UMEÅ UNIVERSITY

Valid causal inference in high-dimensional and complex settings

Niloofar Moosavi

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå 2022

Doctoral Thesis
Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå University
901 87 Umeå, Sweden

Statistical Studies No. 56
ISBN 978-91-7855-881-0 (print)
ISBN 978-91-7855-882-7 (pdf)
ISSN 1100-8989
Electronic version available at `http://umu.diva-portal.org`

Printed by: cityprint i Norr AB
Umeå, Sweden 2022

Contents

List of papers

The following papers are included in the thesis:

I. Moosavi, N., Häggström, J. and de Luna, X. (2021). The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. *To appear in Statistical Science, ArXiv preprint arXiv:2105.02071.*

II. Moosavi, N., Gorbach, T. and de Luna, X. (2022). Valid causal inference: model selection and sensitivity to unobserved confounding in high-dimensional settings. *Submitted.*

III. Ghasempour, M., Moosavi, N.and de Luna, X. (2022). Convolutional neural networks for valid and efficient causal inference. *Working paper.*

IV. Moosavi, N., Gorbach, T. and de Luna, X. (2022). A note on sensitivity analysis for post-machine learning causal inference. *Working paper.*

Abstract

The objective of this thesis is to consider some challenges that arise when conducting causal inference based on observational data. High dimensionality can occur when it is necessary to adjust for many covariates, and flexible models must be used to meet convergence assumptions. The latter may require the use of a novel machine learning estimator. Estimating nonparametrically-defined causal estimands at parametric rates and obtaining good-quality confidence intervals (with near nominal coverage) are the primary goals. Another challenge is providing a sensitivity analysis that can be applied in high-dimensional scenarios as a way of assessing the robustness of the results to missing confounders.

Four papers are included in the thesis. A common theme in all the papers is covariate selection or nonparametric estimation of nuisance models. To provide insight into the performance of the approaches presented, some theoretical results are provided. Additionally, simulation studies are reported. In paper I, covariate selection is discussed as a method for removing redundant variables. This approach is compared to other strategies for variable selection that ensure reasonable confidence interval coverage. Paper II integrates variable selection into a sensitivity analysis, where the sensitivity parameter is the conditional correlation of the outcome and treatment variables. The validity of the analysis where the sensitivity parameter is small relative to the sample size is shown theoretically. In simulation settings, however, the analysis performs as expected, even for larger values of sensitivity parameters, when using a correction of the estimator of the residual variance for the outcome model. Paper IV extends the applicability of the sensitivity analysis method through the use of a different residual variance estimator and applies it to a real study of the effects of smoking during pregnancy on child birth weight. A real data problem of analysing the effect of early retirement on health outcomes is studied in Paper III. Rather than using variable selection strategies, convolutional neural networks are studied to fit the nuisance models.

Keywords. Causal inference, high dimension, sensitivity analysis, variable selection, convolutional neural network, semiparametric efficiency bound

Sammanfattning (Summary in Swedish)

Syftet med denna avhandling är att överväga några utmaningar som uppstår när man drar kausala slutsatser baserat på observationsdata. Hög dimensionalitet kan uppstå när det är nödvändigt att justera för många kovariater, och flexibla modeller måste användas för att möta konvergensantaganden. Det senare kan kräva användning av en ny maskininlärningalgoritm. Att skatta icke-parametriskt definierade kausala parametrar vid parametriska hastigheter och erhålla konfidensintervall av god kvalitet (med nära nominell täckning) är de primära målen. En annan utmaning är att tillhandahålla en känslighetsanalys som kan tillämpas i högdimensionella scenarier som ett sätt att bedöma resultatens robusthet för ej observerade störfaktorer.

Fyra artiklar ingår i avhandlingen. Ett gemensamt tema i dessa artiklar är val av kovariat eller icke-parametrisk skattning av underliggande modeller. För att ge insikt i hur de utvecklade metoderna fungerar ges några teoretiska resultat. Dessutom rapporteras simuleringsstudier. I artikel I diskuteras kovariatselektion som en metod för att ta bort onödiga variabler. Detta tillvägagångssätt jämförs med andra strategier för variabelval som säkerställer rimlig täckning av konfidensintervall. Paper II integrerar variabelval i en känslighetsanalys, där sensitivitetsparametern är den betingade korrelationen mellan utfalls- och behandlingsvariablerna. Validiteten av analysen där känslighetsparametern är liten i förhållande till stickprovsstorleken visas teoretiskt. I simuleringstudien fungerar emellertid analysen som förväntat, även för större värden på känslighetsparametrar, när man använder en korrigering av den skattade residualvariansen för utfallsmodellen. Paper IV utökar användbarheten av känslighetsanalysmetoden genom att använda en annan skattning av residualvariansen och tillämpar den på en studie av effekterna av rökning under graviditeten på barnets födelsevikt. Registerdata används för att analysera effekten av förtidspensionering på hälsan i Paper III. Istället för att använda variabelval studeras konvolutionella neurala nätverk för att anpassa underliggande modellerna.

Preface

The journey of doing a PhD has been quite a ride. It has been a great opportunity getting to experience such a different environment compared to my home country. The first time experiences I had these years were a lot of fascinating ones, including downhill skiing (and even daring to cross the bump on a children's track) in beautiful Hemavan, cross-country skiing on the Umeå river, skating on frozen Tavelsjö and many more.

I have to admit I was a bit sceptical on the first day at work when my mentor, Gabriel, showed me the PhD room and said something like: This is where you are going to spend the next four years of your life (or maybe this is just how I heard it). However, taking the job actually turned out to be not the worst decision I have ever made, if not the best. I mean what is boring about biking home from work watching northern lights or trying out strong Swedish food at Johan's house (In more precise terms, we were in their yard, but the smell got in and one of the kids cried out his strong feelings)? At last, a special mention goes to IKSU, which made a couch potato enjoy sweating.

In spite of this, I have been experiencing the toughest years of my life. I lost my adorable little nephew, Roham. Corona had prevented me from seeing my family for over a year and I had already missed him deeply. It wouldn't have been possible for me to get through this rough time without the support and help of my dear husband, my close friend Navid, as well as family members, friends and colleagues. Thank you all.

It is not simply the content of my thesis that I will take away from these years, but rather a great deal of personal and professional development. My supervisor, Xavier de Luna, has been a pleasure to work with and learn from. I would like to thank him for giving me the opportunity, giving me the confidence, and guiding me along the way. Also, I must thank my co-supervisor, Jenny Häggström, for her valuable assistance. In addition, I want to thank my co-authors, Tetiana Gorbach and Mohammad Ghasempour, for their positive and friendly collaboration.

Thanks to colleagues and fellow students for being very kind, sharing experiences and providing generous mental support. I wish you all the best.

Niloofar Moosavi
Umeå, September 2022

1. Introduction

The focus of this thesis is to make inferences about the causal effects of interventions. It is common to conduct random experiments in order to study causality. The randomization of treatments or interventions is, however, not always feasible or ethical. Moreover, we cannot take advantage of some large observational data sets that are available, such as Swedish register data, by relying only on randomization.

Deriving causal inferences from an observational study involves addressing different challenges. The researcher typically must adjust for a set of pretreatment covariates in order to avoid biased estimation due to non-comparability between treated and control groups. This set of variables is used to model the potential outcomes of different treatment levels and/or the probability of treatment. It is often assumed that the potential outcomes are independent of treatment given the covariates (i.e., given an assumption of no unobserved confounding). However, the size of the set of covariates might be larger than the sample size (high dimensionality). Moreover, to get consistent regression fits, many higher-order terms might be used. In such high-dimensional cases, variable selection is unavoidable. Variable selection as a step in estimation of a causal parameter is discussed in Paper I.

In Papers II and IV, we use a sensitivity analysis to analyze the effect of violation of the assumption of no unobserved confounders. In a sensitivity model, conditional correlations between outcomes and treatment are used to quantify deviations from unconfoundedness. The possibility of weakening some parametric modelling assumptions, previously considered in the sensitivity model, is investigated under some regularity conditions.

A real data problem of analyzing the effect of early retirement on health outcomes is studied in Paper III. To ensure unconfoundedness, we account for several pretreatment covariates, such as ten-year measurements of both hospitalization and outpatient health care, annual income from work and pension, and annual income from unemployment programs. In order to detect time-invariant patterns in the data, convolutional neural networks are used.

The introductory part of the thesis is organized as follows. A brief introduction to causal inference and estimators is given in Section 2. In Section 3, efficiency considerations are discussed. In Sections 4 and 5, variable selection and other complexities are addressed. In Section

6, sensitivity analysis is discussed, and, in Section 7, the papers are summarized.

2. Causal inference

Correlation was introduced in the 19th century (e.g. Stigler, 1989) and has been the subject of a great deal of debate regarding its interpretation in relation to causality. It is possible to interpret correlation in an unsatisfactory manner, as illustrated by the Simpson paradox, in which correlation between two random variables has the opposite sign in sub-populations compared to the whole population. The fact that correlation does not always imply causation is well known today. For this reason, and in order to achieve valid causal inferences, randomization and inference based on randomization were advocated by Fisher (Fisher, 1992). Randomization ensures that those who receive treatment and those who do not represent the same population. When treatment is not randomized, differences in pretreatment covariates could account for the difference in the outcome between treated and control groups. In this regard, Fisher speculated that an observed association between smoking and lung cancer might be explained by an unobserved genetic variant (Fisher, 1958). For this example, a counter-argument was presented by Cornfield et al. (1959). In order for Fisher's speculation to be valid, the genetic factor would have to be a confounder that is 10 times more prevalent in smokers than non-smokers and those with the confounder would have to have a 10 times greater risk of developing lung cancer. The existence of such a genetic confounder seemed unlikely.

As a formal method for discussing causality in observational studies, Rubin (1974) used the language of potential outcomes. This was introduced by Neyman (1923) in the context of randomized experiments. With binary treatment assignments, $Y(1)$ and $Y(0)$ are the potential outcomes under each level of treatment. The average treatment effect (ATE) is then defined as the expectation of the difference between those two variables, $\text{ATE} = E\left(Y\left(1\right) - Y\left(0\right)\right)$. Identification of this parameter can be shown using a set of assumptions. A typical set of identifiability assumptions includes (i) the result of one treatment level is observed for each individual, (ii) every individual has a positive probability of receiving each of the treatments and (iii) there are no unobserved confounders of the relationship between treatment and outcome. When identifiability

holds, the estimation becomes a pure statistical problem. The outcome regression (OR) estimator of the average treatment effect requires estimations of the potential outcome models $m_t(X) = E(Y(t)|X)$ and is based on the following estimating equation:

$$E\left(\Psi_{\text{OR}}\left(O, m_1(X), m_0(X)\right)\right) - \mathbf{ATE} = 0,$$

where $O = (X, Y, T)$, X is a set of covariates, Y is the observed outcome, T is the treatment and $\Psi_{\text{OR}} = m_1(X) - m_0(X)$. Typically, when drawing an inference from this estimator, it is assumed that parametric outcome models are correctly specified. Paper I, however, investigates the performance of this estimator in a high-dimensional situation.

There are other estimators of the average treatment effect that require estimation of the probability of receiving the treatment $p(X) = E(T|X)$ (See e.g. Kang and Schafer, 2007). One example is the double robust (DR) estimator, which is based on the following estimating equation:

$$E\left(\Psi_{\text{DR}}\left(O, m_1(X), m_0(X), p(X)\right)\right) - \mathbf{ATE} = 0,$$

where

$$\Psi_{\text{DR}} = \Psi_{\text{OR}} + \frac{T}{p(X)}\left(Y - m_1(X)\right) - \frac{1-T}{1-p(X)}\left(Y - m_0(X)\right).$$

This estimator is considered in this thesis because of the availability of $\sqrt{n}$-inference under weak conditions and for efficiency considerations. This is discussed in the following sections.

3. Efficiency and superefficiency

The well-known Cramer Rao bound provides an efficiency bound for an estimator, but only for unbiased estimators. The asymptotic variance bound for consistent estimators is nontrivial if we do not restrict ourselves to specific estimators. Dropping the unbiasedness assumption, we have the following example of the Hodge estimator (Le Cam, 1953), which can be considered an alternative to the mean of a sample of size n ($\bar{X}_n$) as an estimator of the expectation $\mu = E(X)$:

$$\hat{\mu}_n = \begin{cases} \bar{X}_n, & \text{if } \left|\bar{X}_n\right| \geq n^{-1/4} \\ 0, & \text{if } \left|\bar{X}_n\right| < n^{-1/4} \end{cases}.$$

The asymptotic variance is equal to zero for this estimator when the true parameter μ is zero, but it behaves erratically when μ is close to zero. In other words, this superefficiency is gained at the expense of poor estimation in a neighborhood. This is unsatisfactory, considering that the true value of the parameter is not known (Tsiatis, 2006, Section 3.1).

For the parametric case, when the parameters indexing the model are finite-dimensional, the convolution theorem (Hájek, 1970) suggests that the Cramer Rao bound holds if we limit ourselves to regular estimators (roughly, those whose convergence to their asymptotic distribution holds in a uniform sense and not only pointwise), under some other regularity conditions (Van der Vaart, 2000, Theorem 8.8).

Discussing efficiency is more complicated in the nonparametric (semiparametric) setting, where the model consists of an infinite dimensional nuisance parameter (and a low-dimensional nuisance parameter). However, even in such a case, by limiting ourselves to regular asymptotic linear estimators, we can characterize the variance bound. An asymptotic linear estimator $\hat{\theta}$ of a parameter θ is such that

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \Sigma_{i=1}^{n} \Phi(O_i, \eta) + o_P(1), \tag{1}$$

where Φ is a mean zero finite variance function of an observation O_i and a (possibly infinite-dimensional) nuisance parameter η and the last term converges to zero in probability. Influence functions of θ are Φ functions that satisfy the above presentation for an estimator of θ. The semiparametric efficiency bound is the $1/n$-scaled variance of the "efficient" influence functions – that is, the influence function with the smallest variance.

The efficient influence function of ATE is of the form $\Psi_{DR} - \mathbf{ATE}$. Under $E(Y(t)|X) = E(Y(t)|U)$, only $U \subset X$ is required for identification of the average treatment effect, i.e., it is not necessary to use the entire set X (De Luna et al., 2011). This information (called the exclusion restriction in the outcome relationship in Hahn (2004)) results in a smaller or equal variance bound for the parameter. In other words, omitting variables known to be instrumental variables can reduce the variance of our estimator. In the absence of such prior knowledge, attempting to achieve this lower variance results in superefficiency. A superefficient estimator is constructed by selecting out instrumental variables using a variable selection strategy or through using probability of

treatment conditional on conditional (on covariates) mean outcome instead of probability of treatment conditional on the covariates (Benkeser et al., 2020; Moosavi et al., 2021).

4. Low-dimensional setting

Here, we consider the case where the set of covariates X is low-dimensional and the nuisance models are indexed using low-dimensional parameters (parametric modelling). If m-estimators are used to estimate nuisance models in OR and DR, these estimators are asymptotic linear. Asymptotic distributions of these estimators can be found using a vector form of the Taylor expansion where the vector is found by stacking the estimation equation functions of the nuisance and main parameters (Stefanski and Boos, 2002; Vermeulen and Vansteelandt, 2015). While OR has a lower variance than DR, DR has the advantage of being consistent when only one nuisance model is specified correctly (double robust property).

It is possible to interpret double robustness differently when it comes to inference and asymptotic distribution. For DR, if both models are correctly specified, the first-order asymptotic is not affected by nuisance parameter estimation, meaning it is asymptotic linear with influence function $\Psi_{\text{DR}} - \mathbf{ATE}$. For OR, however, the first-order asymptotic is affected by the fact that the nuisance parameter is being estimated, so the influence function is $\Psi_{\text{OR}} - \mathbf{ATE} + \mathbf{extra\ term}$, where the **extra term** is a mean zero term.

5. High-dimensional and complex settings

High dimensionality can occur for two reasons. It may be because many covariates are considered during the analysis in order to ensure that there are no unobserved confounders. The second reason could be that we may need to consider increasing numbers of basis functions or higher-order terms in order to accomplish a consistent nonparametric estimation of the nuisance models. This may be more than the sample size, resulting in high dimensionality.

5.1. More covariates than the sample size

Data cannot be used to test the assumption that no unobserved confounders are present, so one may select a large number of pre-treatment covariates, sometimes greater than the sample size, to ensure the assumption is valid. A post-variable-selection method is sometimes used when only a low-dimensional set of variables among the initial set is thought to be relevant for outcome models. When the variable selection consistently finds the true subset, this estimator performs as well as an oracle estimate based on the true low-dimensional subset. Inference based on this property implies ignoring the finite sample bias and variation caused by variable selection. Although asymptotic results hold (at least pointwise asymptotic), post-selection estimators may simply be variations of Hodge's estimators, which means that the coverage of confidence intervals in finite samples may be quite low when the post-selection estimator is used (Leeb and Pötscher, 2005). As a way to mitigate this issue, it is possible to perform double selection, which means not just selecting variables that are highly associated with the outcome, but also selecting those variables that are highly associated with the treatment variable (Moosavi et al., 2021; Belloni et al., 2014).

5.2. Nonparametric estimation of nuisance models

The high dimensionality may be due to the inclusion of a large number of basis functions in the model. The task of performing $\sqrt{n}$-inference is challenging in this case and other cases where nuisance models are estimated nonparametrically, which requires estimating causal parameters at a parametric rate. It is possible to use the DR estimator for this purpose, which can be viewed as a debiased version of the OR estimator. A valid $\sqrt{n}$-inference for the DR estimate requires the nonparametric estimators of nuisance parameters to perform well enough, which is achieved by appropriate estimators under different regularity conditions such as sparsity or smoothness (Farrell, 2015; Chernozhukov et al., 2018; Farrell et al., 2021).

5.2.1 Approximate sparsity and Lasso

Whenever the number of parameters exceeds the sample size, it is impossible to fit regressions by minimizing the usual loss functions (least square/negative log-likelihood). In spite of this, the lasso fit, which

is conducted by adding an extra term $\lambda\Sigma_{i=1}^{p}|\beta_i|$ to the loss function (Tibshirani, 1996), with λ as a hyperparameter, makes it possible to solve the optimization problem in a unique way. A sparsity assumption must be made to derive desirable properties for this estimator. The exact sparsity assumption indicates that there are only a small number of parameters with non-zero values. Approximate sparsity, arguably more suitable for nonparametric estimation, requires that there is a low-dimensional number among p terms of covariates/basis functions (where p can be larger than sample size and grows in size with n) that can approximate regression functions relatively well (Belloni et al., 2014; Farrell, 2015; Moosavi et al., 2021).

5.2.2 Smoothness and neural network

There are several areas where neural networks are widely used, because they can be applied to analyze information as diverse as text, image, voice, and other kinds of data. While the first mathematical models were introduced in 1994, their usefulness was only realized later on, as more and more large data sets were collected and computers were able to handle heavier computations. A great deal of progress has been made in the field of applications since then, but the theory is still far behind. Nevertheless, there also has been a growing body of literature devoted to finding theoretical explanations for the success of these estimators. For example, the idea of universal approximation is concerned with estimating a continuous function based on neural network estimation. Moreover, a Sobolev smoothness condition is typically used to find the rate of convergence for this estimator.

Figure 1 illustrates a simple feed-forward network. Computation nodes (in black), which are nodes in the output (last) layer or hidden layers (layers between the input and output nodes), are calculated by using nodes from a previous layer:

$$f = \sigma(\Sigma_{i=1}^{3} w_i x_i + b).$$

In the first step, the incoming nodes (x_i) are multiplied by the weights (w_i), illustrated as connection arrows, and then a bias term (b) is added. A nonlinear function (σ) is then applied to the result. Among the most popular nonlinear functions are rectified linear unit, tanh, sigmoid, and softmax. Binary (multinomial) logistic regression is

equivalent to a network with one layer of a sigmoid (softmax) computation unit. There is, however, a need to have multiple hidden layers (deep networks) to ensure that the neural networks perform at their best – that is, they can produce more flexible functions and, consequently, achieve more accurate predictions through the use of multiple hidden layers. There have been several families of deep neural networks introduced for different applications. The convolutional neural network is a successful one that can efficiently discover local-time invariant features (Goodfellow et al., 2016).

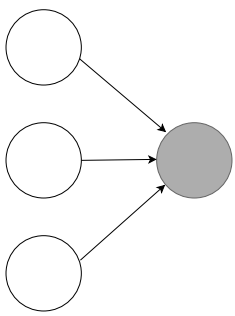

Figure 1. *An illustration of a computation unit in a neural network*

6. Sensitivity analysis

The assumption of no unobserved confounders cannot be empirically tested, yet the results may be highly sensitive to this assumption. It is therefore essential to have a mechanism in place to assess how sensitive results may be to unobserved confounders. Sensitivity parameters are used to quantify deviations from this assumption. According to some proposed methodologies, a given value of a sensitivity parameter can be used to provide a point estimation of the causal parameter. This can be used to assess how robust the results are to different values of the sensitivity parameter or to obtain interval estimation using a range of plausible sensitivity parameter values. An uncertainty interval is a method of conducting sensitivity analysis that takes into account the sampling variability in addition to the uncertainty around the value of the sensitivity parameter (Vansteelandt et al., 2006; Genbäck and de Luna, 2019).

An early example of sensitivity analysis is Cornfield et al. (1959),

which explores conditions for an unobserved confounder to explain away the observed smoking-lung cancer relationship (see Section 2). They do not, however, account for any observed confounders. Another example of sensitivity analysis is presented in Rosenbaum et al. (2010). The odds ratio for receiving treatment (conditional on observed covariates) parameterizes the unobserved confounding (an odds ratio different than 1 means the assumption is violated). If a bound is given on the odds ratio, an interval of p-values is provided instead of a single p-value for the causal parameter. Our approach however, is based on Genbäck and de Luna (2019), where the sensitivity parameter is a conditional correlation between outcome and treatment (ρ). For a given value of the correlation parameter, an estimate of the confounding bias can be found ($\hat{b}(\rho)$). The bias estimate is then subtracted from the lower and upper bounds of the naive (based on the unconfoundedness assumption) 95% confidence interval. A sensitivity analysis based on a range of plausible ρ values ($\mathcal{P}$) is performed by taking the union of corrected confidence intervals (called the uncertainty interval), as below

$$\mathbf{UI} = \bigcup_{\rho \in \mathcal{P}} \{\text{Naiv}_{\text{lower}} - \hat{b}(\rho); \text{Naiv}_{\text{upper}} - \hat{b}(\rho)\}.$$

Other sensitivity analysis approaches have been proposed; see, e.g., Scharfstein et al. (2021) for a review.

7. Summary of papers

7.1. Paper I

Our first study aims to investigate the problem of conducting causal inference in the scenario where the corresponding covariates are chosen using a data-adaptive method, for example, a lasso regression method. There are different purposes for variable selection. These methods can help identify variables that do not contribute to bias but amplify variance within analyses, so that redundant variables can be omitted. Another reason is to select important terms from a potentially high-dimensional set of basis functions.

Naive strategies based on preliminary model selection for nuisance models have finite sample distributions that are poorly approximated by their asymptotic distributions. Recent literature has proposed estimators that converge uniformly over a class of data-generating processes in

order to solve this problem. An overview of the literature on uniformly valid causal inference is presented in this paper. However, there may be a price to pay for uniformly valid procedures in terms of inflated variability. In that regard, we present a double-selection outcome regression estimator that is uniformly asymptotically unbiased. Sparsity conditions are considered for the nuisance models. A double robustness property applies, allowing one nuisance model to be more sparse while the other is less sparse. Unlike the naive single-selection outcome regression estimator, this estimator does not have a large bias, which is seen by comparing the coverage of their confidence intervals.

7.2. Paper II

This paper investigates a sensitivity analysis using double robust estimators of the average causal effect. Under unconfoundedness, it is possible to obtain good confidence intervals using the double robust estimator, even when the variables to be included in the nuisance models have been chosen through a variable selection process. This holds even when the variable selection has not been taken into account when constructing the confidence intervals. The reason is that this estimator has uniform convergence over a class of data-generating processes; our nuisance estimators, such as maximum likelihood refit using an estimated set of covariates, are well-suited to this situation.

To address the possible violation of the assumption of no unobserved confounders, this paper employs an expression of confounding bias of a double robust estimator that is valid under weak modelling assumptions. In this model, the sensitivity parameter is the correlation between error terms of nuisance models. An estimate of the confounding bias can be subtracted from the double robust estimator. In this case, valid inference can be obtained if the sensitivity parameter is small enough that the finite sample bias and the variance of the estimation of confounding bias are negligible. Additionally, an existing formula is used to improve the estimation of the error variance of the outcome model as part of the confounding bias estimation. Originally proposed for a parametric setting, the formula performs well when a doubly selected set of variables is considered.

7.3. Paper III

This paper presents a bound on the approximation error associated with a class of functions generated by a multichannel convolutional neural network structure. Basically, approximation error measures the distance between a goal function and the best candidate in a function space. It is assumed that the goal function belongs to a Sobolev space. It is possible to use the bound on approximation error to calculate a bound for the estimation error, which, instead of the best candidate, corresponds to a fitted function that minimizes loss across a data set. Following this, it is shown that a well-chosen number of channels and layers for the convolutional neural network architecture (more precisely, a specific growth rate) can yield the rate of convergence $o_P(n^{-1/4})$ for both linear and logistic models. In this case, we can construct valid confidence intervals for causal parameters of interest using a double robust estimator. We then use this methodology to estimate the causal effect of early retirement on health outcomes for those who retire early, using Swedish register data.

7.4. Paper IV

When the outcome of a linear regression is missing not at random, the variance of its error term cannot be consistently estimated using the mean squared residual of the ordinary least square fit. Previously, a bias correction has been proposed for this estimator assuming linear models. This paper presents a new correction formula that does not require linearity and shows that it yields a consistent estimate using a nonparametric consistent regression fit. This new correction is then used for performing sensitivity analysis as described in paper II. A R package is provided online for this purpose, and a simulation setting from paper II is re-examined. Finally, the code is used to study the effect of smoking during pregnancy on child birth weight.

8. Further research

Many causal effect estimation procedures are based on strong parametric models. A causal analysis under weaker modelling assumptions has been explored in this thesis. As theories suggest, the efficient influence functions can be used for obtaining $\sqrt{n}$ estimators under weak

modelling assumptions. Further research could investigate the possibility of applying similar theories to new estimands, such as direct and indirect effects of treatment. Another possibility is to study whether the variance estimator in paper IV may allow weaker conditions (e.g., on the sensitivity parameter) for the sensitivity analysis in paper II. Finally, we have plans to merge *hdm.ui* and *ui* into a single package.

References

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Benkeser, D., W. Cai, and M. J. van der Laan (2020). A nonparametric super-efficient estimator of the average treatment effect. *Statistical Science 35*(3), 484–495.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Cornfield, J., W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute 22*(1), 173–203.

De Luna, X., I. Waernbaum, and T. S. Richardson (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika 98*(4), 861–875.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Fisher, R. A. (1958). Cancer and smoking. *Nature 182*(4635), 596–596.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pp. 66–70. Springer.

Genbäck, M. and X. de Luna (2019). Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics 75*(2), 506–515.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT Press.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics 86*(1), 73–76.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 14*(4), 323–330.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. in Statist. 1*, 277–330.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Moosavi, N., J. Häggström, and X. de Luna (2021). The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. *To appear in Statistical Science. ArXiv preprint arXiv:2105.02071*.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (translated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences 10*, 1–51.

Rosenbaum, P. R., P. Rosenbaum, and Briskman (2010). *Design of observational studies*, Volume 10. Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Scharfstein, D. O., R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid (2021). Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*.

Stefanski, L. A. and D. D. Boos (2002). The calculus of m-estimation. *The American Statistician 56*(1), 29–38.

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, 73–79.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Vansteelandt, S., E. Goetghebeur, M. G. Kenward, and G. Molenberghs (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 953–979.

Vermeulen, K. and S. Vansteelandt (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association 110*(511), 1024–1036.