \frac{1}{\Pr(S=1|D,M,X)} \cdot \left(\frac{1}{1-\Pr(D=1|M,X)} - 1\right)\right].
\end{aligned}$$

The direct and indirect effects of interest are obtained as differences between two out of the four mean potential outcomes. In order to ease notation, we denote the various propensity scores in (2.1) by

$$p^{A1} = \Pr(D=1|X), \quad p^{A2} = \Pr(D=1|M,X), \quad p^{A3} = \Pr(S=1|D,M,X).$$

The Figure 2.2 represents the causal structure in this problem.

Assumptions A1-A3 are arguably very strong. It is natural to consider sensitivity analysis and ask how important these are and to what extent are the results driven by particular assumptions. This is the research question of the present paper.

2.1.2 Main results

Suppose that we have doubts about assumption A3, e.g. that there exists an important confounder U that jointly influences both the decision to work (S) and wage itself (Y). In such case the *true* propensity score $q^{A3} = \Pr(S=1|D,M,X,U)$ is different from the one that we can estimate, $p^{A3} = \Pr(S=1|D,M,X)$. This

[1] With an extra assumption that $\Pr(S=1|D=d, M=m, X=x) > 0$, for d, m, x in their support.

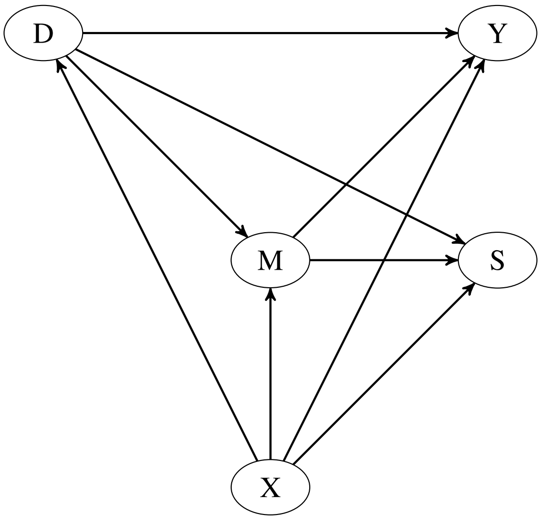


Figure 2.2: Causal paths under conditional exogeneity and missing at random given pre-treatment covariates

motivates the approach in this paper: consider all the possible true probabilities q^{A3} that are not too different from observable probabily p^{A3}. In an ideal situation the distance between the two could be relaxed by some interpretable parameter. We consider the following relaxation via a parameter ϵ^{A3} :

$$|q^{A3} - p^{A3}| \quad \leq \quad \epsilon^{A3}\sqrt{p^{A3}(1 - p^{A3})}.$$

The scaling term $\sqrt{p^{A3}(1 - p^{A3})}$ serves two purposes. Firstly, it ensures that the distances between the true propensity scores and the estimated propensity scores are comparable across different values of p^{A3}, such as 0.99 or 0.5. Secondly, it is symmetric, which means that distances from $p^{A3} = 0.95$ and $p^{A3} = 0.05$ are treated similarly.

For the sake of exposition, we now consider the identification and estimation of $E[Y(1, M(1))]$. Under Assumptions A1-A3, we get that

$$E[Y(1, M(1))] = E\left[\frac{Y \cdot D \cdot S}{\Pr(D = 1|X) \cdot \Pr(S = 1|D, M, X)}\right] = E\left[\frac{Y \cdot D \cdot S}{p^{A1} \cdot p^{A3}}\right],$$

but in case that some confounder (or a vector of confounders) U implies violations of A1 and A2, then the true value of $E[Y(1, M(1))]$ is instead

$$E[Y(1, M(1))] = E\left[\frac{Y \cdot D \cdot S}{\Pr(D = 1|X, U) \cdot \Pr(S = 1|D, M, X, U)}\right] = E\left[\frac{Y \cdot D \cdot S}{q^{A1} \cdot q^{A3}}\right],$$

where $q^{A1} = \Pr(D = 1|X, U)$ and ϵ^{A1} is defined analogously to ϵ^{A3}.

It is therefore in principle possible to find bounds on $E[Y(1, M(1))]$ solving the following (population) optimization problem:

$$\begin{aligned} \min_{q^{A1}, q^{A3}}/\max E[Y(1, M(1))] \quad &= \quad E\left[\frac{Y \cdot D \cdot S}{q^{A1} \cdot q^{A3}}\right]. \\ &s.t. \\ |q^{A1} - p^{A1}| \quad &\leq \quad \epsilon^{A1}\sqrt{p^{A1}(1 - p^{A1})}, \\ |q^{A3} - p^{A3}| \quad &\leq \quad \epsilon^{A3}\sqrt{p^{A3}(1 - p^{A3})}. \end{aligned}$$

The finite sample counterpart is this optimization problem:

$$\begin{aligned} \min/\max_{q^{A1},q^{A3}} \quad & \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \frac{1}{q_i^{A1}} \cdot \frac{1}{q_i^{A3}} \Big/ \sum_{i=1}^{n} \frac{D_i}{q_i^{A1}} \frac{S_i}{q_i^{A3}} \\ & s.t. \\ \forall i: |q_i^{A1} - \hat{p}_i^{A1}| \quad \leq \quad & \epsilon^{A1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})}, \quad |q_i^{A3} - \hat{p}_i^{A3}| \leq \epsilon^{A3}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}, \\ q_i^{A1} \in [0,1], \quad & q_i^{A3} \in [0,1]. \end{aligned}$$

It is not immediately clear how this problem would be solved in practice, especially if the sample size is large. Luckily, after some manipulations, this optimization problem can be shown to be equivalent to the following optimization problem, which is a linear program and therefore computationally attractive:

$$\begin{aligned} \min/\max_{\omega,t} \quad & \sum_{i=1}^{n} Y_i \cdot D_i \cdot S_i \cdot \omega_i \\ & s.t. \\ \forall i: \omega_i \quad \leq \quad & t \Big/ \left(\left(\hat{p}_i^{A1} - \epsilon^{A1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} - \epsilon^{A3}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\ \omega_i \quad \geq \quad & t \Big/ \left(\left(\hat{p}_i^{A1} + \epsilon^{A1}\sqrt{\hat{p}_i^{A1}(1-\hat{p}_i^{A1})} \right) \cdot \left(\hat{p}_i^{A3} + \epsilon^{A3}\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})} \right) \right) \\ \sum_{i=1}^{n} D_i \cdot S_i \cdot \omega_i \quad = \quad & 1, \quad \omega_i \geq 0, \quad t \geq 0. \end{aligned}$$

There is one important challenge that remains and that is that how to set the relaxation parameters ϵ^{A1} and ϵ^{A3} in a meaningful way. The get a value of ϵ^{A3} that is interpretable, consider the following approach. Suppose that we remove the most important predictor (e.g. in the sense of reduction in deviance) in X in a logit regression of S on D, M, X, and denote these predictions as $\hat{p}_{i,X1}^{A3}$. This would lead to the following scaled in-sample differences: $\epsilon_{i,X1}^{A3} = \frac{|\hat{p}_{i,X1}^{A3} - \hat{p}_i^{A3}|}{\sqrt{\hat{p}_i^{A3}(1-\hat{p}_i^{A3})}}$. One way to set ϵ_{X1}^{A3} is to calculate an average of $\epsilon_{i,X1}^{A3}$ in the subpopulation with $D_i = 1$ and $S_i = 1$:

$$\epsilon_{X1}^{A3} = \sum_{i=1}^{n} \frac{D_i \cdot S_i \cdot \epsilon_{i,X1}^{A3}}{\sum_{i=1}^{n} D_i \cdot S_i}.$$

In a similar way, ϵ_{Xj}^{A3} and ϵ_{Mj}^{A3} could be calculated based on the omission of the j-th most important predictor from X and M.

2.1.3 Empirical application

The paper looks at the decomposition of the U.S. gender wage gap using data from the National Longitudinal Survey of Youth (1979). Huber and Solovyeva (2020b) considered five different wage decomposition techniques to examine the sensitivity of the direct/indirect effects estimators, see also Huber (2015) for a discussion on identification issues in mediation analysis. Dataset consists of 6,658 observations, D is gender (0 - female, 1 - male), Y is logaritm of average hourly wage during one year. Selection indicator S is set to one for people who worked at least 1,000 hours (about 80% of the sample). Vector of mediators M consists of variables that were measured after the birth and *before* the treatment itself such as education, employment variables, occupation, marital status, length of marriage, regional dummies, history of health problems and others. Conditioning covariates in X consists of variables that were determined prior to birth such as race, religion, year of birth, birth order, parental place of birth and parental education. There still might exist confounders that we do not observe and that may cause violation of exogeneity assumptions, such as risk preferences, attitudes towards competition, motivation or other socio-psychological factors.

Table 2.1 shows the importance of different predictors in the propensity score estimator. They are used to set up the relaxation parameters, that correspond to average deviation in propensity score that missing the 1st, 2nd and the 3rd most important regressor from X or M would lead to.

Assumption A1		$P(D = 1\|X)$
Most important X	1st	Mothers educ. missing
	2nd	Mothers educ. high school graduate
	3rd	Religion missing
Assumption A2		$P(D = 1\|M, X)$
Most important M	1st	Farmer or laborer
	2nd	Industry: Professional services
	3rd	Clerical occupation
Most important X	1st	White
	2nd	Fathers educ. college/more
	3rd	Mothers educ. missing
Assumption A3		$P(S = 1\|D, M, X)$
Most important M	1st	Employed full time
	2nd	Employment status: employed
	3rd	Operator (machines, transport)
Most important X	1st	Fathers educ. college/more
	2nd	Mothers educ. some college
	3rd	Protestant

Table 2.1: Covariates and mediators with the highest predictive power in propensity score estimations measured by the change in deviance.

Considering the male (or female) wages as the reference, one can take $\Delta = \theta(0) + \delta(1)$ (or $\Delta = \theta(1) + \delta(0)$) as the preferred decomposition of direct (unexplained) and indirect (explained) effects, see Sloczynski (2013) for a detailed discussion.

Several interesting lessons can be learnt: the bounds are in general not symmetric around the point identified effect. Also, the most important regressor in terms of prediction in the propensity score estimation need not be the one that leads to the widest bounds. By applying the proposed methods, we find that the omission of a confounder, that has the same predictive power as the first or second most important mediator entering the treatment propensity score, would render all the natural indirect effects for males insignificant. The effect of the choice of the link function does not matter in most of the specifications. Most importantly, bounds for the natural direct effects (for males and females), thus the unexplained component of the gender wage gap decomposition, do not include zero and are highly significant for the most of the specifications, except for some of those where an important mediator was missing, which leads to a violation of conditional mediator exogeneity. This provides a robust evidence on the existence of unexplained gender wage gap.

2.1.4 Conclusions

This paper presented a computationally feasible method to study sensitivity to identifying assumptions in a mediation analysis with a sample selection problem. This method was applied to NLSY 1979 dataset to study gender wage gap decomposition (outcomes are only observable for those who work), where the results are sensitive to non-ignorable mediator selection and to sample selection.

There are different ways how the sensitivity analysis may be done. There is a trade-off between having an assumption relaxation that is elegant and easy to interpret and, at the same time, the method is still computationally feasible. This paper attempted to find a fine balance between the two.

2.2 Sensitivity of the bounds on the ATE in the presence of sample selection (Lafférs and Nedela 2017)

In many empirical applications, the outcome is not observed for a specific subpopulation. Ignoring this may lead to misleading results if the missingness is not random - problem that is called a *sample selection bias*. There are different ways how sample selection bias can be addressed. Earlier literature was built around restrictive parametric models (Heckman (1979)), and typically, some exogenous variation in the outcome missingness mechanism is needed to identify causal treatment effects. Lee (2009), an influential paper, took a different approach. Lacking an external instrument and not willing to assume a parametric structure on the problem, Lee estimated only bounds on treatment effects instead.

This note builds up on these results. It shows how bounding the average treatment effects under sample selection (Lee (2009)) can be reformulated as a optimization problem. This gives insights into the original problem but more importantly it provides opportunity to conduct sensitivity analysis with respect to some violations of the identifying assumptions of treatment exogeneity and monotonicity, which are relaxed in a novel interpretable manner.

This note contributes to the liteature on the crossroads of identification and linear programming, that was pioneered by Balke and Pearl (1997) and later explored by many others (e.g. Honoré and Tamer (2006)). Formulating identification as an optimization problem is a principled approach that has also been explored in Lafférs (2019).

2.2.1 Setup and main results

Let Y be the outcome only observed if $S = 1$, D be a binary treatment and let us use the potential outcome notation (Rubin (1974)). Lee (2009) studied the effect of U.S. Job Corps training on wages, where wages are observed only for those who work. These are the identifying assumptions:[2]

(C1) Compatibility with the observed data: $Y = S \cdot \left(Y^1 D + Y^0(1-D)\right)$, and $S = S^1 D + S^0(1-D)$;
(C2) Treatment exogeneity: D is statistically independent of $\left(Y^1, Y^0, S^1, S^0\right)$,
(C3) Monotonicity of selection in D: $S^1 \geq S^0$ with probability 1.

Assumption (C2) is satisfied if treatment is randomized and Assumption (C3) states that treatment cannot have a negative effect on the selection.

In his study, Lee (2009) derived analytical formulas for the following quantity $E\left(Y^1 - Y^0 | S^1 = 1, S^0 = 1\right)$, that is the average treatment effect of "always-takers", so those who would be employed in both states when treated and non-treated, thus capturing the *pure wage effect*.

[2] While (C1) is only implicitly assumed in Lee (2009), for the approach in this paper, it is useful to state it explicitly.

This note looks at the identification problem as to a optimization problem that searches through the space of probability distributions. Let Φ be the set of all probability distributions of π of $U = (Y^1, Y^0, S^1, S^0, D)$ that satisfy (C1)-(C3):

$$\Phi = \{\pi \in \Pi(\mathcal{U}) : \pi \text{ satisfies (C1)-(C3)}\}.$$

The problem of bounding the average treatment effects of the always-takers is equivalent to finding the probability distribution π that solves the subsequent optimization problem:

$$\min/\max_{\pi \in \Phi} E_\pi\big(Y^1 - Y^0 | S^1 = 1, S^0 = 1\big),$$

where the dependence on π was made explicit. Condition (C1) states that π has to be compatible with observable distributon of (Y, S, D). This optimization problem can be formulated as a linear program and it replicates the analytical formulas derived in Lee (2009).

Suppose that researcher would be concerned about the validity of the Assumption (C2). The motivation is that even under randomization there is often a non-response/sample attrition problem. In the dataset of Lee (2009) the non-response rate is over 40% and if the sample attrition is correlated with the treatment assignment, and the outcome then Assumption (C2) is violated. Furthermore, Assumption (C3) could also be violated if treatment group members who undergo training choose to wait for a better job.

We can relax the exogeneity and monotonicity assumptions using relaxation parameters α_E, α_M in the following way:

(rC2) Relaxed exogeneity:
Total variation distance between distributions $\pi(.|D = 0)$ and $\pi(.|D = 1)$ is smaller than α_E :

$$TV\big(\pi(.|D = 0), \pi(.|D = 1)\big) \leq \alpha_E.$$

(rC3) Relaxed monotonicity: No more than α_M proportion of population violate monotonicity:

$$\pi(S^1 \geq S^0) \geq 1 - \alpha_M.$$

The total variation distance is the maximum of absolute difference between $\pi(A|D = 0)$ and $\pi(A|D = 1)$ across all possible events A, therefore (rC2) puts an upper bound on how different these two distributions could be.

Reformulating the relaxed problem as a *tractable* optimization problem is not straightforward, because the share of always-takes $\pi(S^1 = 1, S^0 = 1)$ is no longer identified under the relaxed assumptions. It can be, however, treated as an additional unknown free variable in the optimization. With some manipulation and reparametrization, this relaxed problem can also be formulated as a linear program.

2.2.2 Empirical application

We apply this methodology to the original Lee (2009) dataset and simultaneously considered relaxation of both exogeneity and monotonicity. Outcome variable was discretized[3] and subsampling was used for confidence bounds (Politis, Romano, and Wolf (1999)). A mild deviation from identifying assumptions $\alpha_E = \alpha_M = 0.01$ doubles the width of the identified set, while larger deviations $\alpha_E = \alpha_M = 0.05$ lead to very wide bounds, where

[3] The discretization for $\alpha_E = \alpha_M = 0$, where we have analytical formulas, led to an error of order only 10^{-5}.

the lower bound drops from -1.7% to -29%. This suggests that the results are very sensitive to the identifying assumptions.

Sensitivity Analysis of the Bounds on ATE_C
[Lower bound, Upper bound]
(90% confidence bounds)

			α_M	
		0.00	0.01	0.05
	0.00	[-0.0171, 0.0931]	[-0.0545, 0.1217]	[-0.1286, 0.2036]
		(-0.0252, 0.1043)	(-0.0663, 0.1333)	(-0.1431, 0.2179)
α_E	0.01	[-0.0539, 0.1270]	[-0.0871, 0.1541]	[-0.1607, 0.2359]
		(-0.0664, 0.137)	(-0.0985, 0.1667)	(-0.1752, 0.2518)
	0.05	[-0.1821, 0.254]	[-0.2113, 0.2807]	[-0.2893, 0.3641]
		(-0.2009, 0.2688)	(-0.2266, 0.2952)	(-0.309, 0.3796)

Table 2.2: Bounds on the average treatment effect of a Job Corps program on log hourly wages for individuals who would be employed regardless of treatment status.

2.2.3 Conclusions

We provide a method for sensitivity analysis of bounds on ATE under the sample selection. Assumptions are allowed to be relaxed simultaneously using interpretable parameters. This method can be applied whenever Lee (2009) bounds are estimated.

2.3 Limitations and future research avenues

While the methods presented in this chapter can give interesting insights into empirical practice, applied researchers are often hesitant to try new methods, especially if they are computationally based. Providing well-tuned implementations in popular software packages, coupled with detailed documentation, may bring these methods closer to practitioners.

The sensitivity method presented in Huber and Lafférs (2022) can readily be applied to *any* estimator that is based on inverse propensity score weighting and potentially extending its scope of usefulness.

The total variation distance metric that was used to relax exogeneity is both interpretable and computationally convenient and it has a potential to be used within a different context too.

Chapter 3

Authorship contribution statements

The authors' list is in alphabetical order.

Causal mediation analysis with double machine learning (Farbmacher, Huber, Lafférs, Langen, and Spindler 2022)

Quantitative authorship contribution of Lukáš Lafférs: 20%.

- Corresponding author: Henrika Langen
- Conceptualization: Helmut Farbmacher, Martin Huber, Henrika Langen, Lukáš Lafférs, Martin Spindler
- Methodology: Helmut Farbmacher, Martin Huber, Henrika Langen, Lukáš Lafférs, Martin Spindler
- Software: Martin Huber, Henrika Langen
- Data curation: Martin Huber, Henrika Langen
- Data analysis: Martin Huber, Henrika Langen
- Writing (original draft): Helmut Farbmacher, Martin Huber, Henrika Langen, Lukáš Lafférs, Martin Spindler
- Writing (review): Helmut Farbmacher, Martin Huber, Henrika Langen, Lukáš Lafférs, Martin Spindler

Evaluating (weighted) dynamic treatment effects by double machine learning (Bodory, Huber, and Lafférs 2022)

Quantitative authorship contribution of Lukáš Lafférs: 33%.

- Corresponding author: Lukáš Lafférs
- Conceptualization: Hugo Bodory, Martin Huber, Lukáš Lafférs
- Methodology: Hugo Bodory, Martin Huber, Lukáš Lafférs
- Software: Hugo Bodory, Martin Huber
- Data curation: Hugo Bodory, Martin Huber
- Data analysis: Hugo Bodory

- Writing (original draft): Hugo Bodory, Martin Huber, Lukáš Lafférs
- Writing (review): Hugo Bodory, Martin Huber, Lukáš Lafférs

Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition (Huber and Lafférs 2022)

Quantitative authorship contribution of Lukáš Lafférs: 50%.

- Corresponding author: Lukáš Lafférs
- Conceptualization: Martin Huber, Lukáš Lafférs
- Methodology: Martin Huber, Lukáš Lafférs
- Software: Lukáš Lafférs
- Data curation: Martin Huber
- Data analysis: Lukáš Lafférs
- Writing (original draft): Martin Huber, Lukáš Lafférs
- Writing (review): Martin Huber, Lukáš Lafférs

Sensitivity of the bounds on the ATE in the presence of sample selection (Lafférs and Nedela 2017)

Quantitative authorship contribution of Lukáš Lafférs: 50%.

- Corresponding author: Lukáš Lafférs
- Conceptualization: Lukáš Lafférs, Roman Nedela
- Methodology: Lukáš Lafférs, Roman Nedela
- Software: Lukáš Lafférs, Roman Nedela
- Data analysis: Lukáš Lafférs, Roman Nedela
- Writing (original draft): Lukáš Lafférs, Roman Nedela
- Writing (review): Lukáš Lafférs

Bibliography

Athey, Susan. 2019. "21. The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence,* 507–552. University of Chicago Press.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics* 11:685–725.

Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." *Journal of the Royal Statistical Society Series B* 80:597–623.

Balke, Alexander, and Judea Pearl. 1997. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association* 92 (439): 1171–1176.

Blackwell, Matthew, and Anton Strezhnev. 2020. "Telescope Matching for Reducing Model Dependence in the Estimation of the Effects of Time-varying Treatments: An Application to Negative Advertising." *working paper, Harvard University.*

Bodory, Hugo, Martin Huber, and Lukáš Lafférs. 2022. "Evaluating (weighted) dynamic treatment effects by double machine learning." *The Econometrics Journal* 25 (3): 628–648.

Bühlmann, Peter, and Sara Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Bureau of Labor Statistics, U.S. Department of Labor. 2001. *National Longitudinal Survey of Youth 1979 cohort, 1979-2000 (rounds 1-19).* Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning." *The Econometrics Journal* 21 (1).

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. "A crash course in good and bad controls." *Sociological Methods & Research,* 00491241221099552.

Claeskens, Gerda, Nils Lid Hjort, et al. 2008. "Model selection and model averaging." *Cambridge Books.*

Cochran, William G. 1957. "Analysis of Covariance: Its Nature and Uses." *Biometrics* 13:261–281.

Conti, Gabriella, James J. Heckman, and Rodrigo Pinto. 2016. "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour." *The Economic Journal* 126:F28–F65.

Farbmacher, Helmut, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. 2022. "Causal mediation analysis with double machine learning." *The Econometrics Journal* 25 (2): 277–300.

Flores, Carlos A., and Alfonso Flores-Lagunes. 2009. “Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness.” *IZA DP No. 4237.*

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2022. *Contamination bias in linear regressions.* Technical report. National Bureau of Economic Research.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. “Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes.” *American Economic Review* 103:2052–2086.

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47 (1): 153–161.

Hong, Guanglei, Xu Qin, and Fan Yang. 2018. “Weighting-Based Sensitivity Analysis in Causal Mediation Studies.” *Journal of Educational and Behavioral Statistics* 43:32–56.

Honoré, Bo E, and Elie Tamer. 2006. “Bounds on parameters in panel dynamic discrete choice models.” *Econometrica* 74 (3): 611–629.

Huber, Martin. 2015. “Causal pitfalls in the decomposition of wage gaps.” *Journal of Business and Economic Statistics* 33:179–191.

Huber, Martin, and Lukáš Lafférs. 2022. “Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition.” *Econometric Reviews* 41 (10): 1141–1163.

Huber, Martin, Michael Lechner, and Giovanni Mellace. 2017. “Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism.” *The Review of Economics and Statistics* 99:180–183.

Huber, Martin, and Anna Solovyeva. 2020a. “Direct and indirect effects under sample selection and outcome attrition.” *Econometrics* 8 (4): 44.

———. 2020b. “On the sensitivity of wage gap decompositions.” *Journal of Labor Research* 41:1–33.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects.” *Statistical Science* 25:51–71.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. “An Introduction to Statistical Learning: With Applications in R.”

Judd, C M, and D A Kenny. 1981. “Process Analysis: Estimating Mediation in Treatment Evaluations.” *Evaluation Review* 5:602–619.

Keele, Luke, Dustin Tingley, and Teppei Yamamoto. 2015. “Identifying mechanisms behind policy interventions via causal mediation analysis.” *Journal of Policy Analysis and Management* 34:937–963.

Lafférs, Lukáš. 2019. “Identification in models with discrete variables.” *Computational Economics* 53 (2): 657–696.

Lafférs, Lukáš, and Roman Nedela. 2017. “Sensitivity of the bounds on the ATE in the presence of sample selection.” *Economics Letters* 158:84–87.

Lechner, M. 2009. “Sequential Causal Models for the Evaluation of Labor Market Programs.” *Journal of Business and Economic Statistics* 27:71–83.

Lechner, Michael, and Ruth Miquel. 2010. “Identification of the effects of dynamic treatments by sequential conditional independence assumptions.” *Empirical Economics* 39:111–137.

Lee, David S. 2009. “Training, wages, and sample selection: Estimating sharp bounds on treatment effects.” *The Review of Economic Studies* 76 (3): 1071–1102.

Leeb, Hannes, and Benedikt M Pötscher. 2005. “Model selection and inference: Facts and fiction.” *Econometric Theory* 21 (1): 21–59.

Lewis, Greg, and Vasilis Syrgkanis. 2020. “Double/Debiased Machine Learning for Dynamic Treatment Effects.” *arXiv preprint 2002.07285.*

Maciosek, Michael V, Ashley B Coffield, Thomas J Flottemesch, Nichol M Edwards, and Leif I Solberg. 2010. “Greater use of preventive services in US health care could save lives at little or no cost.” *Health Affairs* 29 (9): 1656–1660.

Manski, Charles F. 2003. *Partial identification of probability distributions.* Vol. 5. Springer.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine learning: an applied econometric approach.” *Journal of Economic Perspectives* 31 (2): 87–106.

Newey, Whitney K. 1994. “The asymptotic variance of semiparametric estimators.” *Econometrica: Journal of the Econometric Society,* 1349–1382.

Neyman, Jerzy. 1959. “Optimal asymptotic tests of composite hypotheses.” *Probability and statistics,* 213–234.

Pearl, J. 2001. “Direct and indirect effects.” In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* 411–420. San Francisco: Morgan Kaufman.

Pearl, Judea. 2009. *Causality.* Cambridge university press.

Petersen, M L, S E Sinisi, and M J van der Laan. 2006. “Estimation of Direct Causal Effects.” *Epidemiology* 17:276–284.

Politis, DN, JP Romano, and M Wolf. 1999. *Subsampling.*

Robins, J M. 1986. “A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect.” *Mathematical Modelling* 7:1393–1512.

———. 1998. “Marginal Structural Models.” In *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science,* 1–10.

———. 2003. “Semantics of causal DAG models and the identification of direct and indirect effects.” In *In Highly Structured Stochastic Systems,* edited by P.J. Green, N.L. Hjort, and S. Richardson, 70–81. Oxford: Oxford University Press.

Robins, J M, and Sander Greenland. 1992. “Identifiability and Exchangeability for Direct and Indirect Effects.” *Epidemiology* 3:143–155.

Robins, J M, M A Hernan, and B Brumback. 2000. “Marginal Structural Models and Causal Inference in Epidemiology.” *Epidemiology* 11:550–560.

Robins, J. M., S. Greenland, and F.-C. Hu. 1999. “Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome.” *Journal of the American Statistical Association* 94:687–700.

Robins, J. M., A. Rotnitzky, and L.P. Zhao. 1994. “Estimation of Regression Coefficients When Some Regressors Are not Always Observed.” *Journal of the American Statistical Association* 90:846–866.

Robins, J. M., and Andrea Rotnitzky. 1995. “Semiparametric Efficiency in Multivariate Regression Models with Missing Data.” *Journal of the American Statistical Association* 90:122–129.

Robins, James M, and Andrea Rotnitzky. 1995. “Semiparametric efficiency in multivariate regression models with missing data.” *Journal of the American Statistical Association* 90 (429): 122–129.

Rosenbaum, P. 1984. “The consequences of adjustment for a concomitant variable that has been affected by the treatment.” *Journal of Royal Statistical Society, Series A* 147:656–666.

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

Sloczynski, T. 2013. “Population average gender effects.” *IZA Discussion Paper No. 7315.*

Tamer, Elie. 2010. “Partial identification in econometrics.” *Annu. Rev. Econ.* 2 (1): 167–195.

Tchetgen Tchetgen, E. J., and I. Shpitser. 2011. “Semiparametric Estimation of Models for Natural Direct and Indirect Effects.” *Harvard University Biostatistics Working Paper 129.*

———. 2012. “Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis.” *The Annals of Statistics* 40:1816–1845.

Van der Laan, Mark J, Eric C Polley, and Alan E Hubbard. 2007. “Super learner.” *Statistical applications in genetics and molecular biology* 6 (1).

VanderWeele, Tyler J. 2009. “Marginal Structural Models for the Estimation of Direct and Indirect Effects.” *Epidemiology* 20:18–26.

Vansteelandt, S., and T. J. VanderWeele. 2012. “Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions.” *Biometrics* 68:1019–1027.

Varian, Hal R. 2014. “Big data: New tricks for econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28.

Viviano, Davide, and Jelena Bradic. 2021. “Dynamic covariate balancing:estimating treatment effects over time.” *arXiv preprint 2103.01280.*

Wager, Stefan, and Susan Athey. 2018. “Estimation and inference of heterogeneous treatment effects using random forests.” *Journal of the American Statistical Association* 113 (523): 1228–1242.

Appendix A

Essays

A.1 Causal mediation analysis with double machine learning (Farbmacher, Huber, Lafférs, Langen, and Spindler 2022)

A.2 Evaluating (weighted) dynamic treatment effects by double machine learning (Bodory, Huber, and Lafférs 2022)

A.3 Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition (Huber and Laffërs 2022)

A.4 Sensitivity of the bounds on the ATE in the presence of sample selection (Lafférs and Nedela 2017)