ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Recent Approaches to Instrumental Variable Estimation of Nonlinear Causal Effects

Master Thesis

Samuel Joray

April 15,2024

Advisors: Prof. Dr. N. Meinshausen, A. Holovchak
Seminar for Statistics, ETH Zürich

Abstract

Causal effect estimation requires an understanding of the underlying problem structure. When confronted with an unobserved confounder, instrumental variable (IV) regression techniques accurately predict the true causal effect. It is based on a variable, called IV, that impacts the treatment without affecting the confounder. This thesis delves into IV methods for estimating nonlinear causal effects. We extensively explore the two primary branches of IV methods: two-stage least squares and control functions. Through theoretical analysis and simulation studies, we comprehensively compare six recent algorithms: 2SLS, CF, DeepIV, DeepGMM, GCFN, and DeepCF. Including one that we propose, named DeepCF, which leverages principles from the CF algorithm and employs deep learning techniques to improve its generalisability. From the theory of GCFN, we show that DeepCF learns the true causal effect. Additionally, we establish connections between various estimators, including those between 2SLS and CF, 2SLS and DeepIV, and DeepIV and DeepGMM. Through simulations, we provide a comprehensive assessment of the performance of each method, highlighting their respective strengths and weaknesses.

Contents

Notation

$\mathbb{N} = \{0, 1, 2, \dots\}$	
$\mathbb{R}$	denotes the real numbers.
$plim(\hat{\theta}_n) = \theta$	denotes the convergence in probability of $\hat{\theta}_n$ to θ as $n \to \infty$.
X	denotes a random variable.
x	denotes the realisation of the random variable X.
$X \perp\!\!\!\perp Y$	X is independent of Y.
$X \not\perp\!\!\!\perp Y$	X depends on Y.
$X \perp\!\!\!\perp Y \mid Z$	X is independent of Y conditionally on Z.
$p(x)$	denotes the density of X when $X = x$.
$p(x, y)$	denotes the joint density of X and Y when $X = x$ and $Y = y$.
$p(x \mid y)$	denotes the conditional density of X given Y when $X = x$ and $Y = y$.
$\|\|x\|\|$	denotes the 2-norm of a vector x.
$\mathcal{U}[a, b]$	denotes the uniform distribution on $[a, b]$.
$\mathcal{N}(\mu, \sigma^2)$	denotes the normal distribution with mean μ and variance σ^2.

Chapter 1

Introduction

1.1 Motivation

Assuming that an airline company is interested in the impact of airline ticket prices on demand[1], we observe that prices are high when demand is high and low when demand is low. Does that mean that airlines should increase prices to increase demand?

Certainly not! The reason why the positive correlation between prices and demand is not a causation is that other variables, such as holidays or conferences, affect both prices and demand. Therefore, if prices are high because of holidays, the demand will also be high due to holidays, but not due to prices! We are in the presence of a confounder: holidays and conferences. If we do not know when the holidays or conferences are, then we are in the so-called unobserved confounder problem.

Can we still estimate the effect of price on demand if we do not observe these confounders? No, unless we have access to other variables such as the price of the fuels. Fuel prices influence the ticket price, but do not influence when the holidays or conferences are. This variable is called an instrumental variable. Using instrumental variables to solve the unobserved confounder problem is what this thesis is all about!

1.2 Broader Context

Causal inference is the process of determining whether and to what extent a cause-and-effect relationship exists between two variables. Not like traditional machine learning, it provides answers to problems that data-driven algorithms cannot address by making specific assumptions about the problem, which must correspond to reality and are not provided by the data.

[1]This example also is used in Hartford et al. (2017) as an introductory example.

Usually, we have a treatment, which can be a drug, a policy or prices, and an outcome, such as the health of the patient, impact on the population or demand, respectively. A problem often encountered in practice is when an unobserved variable influences both the treatment and the outcome. These variables might be, for instance, age, socioeconomic factor, or holidays. They are called confounders.

A technique commonly used to mitigate the unobserved confounder problem is Random Control Trials (RCTs), which consist of randomly separating individuals in two groups. The first group consists of the individuals who take the treatment and the second group consists of those who do not. Then we compare the outcome of the two groups and obtain an unbiased causal effect. When these two groups are formed, the effect of unobserved confounders is mitigated since the confounders would influence both groups the same. In practice, it might be unethical, too costly, or infeasible to implement RCTs. Furthermore, statisticians often deal with observed data that are not necessarily part of an experiment.

Instrumental Variables (IVs) can be used to mitigate the unobserved variables problem when we do not have data from an RCT. IVs have a long history in econometrics, epidemiology, medecine, climatology, etc. They can be traced back to Wright (1928), where IVs have been described for the first time. The book is called *The Tariff on Animal and Vegetable Oils* and the breakthrough of IVs appears only in Appendix B, where he describes how IV regression can be used to estimate the coefficient of an endogenous variable. The literature first focused on problems where the relation between variables is linear. Statisticians and econometricians then tried to relax every assumption as much as possible, including nonlinear relations between the variables. Recent developments in machine learning and deep learning have contributed significantly to the field of instrumental variables. A comprehensive and comparative study detailing much of the recent development of instrumental variables can be found in Wu et al. (2022).

The idea behind IVs is to use a variable, an IV, that is independent of the unobserved confounder, but that still influences the treatment. It is used to separate the treatment in two parts, one that is independent of the confounder and one that is dependent. Then, we take the independent part to learn the causal effect with the outcome and since we removed the unobserved confounder, the estimation is now unbiased. This method is known as the Two-Stage Least Squares (2SLS) algorithm.

In practice, it can be challenging to determine whether the instrumental variable is truly independent of the unobserved confounder, as we do not observe it and therefore cannot test the validity of this assumption. Thus, we must argue, with knowledge of the specific problem, that it is indeed independent.

The goal of this thesis is to define and describe the main algorithm used in instrumental variables estimation. We give a strong emphasis on the assumptions and we prove some result on the consistency of the methods.

1.3 Outline of the Thesis

In this thesis, we start in Chapter 2 recalling important definitions and building the basis for causal inference that will be used throughout the thesis.

In Chapter 3, we describe the linear case and the main algorithm used for IV regression, namely the 2SLS algorithm. In the same chapter, we describe a generalisation of this algorithm for the case where the relation between the outcome and treatment is given by a linear combination of known nonlinear functions.

In Chapter 4, we describe the method of Control Function (CF). This method is based on other assumptions which make it powerful in the nonlinear case. We then apply splines to the control function and two-stage least squares algorithms to estimate smooth functions. We then further generalise the concept of control functions with an identification theorem for a general control function.

In Chapter 5, we describe recent approaches that build on deep learning to estimate complex causal effects. A first approach is called DeepIV which is a generalisation of 2SLS using deep learning. This method was one of the first to use deep learning. In this thesis, we propose an algorithm that generalises the CF algorithm described in Chapter 4. We call it DeepCF and we prove that the algorithm gives identification of the true causal effect. We then describe DeepGMM, which is based on the method of moment and has been modified to use deep neural networks. We show that this method uses the same kinds of assumption as in DeepIV. Finally, we describe the general control function algorithm, which creates a control function through an autoencoder, which is then used to control the confounder given the treatment to find an unbiased causal effect.

Last but not least, in Chapter 6, we test all algorithms described in this thesis. We compare them under different settings, compare the convergence, and consider higher-dimensional settings.

1.4 Further Real-Life Examples

In this section, we describe some real-life examples in which instrumental variables have been used to correct for unobserved confounder. The examples described here are mostly from econometrics.

Example 1.1. *Assume we want to find out how the level of study of individuals influences the earnings. If we observed for some people their level of education and their earnings, we would have a wrong idea of the relation between the two variables because the IQ (or any measure of intelligence) of a person is not taken into account, and it may be that more people with a higher IQ do university and earn more money later on, which does not tell whether the university actually makes people earn more or if people earn more because they have more skills in general. In this example, the unobserved confounder is the IQ of an individual.*

An instrumental variable could be the distance between a university and the place where the person lives. It is supposed to influence the decision to go to university. A smaller distance between home and a university reduces the cost of going to the university compared to people who live far away from a university. Furthermore, if it does not influence the skills of a person, then the distance to a university could be an instrumental variable. A study of this problem can be found in Card (1999).

Example 1.2. *Can higher cigarette taxes improve birth outcomes?*

This is the question that has been asked in the study conducted by Evans and Ringel (1999). Researchers tried to assess the causal effect of maternal smoking on birth weight. This is not an easy question since some health problems may impact both maternal smoking and birth weight, therefore, if it is unobserved, it is an unobserved confounder. To control for this confounding effect, the researchers used cigarette taxes that varied across states. This variation has been used to extract the variation in maternal smoking that is uncorrelated with unobserved health problems.

The study was the first to show that, indeed, higher cigarettes taxes have a beneficial impact on infant birth weight. Moreover, the estimate they found for the impact of maternal smoking on birth weight is remarkably close to estimates from a random assignment clinical trial, suggesting that the unobserved confounder has indeed been controlled for and the true causal effect has been estimated.

Another study, in which reverse causality has been investigated and taken into account, is due to Husnain et al. (2018), in which the impact of climate change on agriculture has been investigated.

Example 1.3. *(Assessment of the impact of climate change on agriculture)*

In a study due to Husnain et al. (2018), instrumental variables have been used to assess the impact of climate change on agriculture. For this problem, climate change can be characterised with two variables, temperatures and precipitations, and agriculture is measured by the value-added agriculture, which is a measure of net agricultural output. They took into account many variables that influence agricultural yields, such as fertilisers, population, and agricultural land area. Previous studies on the subject did not take into account the impact of agriculture on climate change, which is a reverse causality problem, that is, climate change impacts agriculture, but also

agriculture impacts climate change. In this study, the researchers took this fact into account by adding two instrumental variables. The longitudes and latitudes, which influence the climate and agriculture only through the climate.

To learn the true causal effect of climate change on agriculture, instrumental variables are a great technique. If we were to try to get the correlation between climate change and agriculture, the estimate would not be the causal effect because it cannot distinguish between the changing climate on agriculture and the agriculture on climate change. As before, we can use the 2SLS algorithm together with the IV to obtain the causal effect of climate change on agriculture.

Chapter 2

Preliminaries

In this chapter, we recall important definitions that we will use throughout this thesis.

2.1 Linear Algebra

Definition 2.1. *For an $m \times n$ matrix A, the **rank** of A is defined to be the dimension of the row space (or the column space), denoted by rank(A).*

Proposition 2.2. *Let A be a $m \times n$ matrix and B be a $n \times k$ matrix with $rank(B) = n$. Then,*

$$rank(AB) = rank(A). \tag{2.1}$$

Definition 2.3. *For a vector $x = (x_1, \ldots, x_n)$, the 2-norm is defined as:*

$$||x|| := \sqrt{\sum_{i=0}^{n} x_i^2}. \tag{2.2}$$

2.2 Probability Theory

The content of this section is based on the book by Peters et al. (2017).

Unless specified differently, we consider a continuous real-valued random variable X that has a finite first and second moment, that is, $-\infty < \mathbb{E}[X] < \infty$ and $\mathbb{E}[X^2] < \infty$, and we assume the existence of a probability density function (pdf) that we write $x \mapsto p(x)$. We denote by $p(x, y)$ the joint pdf of two random variables X and Y.

Definition 2.4. *We call X **independent** of Y if and only if:*

$$p(x, y) = p(x)p(y) \tag{2.3}$$

*for all $x, y \in \mathbb{R}$. Otherwise, X and Y are **dependent**.*

When considering more than two random variables, different notions of independence can be stated. We consider the following.

Definition 2.5. *We call $X_1, \ldots, X_d$ **mututally** independent, if the joint pdf exists, and*

$$p(x_1, \ldots, x_d) = p(x_1) \cdots p(x_d) \tag{2.4}$$

for all real values $x_1, \ldots, x_d$.

Definition 2.6. *We call X_1 **jointly** independent of (X_2, X_3), denoted $X_1 \perp\!\!\!\perp (X_2, X_3)$, if:*

$$p(x_1, x_2, x_3) = p(x_1)p(x_2, x_3) \tag{2.5}$$

for all real values $x_1, \ldots, x_d$.

Definition 2.7. *We call $X_1, \ldots, X_d$ **pairwise independent** if and only if all pairs of random variables are independent, that is, $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$.*

Definition 2.8. *We call X and Y **conditionally independent** on Z, denoted by $X \perp\!\!\!\perp Y|Z$, which should be read as $(X \perp\!\!\!\perp Y)|Z$, if a joint pdf $p(x, y, z)$ exist and*

$$\begin{aligned} p(x|y, z) &= p(x|z), \quad \text{or} \\ p(x, y|z) &= p(x|z)p(y|z) \end{aligned} \tag{2.6}$$

for any real number x, y, z such that $p(z) > 0$.

Proposition 2.9. *We have the following relations between the notions of independence:*

- *Mutual independence implies pairwise independence. However, the converse is generally not true.*
- *If X_1 is jointly independent of (X_2, X_3), then X_1 is independent of X_2 and X_3, that is, $X_1 \perp\!\!\!\perp (X_2, X_3) \Rightarrow X_1 \perp\!\!\!\perp X_2$ and $X_1 \perp\!\!\!\perp X_2$.*

Definition 2.10. *When $\mathbb{E}[X^2] < \infty$, the **variance** of a random variable X is defined as*

$$var(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{2.7}$$

Definition 2.11. *We call X and Y **uncorrelated** if*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y], \tag{2.8}$$

that is

$$\rho_{X,Y} := \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{var(X)var(Y)}} = 0. \tag{2.9}$$

*Otherwise, that is, if $\rho_{X,Y} \neq 0$, X and Y are **correlated**. $\rho_{X,Y}$ is called the **correlation coefficient** between X and Y.*

Proposition 2.12. *If X and Y are independent, then they are also uncorrelated:*

$$X \perp\!\!\!\perp Y \Rightarrow \rho_{X,Y} = 0. \tag{2.10}$$

The converse is generally not true. However, there are examples where uncorrelated implies independent, such as in the case of Gaussian bivariate distributions or binary random variables.

Definition 2.13. *A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ converges to X in probability, denoted $plim(X_n) = X$, if:*

$$\forall \varepsilon > 0 : \mathbb{P}(|X_n - X| > \varepsilon) = 0. \tag{2.11}$$

Definition 2.14. *An estimator $\{\theta_n\}$ is consistent with θ, if:*

$$plim(\theta_n) = \theta. \tag{2.12}$$

Proposition 2.15.

Dependence between two continuous random variables X and Y can be measured by the Shannon mutual information, which is equal to the Kullback-Leibler divergence, denoted by $\mathbf{KL}(P||Q)$ if the random variables X and Y have, respectively, cumulative distribution functions (CDF) P and Q and pdf p and q.

Definition 2.16. *The **Shannon mutual information** between two random variables X and Y, denoted by $I(X,Y)$, is defined by:*

$$I(X,Y) := \int p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) dxdy. \tag{2.13}$$

2.3 Causal Inference

The foundations of modern causal inference have been given by Judea Pearl in Pearl (2000). A common tool used to describe these relations is a graph. In the following, we construct the basis for causal inference. We recall some of the most important definitions. Most of the content in this section can be found in the book by Peters et al. (2017).

Definition 2.17. *Consider finitely many random variables $X = (X_1, \ldots, X_d)$ with index set $\mathcal{V} := \{1, \ldots, d\}$, joint distribution P_X, and density $p(x)$. A **graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of nodes or vertices $\mathcal{V}$ and edges $\mathcal{E} \subseteq \mathcal{V}^2$ with $(v,v) \notin \mathcal{E}$ for all $v \in \mathcal{V}$. A node i is called a parent of j if $(i,j) \in \mathcal{E}$ and $(j,i) \notin \mathcal{E}$. The set of parents of j is denoted by $PA_j^{\mathcal{G}}$. Furthermore, there is a directed edge between two nodes i and j, if $(i,j) \in \mathcal{E}$ or $(j,i) \in \mathcal{E}$ but not both, and we denote it by $i \to j$ or $j \to i$. We call the edge undirected if $(i,j) \in \mathcal{E}$ and $(j,i) \in \mathcal{E}$.*

Definition 2.18. *A **path** in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, \ldots, i_m$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \ldots, m-1$. If $i_k \to i_{k+1}$ for all k, we speak of a **directed path** from i_1 to i_m. Moreover, we call i_j a **descendant** of i_k, if there is a directed path from i_j to i_k. We denote all descendants of i by $DE_i^{\mathcal{G}}$*

Definition 2.19. *A graph $\mathcal{G}$ is called a **Partially Directed Acyclic Graph (PDAG)** if there is no cycle, that is, if there is no pair (j, k) with a directed path from j to k and from k to j. The graph $\mathcal{G}$ is called **Directed Acyclic Graph (DAG)** if, additionally, all edges are directed.*

DAGs[1] are heavily used in causal inference to model the causes and effects. In a DAG, the nodes represent the variables and the edges represent the causal effect. For example, if there is an arrow from node j to k, then we say that j causes k. In a PDAG, if there is an undirected edge from j to k, then this represents an association between the two variables.

Definition 2.20. *In a DAG $\mathcal{G}$, a path between nodes i_1 and i_m is **blocked** by a non-empty set S (with neither i_1 nor i_m in S) whenever there is a node i_k, such that one of the following two possibilities holds:*

1. *$i_k \in S$ and*

$$\begin{aligned} &i_{k-1} \to i_k \to i_{k+1}, \\ &\text{or } i_{k-1} \leftarrow i_k \leftarrow i_{k+1}, \\ &\text{or } i_{k-1} \leftarrow i_k \to i_{k+1}. \end{aligned}$$

2. *Neither i_k nor any of its descendants is in S, i.e., $(i_k \cup DE_{i_k}) \cap S = \varnothing$, and*

$$i_{k-1} \to i_k \leftarrow i_{k+1}.$$

*Furthermore, in a DAG $\mathcal{G}$, we say that two disjoint subsets of vertices A and B are **d-separated** by a third (also disjoint) subset S if every path between nodes in A and B is blocked by S. We then write*

$$A \perp\!\!\!\perp_{\mathcal{G}} B | S. \tag{2.14}$$

Example 2.21. *The left graph of figure 2.1 has the following properties:*

- *X_1 and X_4 are d-separated by $S = \{X_2\}$ or $\{X_2, X_3\}$.*
- *The path $X_2 \to X_3$ is unblocked since X_2 is directly connected with X_3.*
- *The path $X_2 \to X_4 \to X_3$ is blocked by $\{X_1\}$ but unblocked by $\{X_4\}$ because of point 2 in the definition of d-separation 2.20.*

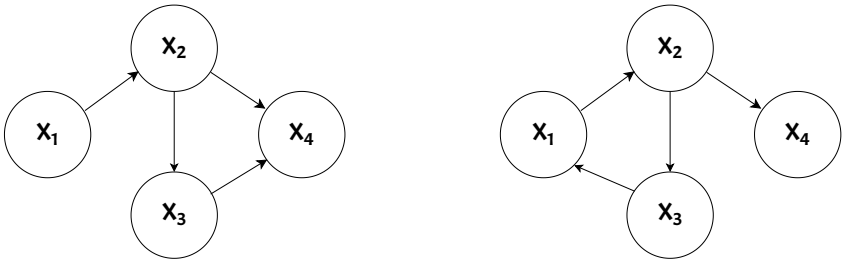


Figure 2.1: Two graphs, the one on the left is a DAG. The graph on the right is not a DAG since it contains a cycle $X_1 \to X_2 \to X_3 \to X_1$.

Next, we see a way to model the causal graph with structural equations. These are a set of assignments that captures the relations between the variables and corresponds to a DAG.

Definition 2.22. *A **Structural Causal Model (SCM)** $\mathcal{C} := (S, P_N)$ consists of a collection S of d assignments*

$$X_j := f_j(PA_j, N_j), \quad j = 1, \dots, d, \tag{2.15}$$

where $PA_j \subseteq \{X_1, \dots, X_d\} \setminus X_j$ are called parents of X_j and $P_N = P_{N_1,\dots,N_d}$ is a joint independent distribution over the noise variables. The distribution $P_X^{\mathcal{G}}$ is called the observational distribution.

Example 2.23. *An SCM that can be associated with the left graph of figure 2.1 is:*

$$\begin{aligned} X_1 &:= f_1(N_1), \\ X_2 &:= f_2(X_1, N_2), \\ X_3 &:= f_2(X_2, N_3), \\ X_4 &:= f_2(X_2, X_3, N_4), \end{aligned} \tag{2.16}$$

where the N_i's are jointly independent.

To separate the effect of a specific variable on the others and know how it affects the system, we can use the concept of intervention.

Definition 2.24. *Consider an SCM $\mathcal{C} := (S, P_{,N})$ and its entailed distribution $P_X^{\mathcal{C}}$. We replace one (or more) of the structural assignments to obtain a new SCM $\tilde{\mathcal{C}}$. Assume that we replace the assignment for X_k by*

$$X_k := \tilde{f}(\widetilde{PA}_k, \tilde{N}_k).$$

[1] In this work, the figures of causal graphs have been made with the website visual-paradigm https://online.visual-paradigm.com/.

We then call the distribution of the new SCM an ***Intervention distribution*** *and denote it by*

$$P_X^{\tilde{\mathcal{C}}} =: P_X^{\mathcal{C};do(X_k := \tilde{f}(\widetilde{PA}_k, \tilde{N}_k))}. \tag{2.17}$$

The causal effect of Y on X can be defined with the help of interventions. In words, there is a causal effect if there is still a dependence between the two variables after an arbitrary intervention on X.

Definition 2.25. *Given an SCM $\mathcal{C}$, there is a* ***total causal effect*** *from X to Y if and only if*

$$X \not\perp\!\!\!\perp Y \text{ in } P_X^{\mathcal{C};do(X:=\tilde{N}_X)}$$

for some random variable $\tilde{N}_X$.

Before defining counterfactuals, we start with an example.

Counterfactuals deal with the following. Assume that you are playing the roulette wheel at the casino and you put all your money on red. This, however, falls on a black number, then you might say "If I had played on black, I would have doubled my money", which is a counterfactual statement given one observation. However, this does not restrict to one observation. If one can observe the game long enough, we could observe that there are equal chances to win as playing red or black, both $\frac{18}{37}$ because the 0 is green, which is the counterfactual distribution given all the observations and the counterfactual statement could be: "If I had not played, I would have saved in average $\frac{1}{37}$ times the money I have played".

Definition 2.26. *Consider an SCM $\mathcal{C} := (S, P_N)$ over nodes $\mathbf{X}$. Given some observations $\mathbf{x}$, we define a* ***counterfactual*** *SCM by replacing the distribution of noise variables:*

$$\mathcal{C}_{\mathbf{X}=\mathbf{x}} := (S, P_N^{\mathcal{C}|\mathbf{X}=\mathbf{x}}),$$

where $P_N^{\mathcal{C}|\mathbf{X}=\mathbf{x}} := P_{N|\mathbf{X}=\mathbf{x}}^{\mathcal{C}}$. The new set of noise variables needs not be jointly independent anymore. Counterfactual statements can now be seen as do-statements in the new counterfactual SCM.

The Markov property is an important assumption that builds a certain relation between a graph and the distribution associated with it. If the Markov property holds, two variables in a graph that are d-separated are also independent.

Definition 2.27. *A DAG $\mathcal{G}$ satisfies the* ***global Markov property*** *if*

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C} \tag{2.18}$$

for all disjoint node sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$.

Following Example 2.21 in figure 2.1 and assuming the Markov property, we have the following conditional independence:

- $X_1 \perp\!\!\!\perp X_4 | X_2$,
- $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$,
- $X_1 \perp\!\!\!\perp X_3 | X_2$,
- $X_1 \perp\!\!\!\perp X_3 | X_2, X_4$.

Definition 2.28. *In a regression setting, an **endogenous** random variables is a variable that depends on the error, while an **exogenous** random variables is one that is independent of the error term.*

Having endogenous variables as predictors often results in inconsistency in the prediction.

Chapter 3

Two-Stage Least Squares

In this chapter, we describe precisely the most famous algorithm, two-stage least squares, for using IVs to control for unobserved confounders, generalise it to nonlinear functions, and prove its consistency.

3.1 General Setup

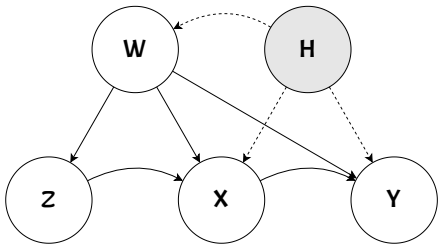


Figure 3.1: DAG of the Instrumental Variables setting

In this thesis, we assume that we already know the underlying structure of the problem, that is, which variables cause what. Here, we are interested to know how much a variable causes another one, which is to quantify and estimate this effect. The problem we are interested in can be summarised with a DAG as in Figure 3.1 and an SCM:

$$\begin{aligned} W &= h_0(H, N_w), \\ Z &= h_1(W, N_z), \\ X &= g(Z, W) + h_2(H, N_x), \\ Y &= f(X, W) + h_3(H, N_y). \end{aligned} \tag{3.1}$$

Here, N_w, N_z, N_x and N_x are independent centred noise. We say that X is the treatment variable, Y is the outcome variable, and W is an observed

confounder that influences both X and Y and can also influence Z. We assume that the noises and the hidden confounders, $h_1(H, N_X)$ and $h_2(H, N_Y)$, are additive in the treatment and outcome. We also assume that $\mathbb{E}[h_2(H, N_X)] = \mathbb{E}[h_3(H, N_Y)] = 0$ and that the mean of W and Z is 0. Later, we will make assumptions on the functions g, f, h_0, h_1, h_2 and h_3 to be able to estimate f well.

This formulation of the setting arises in papers such as Hartford et al. (2017), Wu et al. (2022) and Puli and Ranganath (2021).

To first explain and get an intuition on the assumptions we need, we follow Hartford et al. (2017), Z is called an IV if it satisfies the following assumptions:

Assumption 1. *(Relevance) The treatment variable X depends on Z given W, that is, the distribution of X given W and Z, $p(x|w, z)$, is not constant in z.*

Assumption 2. *(Unconfounded Instrument) Z is independent of the hidden confounders H given W, that is, $Z \perp\!\!\!\perp H|W$.*

Assumption 3. *(Exclusion) Z depends on Y only through X or W, that is, $Z \perp\!\!\!\perp Y|\{X, W\}$.*

Remark 3.1. *We note the following on the assumptions 1-3.*

- *In Assumption 1, if X only slightly depends on Z, we call Z a weak instrument. In practice, it can lead to a biased estimate, as shown in Andrews et al. (2018) and Bound et al. (1995).*
- *Since H is unobservable, we cannot statistically test if assumptions 2 and 3 are satisfied. Therefore, in practice, it is hard to find a valid instrumental variable, and the validity of the instrumental variable must be well justified by arguments from previous studies on the problem.*
- *Since d-separation implies conditional independence by the Markov property, the second assumption does not allow any path between Z and H unless it is blocked by W, in short Z and H are d-separated by W. In the same way, the third assumption does not allow any path between Z and Y, unless blocked by the set $\{X, W\}$. In short Z and Y are d-separated by $\{X, W\}$.*
- *In this work, we mostly consider 1-dimensional variables. However, it can be generalised to higher dimensions. See, for example, Guo and Small (2016) for a generalisation of the case where Z and W are vectors for the additive nonlinear estimator. In Hartford et al. (2017) and Bennett et al. (2020), a generalisation with higher dimensions of X and Z has been tested.*
  In Chapter 6, we consider simulations in which the treatment X and instruments Z have higher dimensions.

Instrumental variables can be used in many different DAGs. We can also adapt assumptions 1, 2, and 3 if we have different variables. However, if

we assume that we only have variables X, Y, Z and W, we can also allow different links between the variables as long as Z is still a valid instrumental variable and the graph is still a DAG.

In the graph on the left of Figure 3.2, H and Z are d-separated by W. Therefore, according to the Markov property 2.27, we have the Assumption 2, that is, $Z \perp\!\!\!\perp H|W$. In the same way, Assumption 3 holds. However, in the graph on the right, the path $Z \to W \leftarrow H$ is not blocked by W. Thus, Assumption 2 does not hold.

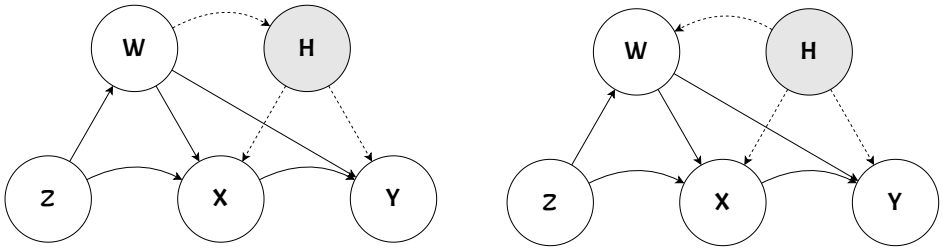


Figure 3.2: Two DAGs, the one on the left has a valid IV where as the one on the right has an invalid IV

We are interested in the causal effect from X to Y and we want to make a counterfactual assertion such as: "If the price X was different, what would have been the demand Y for this flight". This can be answered if we can estimate the function h:

$$h(x, w) = f(x, w) + \mathbb{E}[h_2(H)|W = w]. \tag{3.2}$$

Intuitively, it means taking the conditional expectation of Y given W holding the same distribution of H as the treatment X changes. The function h gives indeed the correct counterfactual prediction since intervening on X gives:

$$\begin{aligned} h(x_1, w_1) &:= \mathbb{E}[Y|do(X = x_1), W = w_1] \\ &= \mathbb{E}[f(X, W) + h_2(H)|do(X = x_1), W = w_1] \\ &= f(x_1, w_1) + \mathbb{E}[h_2(H)|do(X = x_1), W = w_1] \\ &= f(x_1, w_1) + \mathbb{E}[h_2(H)|W = w_1], \end{aligned}$$

where we used the definition of Y in the first equality and in the third equality, we use the fact that intervening on X does not change the value of H.

3.2 Linear Effects

We first consider the setting where the relations between the variables are linear, that is, when all functions g, f, h_0, h_1, h_2 and h_3 are linear. Furthermore,

to understand well what is happening, we do not take W into account. We have the SCM:

$$\begin{aligned} Z &= N_z, \\ X &= \alpha Z + \gamma_2 H + N_x, \\ Y &= \beta X + \gamma_3 H + N_y. \end{aligned} \tag{3.3}$$

In this setting, we assume the two following assumptions, which are similar to Assumption 1 and Assumption 2, respectively.

The following assumption is stronger than Assumption 1.

Assumption 1a. *Z is correlated with X, i.e. $cov(Z, X) \neq 0$.*

The following assumption is weaker than Assumption 2.

Assumption 2a. *Z is uncorrelated with H, that is, $cov(Z, H) = 0$.*

The idea of using covariance instead of independence is that independence implies that covariance is 0 but not the opposite. Indeed, consider any symmetric random variable, such as $X \sim \mathcal{N}(0, 1)$, then $cov(X, X^2) = 0$, but X and X^2 are not independent.

In this case, the counterfactual function h can be calculated directly. We use the definition of h in equation 3.2.

$$h(x) = \beta x + \mathbb{E}[\gamma_3 H] = \beta x \tag{3.4}$$

Hence, we are interested in estimating β consistently. It becomes slightly more complicated when we consider the variable W.

Remark 3.2. *The standard least squares estimator $\hat{\beta}$ of the regression of Y on X is generally biased for β.*

Since $\hat{\beta}$ is the least squares coefficient,[1] *we have:*

$$\hat{\beta} = \frac{\widehat{cov}(X, Y)}{\widehat{var}(Y)}, \tag{3.5}$$

[1]This comes from the estimation $\hat{\beta}$ of the linear regression with one predictor and n data points $X = (x_1, \ldots, x_n)$ and $y = X^T\beta + \varepsilon$. One can show that: $\beta = plim((X^TX)^{-1}Xy) = (plim(X^TX))^{-1}plim(Xy) = \frac{cov(X,y)}{var(X)}$, where $plim$ is the probability limit, that is, $plim_{n\to\infty}(\hat{\theta}_n) = \theta \Leftrightarrow lim_{n\to\infty}\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ for every $\varepsilon > 0$. We used the consistency of the least squares estimate in the first equation and the second equation hold due to Slutsky's Theorem.

where $\widehat{cov}$ and $\widehat{var}$ denote the sample covariance and sample variance respectively. Thus, $\hat{\beta}$ converges to $\frac{cov(X,Y)}{var(X)}$ and we have:

$$\begin{aligned}\frac{cov(X,Y)}{var(X)} &= \frac{cov(X, \beta X + \gamma_3 H + N_y)}{var(X)} = \frac{\beta var(X) + \gamma_3 cov(X,H)}{var(X)} \\ &= \frac{\beta var(X) + \gamma_3\gamma_2 var(H)}{var(X)} = \beta + \gamma_3\gamma_2 \frac{var(H)}{var(X)} \neq \beta.\end{aligned} \tag{3.6}$$

In the first and third equations, we have inserted the equations of Y and X, respectively. Therefore, if $var(H) \neq 0$ the least squares estimator $\hat{\beta}$ is biased for β.

3.2.1 Linear Two-Stage Least Squares

A method to solve the problem described above, in other words, to have a consistent estimate of β, is the 2SLS algorithm. The idea of 2SLS is to extract the variation in the treatment that is uncorrelated with the unmeasured confounder and then regress it with the outcome.

The algorithm first linearly regresses X on Z providing a consistent estimate for α since Z is uncorrelated with H. Therefore, we can consider the term $\gamma_2 H + N_x$ to be the error term in the equation of X in 3.3. Then, we linearly regress Y on the predicted values of X, denoted by $\hat{X} = \hat{\alpha} Z$, giving a consistent estimate for the coefficient β. In fact, in the SCM 3.3, in the equation of Y, we can consider the error as $\gamma_3 H + N_y$. The error term is now uncorrelated with the predictor $\hat{X}$. Thus, the estimate of β is consistent.

We can also see the consistency of $\hat{\beta}$ for β by calculating $\hat{\beta}$ directly.

$$\hat{\beta} = \frac{\widehat{cov}(\hat{\alpha} Z, Y)}{\widehat{var}(\hat{\alpha} Z)}. \tag{3.7}$$

This is consistent with $\frac{cov(\hat{\alpha}Z,Y)}{var(\hat{\alpha}Z)}$ and we have:

$$\frac{cov(\hat{\alpha}Z, Y)}{var(\hat{\alpha}Z)} = \frac{cov(\hat{\alpha}Z, \beta X + \gamma_3 H + N_y)}{var(\hat{\alpha}Z)} = \frac{cov(\hat{\alpha}Z, \beta X)}{var(\hat{\alpha}Z)} = \frac{\hat{\alpha}\beta cov(Z,X)}{\hat{\alpha}^2 var(Z)} = \beta, \tag{3.8}$$

where we used the independence of $\gamma_3 H + N_y$ and Z in the third equation and the first stage in the last equality.

The 2SLS algorithm can be written compactly as in Algorithm 1.

Remark 3.3. *An important note is that we actually do not need that $\hat{\alpha}$ is consistent with α or even that the first stage is linear, but we still need to use least squares regression in the first stage. A reason for this is that $\hat{\alpha}$ cancels out in equation 3.8. A*

Algorithm 1 Linear Two-Stage Least Squares Algorithm

Stage 1
Regress X on Z: $X \sim Z$
Get the predicted values $\hat{X}$
Stage 2
Regress Y on $\hat{X}$: $Y \sim \hat{X}$
The coefficient of $\hat{X}$ is the estimator $\hat{\beta}$

more intuitive idea, which is another proof for the consistency of β, is that we can write the equation of Y in SCM 3.3 as:

$$Y = \beta X + \gamma_3 H + N_y = \beta\hat{X} + (\beta X - \beta\hat{X}) + \gamma_3 H + N_y. \tag{3.9}$$

The second stage then has the error $(\beta X - \beta\hat{X}) + \gamma_3 H + N_y$. A crucial point is that the Least Squares regression produces residuals that are uncorrelated with the predicted values shown in Lemma A.3. Hence, $(\beta X - \beta\hat{X})$ which are the residuals of the first stage multiplied by β are uncorrelated with the predicted values of the first stage $\hat{X}$ which is also the predictor in the second stage. We also know that $\hat{X}$ is uncorrelated with $\gamma_3 H + N_y$ because $\hat{X} = \hat{\alpha}Z$ and Z are uncorrelated with H according to the Assumption 2a. Thus, the predictor in the second stage $\hat{X}$ is uncorrelated with the error $(\beta X - \beta\hat{X}) + \gamma_3 H + N_y$ and therefore $\hat{\beta}$ is consistently estimated.

If the first stage is nonlinear and we perform the 2SLS algorithm, it is still consistent. However, the drawback is that the variance of the error term is larger due to the term $(\beta X - \beta\hat{X})$ resulting in a less accurate estimation. Having a nonlinear estimation in the first stage can solve the problem; more on this in the Remark 3.10.

In the following plots, we compare the performance of the 2SLS algorithm when the first stage is incorrect (Figure 3.3a) and when it is correct (Figure 3.3b). In this example, the SCM is:

$$\begin{aligned} X &= Z^2 + |Z| + Z^3 + H + N_X, \\ Y &= 3X + H + N_Y. \end{aligned} \tag{3.10}$$

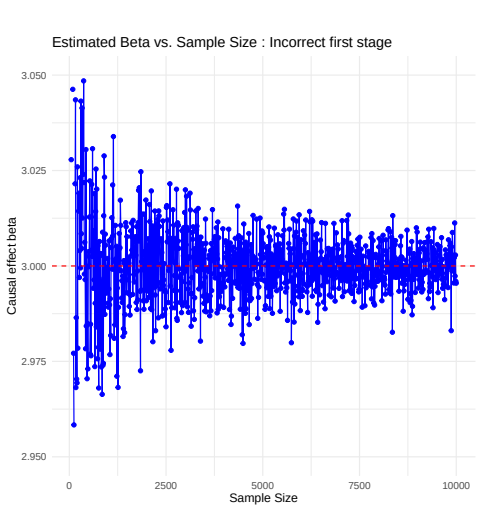
Here, the coefficient of interest is that of X, denoted by $\beta = 3$. In the first regression, when the first stge is incorrect, we perform the linear regression:

$$X \sim Z. \tag{3.11}$$

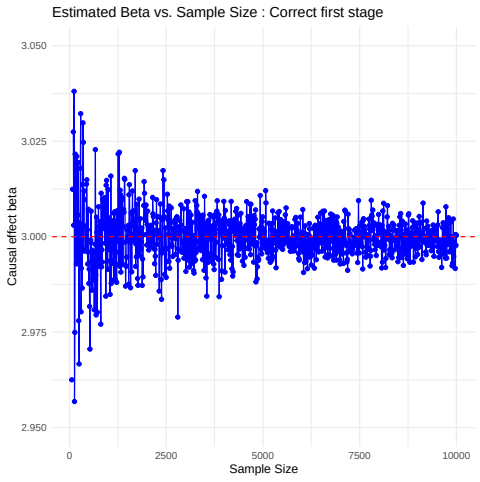
To increase the performance, we can perform in the first stage, the following linear regression:

$$X \sim Z^2 + |Z| + Z^3. \tag{3.12}$$

We see in Figure 3.3b and 3.3a, that the performance increases slightly when the first stage is correctly estimated, because the error in the second stage is smaller, compared



(a) 2SLS algorithm is performed with an incorrect first stage.



(b) 2SLS algorithm is performed with a correct first stage. It converges faster toward zero compared to the plot on the left.

Figure 3.3: The plots suggest consistency in both cases. The Y-axis shows the estimated coefficient $\hat{\beta}$ of $\beta = 3$ and the X-axis is the sample size.

to the case when it is not. However, at least for this example, the differences remain small and both are indeed consistent.

3.2.2 Linear Two-Stage Least Squares with a Covariate

We can extend the linear case by adding one or more covariates W, that is, we would have the following SCM:

$$\begin{aligned} W &= \delta_1 H + N_W, \\ Z &= \delta_2 W + N_Z, \\ X &= \alpha_0 + \alpha_1 Z + \alpha_2 W + \gamma_1 H + N_X, \\ Y &= \beta_0 + \beta_1 X + \beta_2 W + \gamma_2 H + N_Y. \end{aligned} \tag{3.13}$$

Furthermore, we assume that the variables H, W and Z have mean 0. It is associated with the general causal graph as shown at the beginning in the Figure 3.1.

The 2SLS algorithm remains very similar as in the last section. In the first stage, we simply regress X on Z and W and in the second stage we regress Y on X and W.

Since in this case W is correlated with H, the proof is not as straightforward as before.

Theorem 3.4. *Assuming the SCM 3.13, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ of the 2SLS algorithm are consistent with β_0 and β_1.*

Before proving the consistency of the 2SLS algorithm in algorithm 1. We state and prove two lemmata.

The first one is:

Lemma 3.5. *Let A be a non-singular $k \times k$ matrix. The estimated coefficients β of the following multiple linear regression model $y = XA\beta + \varepsilon$ are $\hat{\beta} = A^{-1}\beta$.*

Proof. $\hat{\beta} = ((XA)^T(XA))^{-1}(XA)^T y = A^{-1}(X^TX)^{-1}(A^T)^{-1}A^TX^T = A^{-1}\beta$ □

The second is:

Lemma 3.6. *If a multiple linear regression is given by*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \varepsilon, \tag{3.14}$$

where $x_1, \ldots, x_k$ are uncorrelated with x_{k+1} and ε and where x_{k+1} and ε are correlated. Then we can estimate the coefficients $\beta_0, \ldots, \beta_k$ consistently, but β_{k+1} is inconsistent.

Proof. See Appendix A.1 for the proof. □

Proof. (Theorem 3.4)

The first stage linearly regresses X on Z and W to get the predicted values $\hat{X} = \hat{\alpha}_0 + \hat{\alpha}_1 Z + \hat{\alpha}_2 W$.

$\hat{X}$ is correlated with H, however, since in the second stage we regress on $\hat{X}$ and on W, the part correlated with W in $\hat{X}$ is included in W.

More precisely, we can write Y in the following way by reordering the terms and inserting the structural equation of X:

$$Y = \beta_0 + \alpha_0\beta_1 + \beta_1\alpha_1 N_Z + (\beta_1\alpha_1\delta_2 + \beta_2)W + \gamma_2 H + N_Y, \tag{3.15}$$

since $\hat{\alpha}_1 N_Z$ is uncorrelated with the error $\gamma_2 H + N_Y$, by Lemma 3.6, we know that the intercept $\hat{\beta}_0$ and the coefficient $\hat{\beta}_1$ of $\hat{\alpha}_1 N_Z$, in the linear regression:

$$Y \sim (\hat{\alpha}_1 N_Z) + W, \tag{3.16}$$

are consistent with $\beta_0 + \alpha_0\beta_1$ and β_1. We can now apply the Lemma 3.5 with the following X and A:

$$X = \begin{pmatrix} 1 & \hat{\alpha}_1 N_Z^1 & W^1 \\ & \vdots & \\ 1 & \hat{\alpha}_1 N_Z^n & W^n \end{pmatrix}, A = \begin{pmatrix} 1 & \hat{\alpha}_0 & 0 \\ 0 & 1 & 0 \\ 0 & \hat{\alpha}_2\hat{\delta}_2 + \hat{\alpha}_2 & 1 \end{pmatrix}. \tag{3.17}$$

Now,

$$XA = \begin{pmatrix} 1 & \hat{X}^1 & W^1 \\ & \vdots & \\ 1 & \hat{X}^n & W^n \end{pmatrix}, A^{-1} = \begin{pmatrix} 1 & -\hat{\alpha}_0 & 0 \\ 0 & 1 & 0 \\ 0 & -(\hat{\alpha}_2\hat{\delta}_2 + \hat{\alpha}_2) & 1 \end{pmatrix}, \tag{3.18}$$

by Lemma 3.5, we get that the intercept and the coefficient for $\hat{X}$, in the regression

$$Y \sim \hat{X} + W, \tag{3.19}$$

are equal to the first two entry of the vector $A^{-1}\beta$, which are $\hat{\beta}_0 - \hat{\alpha}_0\hat{\beta}_1$ and $\hat{\beta}_1$ and thus consistent with β_0 and β_1.

□

Remark 3.7. *If we observe X, Y, W and Z, we cannot statistically test the Assumption 3, whether Z enters the equation of Y or not and we cannot also test the Assumption 2, whether Z is directly correlated with H. However, we can statistically test the Assumption 1 with the help of a F-test to know whether it is a weak or strong instrument. In practice, a rule of thumb is used and says that if the F-statistic is smaller than* 10*, then the instrument is weak.*

If we have multiple instruments and we know that at least one satisfies the Assumption 2, then with the so-called J-test often used for overidentifying restrictions we can test if all other instruments also satisfy the Assumption 2.

The linear case is well understood. However, in practice, the relations between the variables are often nonlinear. Therefore, we want to generalise the 2SLS algorithm to be able to capture more complex associations between the variables including the unobserved confounder.

3.3 Nonlinear Additive Effects

In this section, we follow the paper from Guo and Small (2016), the principles of a nonlinear 2SLS can be traced back to Newey and Powell (2003). We assume that the causal effects from Z to X and from X to Y are given by a linear combination of nonlinear functions. We also assume that we observe W that causes both X and Y and that is caused by H. To estimate the causal effect of X over Y, we assume the following SCM:

$$\begin{aligned} X &= g(Z, W, H, N_X), \\ Y &= \beta_0 + \beta_1 f_1(X) + \cdots + \beta_k f_k(X) + \delta_2 W + h_2(H, N_Y), \end{aligned} \tag{3.20}$$

where $f_1, \ldots, f_k$ is a known basis of linearly independent functions, and we assume that $f_1(x) = x$ for identifiability reasons. Since f_1 is linear

$f_2, \ldots, f_k$ must be non-linear otherwise it is not be linearly independent. Furthermore, we assume that W has a linear effect on the outcome Y. N_x, N_y are independent noise, that is, centered and independent with all other variables and we assume that $h_2(H, N_Y)$ has mean 0. We will implicitely assume some condition on g in assumption 1b.

3.3.1 Nonlinear Two-Stage Least Squares

As in the linear case in equation 3.6, we cannot regress directly Y on X and W because H is unobservable and affects X and Y. This can be solved with a generalised version of the linear two-stage least squares in Algorithm 1. From now on, when we say two-stage least squares, we refer to the Algorithm 2 described in this section.

In the first stage, we regress the treatment X on known functions of Z. We denote these functions by $g_1, \ldots, g_k$ and assume that they are linearly independent and assume that $g_1(z) = z$ for identifiability.

We need to modify the Assumptions 1a and 2a as follows.

The following assumption is stronger, but analogous to Assumption 1a. Assumption 1b is crucial for the 2SLS algorithm to work, more details in Remark 3.11.

Assumption 1b. *The instrument is correlated with the treatment in the following sense:*

$$\mathbb{E}\left[\begin{pmatrix} f_1(X) \\ \vdots \\ f_k(X) \end{pmatrix} \begin{pmatrix} g_1(Z) & \ldots & g_k(Z) \end{pmatrix} | W \right] \tag{3.21}$$

has full rank.

Assumption 1b restricts in some sense the function h_1 in SCM 3.20. For example, it assumes that h_1 is not constant in Z otherwise the matrix in assumption 1b is the null matrix and has not full rank.

A weaker assumption than assumption 2 is as follows.

Assumption 2b. *The instrument is uncorrelated with the error in the second stage in the following sense:*

$$\mathbb{E}\left[\begin{pmatrix} g_1(Z) \\ \vdots \\ g_k(Z) \end{pmatrix} h_2(H, N_Y) \right] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.22}$$

The 2SLS algorithm for the nonlinear case is implemented in Algorithm 2.

Algorithm 2 Two-Stage Least Squares Algorithm

Stage 1

Implement the linear regression:

$f_1(X) \sim g_1(Z) + \cdots + g_k(Z) + W$

$\vdots$

$f_k(X) \sim g_1(Z) + \cdots + g_k(Z) + W$

Get the predicted values $(\widehat{f_1(X)}, \ldots, \widehat{f_k(X)})$

Stage 2

Implement the linear regression:

$Y \sim \widehat{X} + \widehat{f_2(X)} + \cdots + \widehat{f_k(X)} + W$

The coefficients of $\widehat{f_1(X)}, \ldots, \widehat{f_k(X)}$ are the estimates $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$ of $\beta = (\beta_1, \ldots, \beta_k)$ and the estimate of the intercept β_0 is denoted by $\hat{\beta}_0$.

Remark 3.8. *A more flexible 2SLS-type of algorithm could allow for a path from W to Z and also from H to W and relax Assumption* 2*b to have:*

$$\mathbb{E}\left[\begin{pmatrix} g_1(Z) \\ \vdots \\ g_k(Z) \end{pmatrix} h_2(H, N_Y) | W \right] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.23}$$

However, Assumption 2*b almost implies independence between W and Z or independence between W and H.*

This is because we only consider W to be additive. If we weaken Assumption 2*b and still assume SCM 3.20, then there might be correlation problems between* $\widehat{f_j(X)}$ *and* $h_2(H, N_Y)$.

We can also note that Z does not enter directly the structural equation of Y in the SCM 3.20.

Theorem 3.9. *Under assumptions* 1*b*, 2*b and assuming SCM 3.20. The estimates* $\hat{\beta} = (\beta_1, \ldots, \beta_k)$ *of the two-stage least squares estimates are consistent.*

The idea of the proof can be found in Guo and Small (2016), we propose a detailed proof.

Proof. (Theorem 3.9)

We prove that the estimate for β provides a consistent estimate. However, in this case, the coefficient of W is inconsistent.

The idea is to prove that all predictors in the second stage are uncorrelated with the error. First, we can write the predicted values, obtained in the first stage, as:

$$\begin{aligned}
\widehat{f_1(X)} &= \hat{\alpha}_0^1 + \hat{\alpha}_1^1 g_1(Z) + \cdots + \hat{\alpha}_k^1 g_k(Z) + \hat{\delta}_1^1 W, \\
&\vdots \\
\widehat{f_k(X)} &= \hat{\alpha}_0^k + \hat{\alpha}_1^k g_1(Z) + \cdots + \hat{\alpha}_k^k g_k(Z) + \hat{\delta}_1^k W.
\end{aligned} \tag{3.24}$$

In fact, all predicted values $\widehat{f_1(X)}, \ldots, \widehat{f_k(X)}$ in 3.24 depend on Z. Indeed, the Assumption 1*b* implies that the matrix of the coefficients in equation 3.24 has full rank (see Remark 3.11 for more details).

Then, we can rewrite the equation of Y in 3.20 as:

$$\begin{aligned}
Y &= \beta_0 + \beta_1 f_1(X) + \cdots + \beta_k f_k(X) + \delta_2 W + h_2(H, N_y) \\
&= \beta_0 + \beta_1 \widehat{f_1(X)} + \cdots + \beta_k \widehat{f_k(X)} + \delta_2 W + \beta_1 (f_1(X) - \widehat{f_1(X)}) \\
&+ \cdots + \beta_k (f_k(X) - \widehat{f_k(X)}) + h_2(H, N_Y),
\end{aligned} \tag{3.25}$$

where $f_j(X) - \widehat{f_j(X)}$ are the residuals of the *jth* regression in the first stage, which are therefore uncorrelated with $g_1(Z), \ldots, g_k(Z), W$. $h_2(H, N_Y)$ is also uncorrelated with $g_1(Z), \ldots, g_k(Z)$ by Assumption 2*b*. However, W is correlated with $h_2(H, N_Y)$.

Since W appears as an additive term in $\widehat{f_j(X)}$ and W is also a predictor in the equation of Y. We can use Lemma 3.5 with

$$X = \begin{pmatrix}
1 & \widehat{X^1} & \widehat{f_2(X^1)} & \ldots & \widehat{f_k(X^1)} & W^1 \\
1 & \widehat{X^2} & \widehat{f_2(X^2)} & \ldots & \widehat{f_k(X^2)} & W^2 \\
1 & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & \widehat{X^n} & \widehat{f_2(X^n)} & \ldots & \widehat{f_k(X^n)} & W^n
\end{pmatrix}, A = \begin{pmatrix}
1 & 0 & \ldots & 0 & 0 \\
0 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & -\hat{\delta}_1^1 & \ldots & -\hat{\delta}_1^k & 1
\end{pmatrix} \tag{3.26}$$

and we see that the predictors in XA are 1,
$(\hat{\alpha}_0^1 + \hat{\alpha}_1^1 g_1(Z) + \cdots + \hat{\alpha}_k^1 g_k(Z)), \ldots, (\hat{\alpha}_0^k + \hat{\alpha}_1^k g_1(Z) + \cdots + \hat{\alpha}_k^k g_k(Z))$ and W. According to Lemma 3.5, the estimated coefficients are:

$$A^{-1}\beta = \begin{pmatrix}
1 & 0 & \ldots & 0 & 0 \\
0 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \hat{\delta}_1^1 & \ldots & \hat{\delta}_1^k & 1
\end{pmatrix} \begin{pmatrix}
\beta_0 \\ \vdots \\ \beta_k \\ \gamma_2
\end{pmatrix} = \begin{pmatrix}
\beta_0 \\ \vdots \\ \beta_k \\ \beta_1 \hat{\delta}_1^1 + \cdots + \beta_k \hat{\delta}_1^k + \gamma_2
\end{pmatrix}. \tag{3.27}$$

Therefore, we get that the coefficients for the first k predictors and intercept in the two following regressions are the same:

$$Y \sim \widehat{f_1(X)} + \cdots + \widehat{f_k(X)} + W, \tag{3.28}$$

$$\begin{aligned} Y \sim &(\hat{\alpha}_0^1 + \hat{\alpha}_1^1 g_1(Z) + \cdots + \hat{\alpha}_k^1 g_k(Z)) + \dots \\ &+ (\hat{\alpha}_0^k + \hat{\alpha}_1^k g_1(Z) + \cdots + \hat{\alpha}_k^k g_k(Z)) + W. \end{aligned} \tag{3.29}$$

Now, since we can write Y as in 3.25, the error of regression 3.28 is equal to:

$$err = h_2(H, N_y) + \beta_1(f_1(X) - \widehat{f_1(X)}) + \cdots + \beta_k(f_k(X) - \widehat{f_k(X)}). \tag{3.30}$$

We know by Assumption 2*b* that Z is uncorrelated with $h_2(H, N_y)$ and since $(f_j(X) - \widehat{f_j(X)})$ is the residuals in the *jth* regression in the first stage. Therefore, the error is uncorrelated with all predictors $g_1(Z), \dots, g_k(Z)$.

Now, we know that the first k predictors in 3.29 are uncorrelated with the error in equation 3.30, Lemma 3.6 shows that the coefficients of the first k predictors in the regression 3.29 are consistently estimated and thus are those in equation 3.28. This ends the proof.

□

Remark 3.10. *An important note, is that it is not required that the first stage is consistently estimated to have consistency in the second stage. In fact, the error given in equation 3.30 is uncorrelated with the predictors of equation 3.29 no matter how well the predicted values describe the outcome.*

This is why the algorithm still gives consistent estimates even though the fit in all regressions are not exact. The drawback is that if the predicted values in the first stage are poorly estimated, then the error in equation 3.30 has a larger variance and there is a loss of information. We will, in Section 4.1, characterise this loss.

In Guo and Small (2016), they assume that the first stage has the following form:

$$X = \beta_0 + \beta_1 f_1(Z) + \cdots + \beta_k f_k(Z) + h_2(H, N_X), \tag{3.31}$$

which is not necessary as long as Assumption 1*b holds.*

Remark 3.10 gives a sense of why the nonlinear case is more difficult than the linear case. Since we can use OLS, the residuals are uncorrelated with the predictors, which gives nice properties. If we try with a nonlinear estimator and implement again a two-stage algorithm, then it might be biased if the first stage is not sufficiently well approximated. This problem is known as the forbidden regression in Angrist and Pischke (2009).

Remark 3.11. *(Assumption* 1*b is necessary)*

We want to show that Assumption 1*b is necessary for the 2SLS algorithm. For simplicity, we prove this assuming that there is no W. Assumption* 1*b is used in the equation 3.24, so there is no multicollinearity between the* $\widehat{f_j(X)}$*'s or constant predictors in the second stage. To see this, we prove the following result.*

Claim 1. *Assuming infinitely many data points, Assumption 1b holds if and only if the matrix of the estimated coefficient of the first stage*

$$\begin{pmatrix} \hat{\alpha}_0^1 & \hat{\alpha}_1^1 & \dots & \hat{\alpha}_k^1 \\ \hat{\alpha}_0^2 & \ddots & & \vdots \\ \vdots & & & \\ \hat{\alpha}_0^k & \dots & & \hat{\alpha}_k^k \end{pmatrix} \tag{3.32}$$

has rank k.

Proof. Since the α's are calculated from the linear regressions in the first stage, the matrix in equation 3.32 can be written as $((X^T X)^{-1} X^T y)^T$, where

$$X = \begin{pmatrix} 1 & g_1(z_1) & \dots & g_k(z_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_1(z_n) & \dots & g_k(z_n) \end{pmatrix}, y = \begin{pmatrix} f_1(x_1) & \dots & f_k(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \dots & f_k(x_n) \end{pmatrix}. \tag{3.33}$$

Since $(X^T X)^{-1}$ has full rank, we know, by a property of the rank, that $((X^T X)^{-1} X^T y)^T$ has full rank if and only if $X^T y$ does. However, $(\frac{1}{n} X^T y)^T$ is a consistent estimate of the matrix in Assumption 1*b*:

$$\mathbb{E}\left[\begin{pmatrix} f_1(X) \\ \vdots \\ f_k(X) \end{pmatrix} \begin{pmatrix} g_1(Z) & \dots & g_k(Z) \end{pmatrix} \right] = \begin{pmatrix} \mathbb{E}[f_1(X) g_1(Z)] & \dots & \mathbb{E}[f_1(X) g_k(Z)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[f_k(X) g_1(Z)] & \dots & \mathbb{E}[f_k(X) g_k(Z)] \end{pmatrix}. \tag{3.34}$$

Thus, the matrix in Assumption 1*b* has full rank if and only if the matrix in 3.32 does. □

Remark 3.12. *(Counterfactual function 2sls)*

The counterfactual function h in 3.2 can, in this case, be written as:

$$h(x, w) = \beta_0 + \beta_1 f_1(x) + \dots + \beta_k f_k(x) + \delta_2 w + \mathbb{E}[h_2(H) | W = w]. \tag{3.35}$$

The counterfactual function is estimated by

$$\hat{h}(x, w) = \hat{\beta}_0 + \hat{\beta}_1 f_1(x) + \dots + \hat{\beta}_k f_k(x) + \hat{\delta}_2 w. \tag{3.36}$$

To estimate the function h consistently, we need the following additional assumption:

$$\mathbb{E}[h_2(H) | W = w] = \mathcal{L}[h_2(H) | W = w]. \tag{3.37}$$

It implies that the relation between $h_2(H)$ and W must be linear.

The reason is that if $h_2(H)$ and W are correlated, then $\hat{\delta}_2$ is not consistent with δ_2. However, with this assumption, we have:

$$\begin{aligned} plim(\hat{\delta}_2)w &= \frac{cov(Y,W)}{var(W)}w \\ &= \frac{\delta_2 var(W) + cov(h_2(H),W)}{var(W)}w \\ &= \delta_2 w + \frac{cov(h_2(H),W)}{var(W)}w \\ &= \delta_2 w + \mathcal{L}[h_2(H)|W=w] \\ &= \delta_2 w + \mathbb{E}[h_2(H)|W=w]. \end{aligned} \tag{3.38}$$

Thus, the counterfactual function h that we estimate,

$$h(x,w) = \hat{\beta}_0 + \hat{\beta}_1 f_1(x) + \cdots + \hat{\beta}_k f_k(x) + \hat{\delta}_2 w, \tag{3.39}$$

converges in probability for all $x, w \in \mathbb{R}$ to the counterfactual function in equation 3.35.

Chapter 4

Control Function

A different approach for estimating the causal effect in an instrumental variables setting is through control function methods. The idea is to decompose the error into two uncorrelated parts: one correlated with the treatment, denoted e_1, and one that is uncorrelated. The correlated part is called the control function, while the uncorrelated part becomes the new error in the regression $Y \sim f_1(X) + \cdots + f_k(X) + W + e_1$. This decomposition allows for a consistent estimation of the treatment coefficient.

The idea is fundamentally different as with the 2SLS algorithm since we try here to learn the confounder of the second stage, where as in the 2SLS we extract the uncorrelated part in the treatment from the confounder.

In this section, we consider the following SCM, which differs in the error of the first stage from the one in the last section.

$$\begin{aligned} X &= \alpha_0 + \alpha_1 g_1(Z) + \cdots + \alpha_k g_k(Z) + \delta_1 W + h_1(H, N_X), \\ Y &= \beta_0 + \beta_1 f_1(X) + \cdots + \beta_k f_k(X) + \delta_2 W + h_2(H, N_Y), \end{aligned} \tag{4.1}$$

where $f_1, \ldots, f_k$ and $g_1, \ldots, g_k$ are a known basis of linearly independent functions, and we assume that $f_1(x) = x, g_1(z) = z$ for identifiability reasons. N_x, N_y are independent noise, that is, centred and independent of all other variables and we assume that $\mathbb{E}[h_1(H, N_X)] = \mathbb{E}[h_2(H, N_Y)] = 0$.

The Assumption 1*b* is not required for the CF algorithm. Instead, we need to be able to consistently estimate the error $h_1(H, N_X)$ by the residuals of the first stage. Therefore, there is a trade-off between:

- 2SLS: Not assuming any form for the treatment X and assuming Assumption 1*b*.
- CF: Assuming a specific form for the treatment X (as in SCM 4.1) without assuming assumption 1*b*.

In the paper by Guo and Small (2016), they still assumed Assumption 1*b* for the CF algorithm.

For CF, we need two additional assumptions:

Assumption 4. *$h_1(H, N_X)$, $h_2(H, N_Y)$ are pairwise independent of Z, W, that is*

$$\begin{aligned} &Z \perp\!\!\!\perp h_1(H, N_X), Z \perp\!\!\!\perp h_2(H, N_Y) \text{ and} \\ &W \perp\!\!\!\perp h_1(H, N_X), W \perp\!\!\!\perp h_2(H, N_Y). \end{aligned} \tag{4.2}$$

Remark 4.1. *Assumption 4 gives restrictions on $h_1(H)$, whereas in the 2SLS algorithm we did not impose anything on $h_1(H)$. We will see in Example 4.5 a case where Assumption 4 is not satisfied.*

Assumption 5. *The relation between $h_1(H, N_X)$ and $h_2(H, N_Y)$ must be linear, that is, $\mathbb{E}[h_2(H, N_Y)|h_1(H, N_X)] = \rho h_1(H, N_x)$, where $\rho = \frac{\mathbb{E}[h_2(H,N_Y)h_1(H,N_X)]}{\mathbb{E}[h_1(H,N_X)^2]}$.*

Remark 4.2. *We list some remarks about Assumptions 4 and 5.*

- *A sufficient condition for the Assumption 5 is when $h_2(H, N_Y) = h_2(H) + N_Y$, $h_1(H, N_X) = h_1(H) + N_X$ and $h_2(H) = \gamma h_1(H)$. Indeed,*

$$\begin{aligned} &\mathbb{E}[h_2(H) + N_Y|h_1(H) + N_X] \\ &= \mathbb{E}[\gamma h_1(H)|h_1(H) + N_X] \\ &= \gamma'(h_1(H) + N_X). \end{aligned} \tag{4.3}$$

  The second equation holds because the relation between $\gamma h_1(H)$ and $h_1(H) + N_X$ is linear, and therefore the conditional expectation is exactly the fitted values. γ' is not equal to γ and more generally, γ' is equal to:

$$\begin{aligned} \gamma' &= \frac{cov(h_1(H) + N_x, \gamma h_1(H))}{var(h_1(H) + N_x)} = \frac{\gamma var(h_1(H))}{var(h_1(H)) + var(N_X)} \\ &= \frac{\mathbb{E}[h_1(H)\gamma h_1(H)]}{\mathbb{E}[(h_1(H) + N_X)^2]} = \rho. \end{aligned} \tag{4.4}$$

  Thus, the Assumption 5 is satisfied.

- *In Section 4.3 and in Remark 4.10, we will also see that it is possible to relax Assumption 4 by allowing W to depend on Z and H. However, the independence between Z and H is necessary for the CF algorithm. It is the main difference from the 2SLS algorithm, where they do not need the full independence.*

- *In Section 4.3 and in Remark 4.10, we will show that it is possible, using a more general method of CF based on the same principle, to avoid assuming Assumption 5.*

Remark 4.3. *The equation of X in the SCM 3.20 must be correct (or well approximated). This differs from the 2SLS algorithm, where a nonlinear relationship was allowed to have consistency even if we used a linear model for the first stage (see Remark 3.10).*

Algorithm 3 Control Function Algorithm

Stage 1

Implement the linear regression:
$X \sim g_1(Z) + \cdots + g_k(Z) + W$
Get the predicted values $\hat{X}$ and the residuals $e_1 = X - \hat{X}$.

Stage 2

Implement the linear regression:
$Y \sim f_1(X) + \cdots + f_k(X) + W + e_1$
The coefficient of $f_1(X), \ldots, f_k(X)$ are the estimates $\hat{\beta}$ of β

Theorem 4.4. *Assuming the SCM 3.20 and assumptions 4 and 5. The CF method in algorithm 3 gives a consistent estimate of $\beta_0, \ldots, \beta_k$.*

Proof. We first define $e = h_2(H, N_Y) - \rho(h_1(H, N_X))$ and rewrite the equation of Y in the SCM 4 as:

$$Y = \beta_0 + \beta_1 f_1(X) + \cdots + \beta_k f_k(X) + \delta_2 W + \rho(h_1(H, N_X)) + e. \tag{4.5}$$

We want now to show that all predictors are uncorrelated with the error e. Due to Assumption 4, e is uncorrelated with W. To see that e is uncorrelated with $f_i(X)$ for $i = 1, \ldots, k$:

$$\begin{aligned} \mathbb{E}[f_i(X)e] &= \mathbb{E}[\mathbb{E}[f_i(X)e|X, W, Z]] \\ &= \mathbb{E}[f_i(X)\mathbb{E}[e|h_1(H, N_X), W, Z]] \\ &= \mathbb{E}[f_i(X)\mathbb{E}[e|h_1(H, N_X)]], \end{aligned} \tag{4.6}$$

where we used Assumption 4 in the third equation. According to Assumption 5, we have:

$$\begin{aligned} &\mathbb{E}[e|h_1(H, N_X)] = \mathbb{E}[h_2(H, N_Y) - \rho h_1(H, N_X)|h_1(H, N_X)] \\ &= \mathbb{E}[h_2(H, N_Y)|h_1(H, N_X)] - \rho h_1(H, N_X) = 0. \end{aligned} \tag{4.7}$$

Therefore, we have $\mathbb{E}[f_i(X)e] = 0$.

In the first stage, the error $h_1(H, N_X)$ converges to the residuals e_1. In fact, it converges because, according to Assumption 4, each predictor is uncorrelated with error $h_1(H, N_X)$.

Therefore, all predictors are uncorrelated with the error e. Finally, since the residuals e_1 are approximations that converge to the error e and therefore are not exactly the same. We need to use Theorem A.2, see Appendix A. In this case, $\hat{\theta} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_k)$, $x_1 = (X, g_1(Z), \ldots, g_k(Z))$,

$$f(\hat{\theta}, x_1) = X - (\hat{\alpha}_0 + \hat{\alpha}_1 g_1(Z) + \cdots + \hat{\alpha}_k g_k(Z)), \gamma = \rho \tag{4.8}$$

and

$$x_2 = \begin{bmatrix} 1 \\ f_1(X) \\ \vdots \\ f_k(X) \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{4.9}$$

If the two following regularity conditions are satisfied:

1. $lim_{n\to\infty} A^T A = Q_0$, where Q_0 is a positive definite matrix.
2. $lim_{n\to\infty} \frac{1}{n} A^T F^* = Q_1$, where Q_1 is a matrix with entry in $\mathbb{R}$.

Then the estimate $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_k)$ of the CF algorithm is consistent. For more details on A, F^* see in the appendix in Theorem A.2. □

Example 4.5. *We show an example that illustrates the fact that the error of the first stage in 2SLS does not need to be independent of Z. In opposition with the CF algorithm where we require that the errors of the first and second stage are independent with Z as in Assumption 4. Consider the following SCM:*

$$\begin{aligned} X &= Z + ZH + H, \\ Y &= sin(X) + H, \end{aligned} \tag{4.10}$$

where $Z \sim \mathcal{N}(0, 0.5)$ *and* $H \sim \mathcal{N}(0, 2)$

As we can see in Figure 4.1a, the CF algorithm is biased because the error in the first stage is $ZH + H$ *and* ZH *is uncorrelated with the error in the second stage* H *but is correlated with* $sin(X)$*. As we can see in Figure 4.1b the 2SLS is consistent. Both algorithms are performed using the basis:*

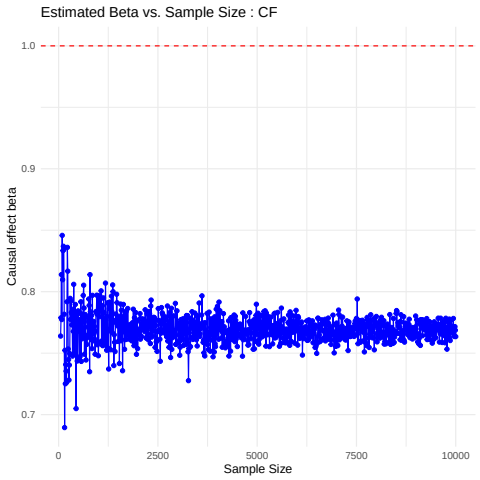
$$\begin{aligned} g_1(Z) &= Z, \\ f_1(X) &= sin(X). \end{aligned} \tag{4.11}$$

Remark 4.6. *(Counterfactual Function for the Control Function Algorithm)*

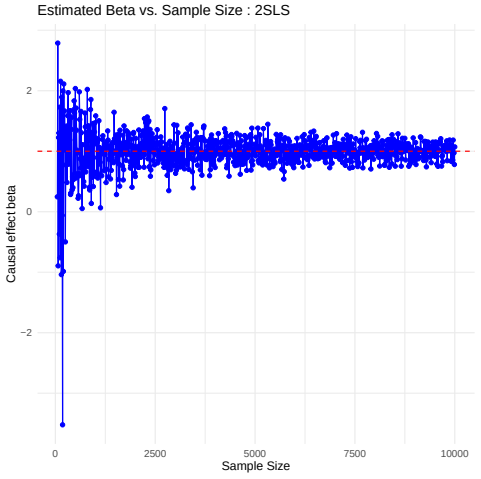
For the CF algorithm, the counterfactual function is given by

$$h(x, w) = \beta_0 + \beta_1 f_1(x) + \cdots + \beta_k f_k(x) + \delta_2 w. \tag{4.12}$$

Since we assume independence between confounders H and instruments Z, and between confounders H and covariates W in Assumption 4. The conditional expectation



(a) CF algorithm is biased because the error in the first stage depends on Z, even though it is uncorrelated. The Y-axis shows the estimated coefficient $\hat{\beta}$ of $\beta = 1$.



(b) 2SLS algorithm is consistent even though the error in the first stage depends on Z. The Y-axis shows the estimated coefficient $\hat{\beta}$ of $\beta = 1$.

$\mathbb{E}[h_2(H)|W = w]$ *is a constant zero function. Thus, the estimated counterfactual function,*

$$\hat{h}(x, w) = \hat{\beta}_0 + \hat{\beta}_1 f_1(x) + \cdots + \hat{\beta}_k f_k(x) + \hat{\delta}_2 w, \tag{4.13}$$

converges to the true counterfactual function 4.12.

Example 4.7. *In the following, we look at an example that illustrates that the control function algorithm does not require Assumption* 1*b as for the 2SLS algorithm. Let us consider the following example. First, consider the following SCM:*

$$\begin{aligned} X &= Z + Z^3 + H, \\ Y &= X + cos(X) + H, \end{aligned} \tag{4.14}$$

where $H \sim \mathcal{U}[-\frac{\pi}{2}, \frac{\pi}{2}]$ *and* $Z \sim \mathcal{U}[-1, 1]$*. For simplicity in the calculation, we do not consider noises* N_X, N_Y*.*

Assume that we take $g_1(z) = z$*,* $g_2(z) = Z^3$ *and* $f_1(x) = x$*,* $f_2(x) = cos(x)$*. Then, the Assumption* 1*b is not satisfied because the following matrix has not full rank.*

$$\begin{pmatrix} \mathbb{E}[XZ] & \mathbb{E}[XZ^3] \\ \mathbb{E}[cos(X)Z] & \mathbb{E}[cos(X)Z^3] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Z^2] + \mathbb{E}[Z^4] & \mathbb{E}[Z^4] + \mathbb{E}[Z^6] \\ 0 & 0 \end{pmatrix}, \tag{4.15}$$

where $\mathbb{E}[Z^2], \mathbb{E}[Z^4] > 0$*. Hence, the matrix has rank* $1 < 2$*. One can see, that:*

$$\begin{aligned} \mathbb{E}[cos(X)Z] &= \mathbb{E}[cos(Z + Z^3 + H)Z] \\ &= \int_{-1}^{1} \int_{-\pi}^{\pi} cos(z + z^3 + h) z p(z, h) dh dz. \end{aligned} \tag{4.16}$$

Since Z and H are independent, we can write for $-1 \leq z \leq 1$ and $-\pi \leq h \leq \pi$:

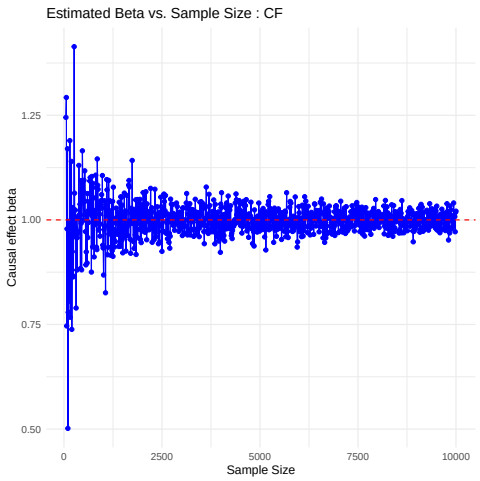
$$p(z,h) = p(z)p(h) = \frac{1}{4\pi}. \tag{4.17}$$

Thus, we get that the integral in the middle vanishes because we integrate over the whole period of cos with respect to h.

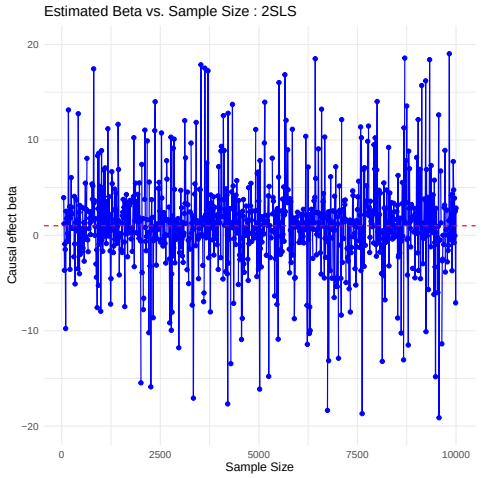
$$\int_{-\pi}^{\pi} cos(z + z^3 + h)zp(z,h)dh = \frac{z}{4\pi} \int_{-\pi}^{\pi} cos(z + z^3 + h)dh = 0. \tag{4.18}$$

Therefore, $\mathbb{E}[cos(X)Z] = 0$ and analogously we find $\mathbb{E}[cos(X)Z^3] = 0$.

The problem is that the predictor $\widehat{cos(X)}$ of the second stage is almost constant, which can be seen as multicollinearity between the predictor and the intercept. As we can see in Figures 4.2a and 4.2b, the 2SLS method is inconsistent, whereas the control function algorithm is consistent.



(a) CF algorithm is consistent. The Y-axis shows the estimated coefficient $\hat{\beta}_2$ of $\beta_2 = 1$ of the predictor $\widehat{cos}(X)$.



(b) 2SLS algorithm is inconsistent because Assumption $1b$ does not hold. The Y-axis shows the estimated coefficient $\hat{\beta}_2$ of $\beta_2 = 1$ of the predictor $\widehat{cos}(X)$.

4.1 Link between Two-Stage Least Squares and Control Function

The CF algorithm is better than the 2SLS if assumptions 4 and 5 hold because it loses less information than the 2SLS algorithm. In the following, we describe and characterise this loss of information.

Theorem 1 in Guo and Small (2016) states and proves the following theorem:

Theorem 4.8. *The regression:*

$$Y \sim \widetilde{X} + \widetilde{f_1(X)} + \cdots + \widetilde{f_k(X)} + W \tag{4.19}$$

gives the same estimate as in the second stage of the CF algorithm for the corresponding coefficients:

$$Y \sim X + f_1(X) + \cdots + f_k(X) + W + e_1, \tag{4.20}$$

where $\widetilde{f_j(X)}$ are the residuals of the regression $f_j(X) \sim e_1$.

Proof. See proof Theorem 1 in Guo and Small (2016). □

Since $\widetilde{f_j(X)}$ are the residuals of $f_j(X) \sim e_1$, $\widetilde{f_j(X)}$ is uncorrelated with e_1. Therefore, $\widetilde{f_j(X)}$ is the projection of $f_j(X)$ on $(span(e_1))^T$ $j = 1, \ldots, k$, where $(span(e_1))^T$ is the orthogonal linear span of e_1, that is, the space of all random variables, which are uncorrelated with e_1. Therefore, Theorem 4.8 shows that the CF algorithm loses information by projecting $f_1(X), \ldots, f_k(X)$ on $(span(e_1))^T$.

In the 2SLS Algorithm 2, we lose information by projecting $f_1(X), \ldots, f_k(X)$ on $span(g_1(Z), \ldots, g_k(Z), W)$. However, since e_1 is the residual of the regression $X \sim g_1(Z) + \cdots + g_k(Z) + W$, we know that e_1 is uncorrelated with $g_1(Z), \ldots, g_k(Z), W$. Hence, we have:

$$span(Z, g_2(Z), \ldots, g_k(Z), W) \subseteq (span(e_1))^T. \tag{4.21}$$

Since we lose less information projecting on a larger space, the CF algorithm performs better than the 2SLS algorithm.

To characterise the extra part in $span(e_1)^T$ that is not included in $span(g_1(Z), \ldots, g_k(Z), W)$, we will show that the CF algorithm is equivalent to the 2SLS algorithm with an augmented set of instrumental variables $\{g_1(Z), \ldots, g_k(Z), error.iv_1, \ldots, error.iv_{k-1}\}$ in Algorithm 4.

First, we need Theorem 2 in Guo and Small (2016):

Theorem 4.9. *The following two regressions give the same estimate of the corresponding coefficients:*

$$f_j(X) \sim g_1(Z) + \cdots + g_k(Z) + W, \tag{4.22}$$

and

$$\widetilde{f_j(X)} \sim g_1(Z) + \cdots + g_k(Z) + W. \tag{4.23}$$

Proof. See Guo and Small (2016) for the proof. □

Algorithm 4 Two-Stage Least Squares Algorithm with an Augmented Set of IVs

Stage 1

Implement the linear regression:

$f_1(X) \sim g_1(Z) + \cdots + g_k(Z) + W + error.iv_1 + \cdots + error.iv_{k-1}$

$\vdots$

$f_k(X) \sim g_1(Z) + \cdots + g_k(Z) + W + error.iv_1 + \cdots + error.iv_{k-1}$

Get the predicted values $(\widehat{f_1(X)}, \ldots, \widehat{f_k(X)})$

Stage 2

Implement the linear regression:

$Y \sim \widehat{f_1(X)} + \cdots + \widehat{f_k(X)} + W$

The coefficient of $\widehat{f_1(X)}, \ldots, \widehat{f_k(X)}$ are the estimates $\hat{\beta}$ of β

The set $\{error.iv_1, \ldots, error.iv_{k-1}\}$ is constructed as follows:

1. $error.iv_1$: take the residuals from the regression in 4.23 for $j = 2$, call it $error.iv_1$.
2. $error.iv_{j-1}, j = 3, \ldots, k$: take the residuals from the regression 4.23 for $j = j$ and linearly regress it with $error.iv_1, \ldots, error.iv_{j-2}$. The residuals are $error.iv_{j-1}$.

In the following, we construct more precisely the augmented set of instrumental variables and we prove that the CF algorithm is equivalent to the 2SLS algorithm with the augmented set of instrumental variables.

We denote the residuals of Equation 4.23 for $j = 2$ by $error.iv_1$.
By Theorem 4.9, we can write $error.iv_1$ as:

$$error.iv_1 = \widetilde{f_2(X)} - (\alpha_0 + \alpha_1 g_1(Z) + \cdots + \alpha_k g_k(Z) + \gamma_1 W), \tag{4.24}$$

where $(\alpha_0 + \alpha_1 g_1(Z) + \cdots + \alpha_k g_k(Z) + \gamma_1 W)$ is the predicted values in equation 4.22.
By definition of $\widetilde{f_2(X)}$ and by equation 4.24, we can write $f_2(X)$ as:

$$\begin{aligned} f_2(X) &= \gamma_c + \gamma e_1 + \widetilde{f_2(X)} \\ &= \gamma_c + \gamma e_1 + \alpha_0 + \alpha_1 g_1(Z) + \cdots + \alpha_k g_k(Z) + \gamma_1 W + error.iv_1. \end{aligned} \tag{4.25}$$

Since $g_1(Z), \ldots, g_k(Z), W$ and $\widetilde{f_2(X)}$ are uncorrelated to e_1, by equation 4.24, $error.iv_1$ is uncorrelated to e_1.

Hence, according to 4.25, γe_1 is the error and
$\gamma_c + \alpha_0 + \alpha_1 g_1(Z) + \cdots + \alpha_k g_k(Z) + \gamma_1 W + error.iv_1$ are the predicted values of the regression:

$$f_2(X) \sim g_1(Z) + \cdots + g_k(Z) + W + error.iv_1. \tag{4.26}$$

Thus, the predicted values of the regression 4.26 are equal to $\widetilde{f_2(X)}$ up to an additive constant.

We can then construct $error.iv_j$, by first defining $error.iv_jprime = \widetilde{f_{j+1}(X)} - (\alpha_0^{j+1} + \alpha_1^{j+1} g_1(Z) + \cdots + \alpha_k^{j+1} g_k(Z) + \gamma_1^{j+1} W)$ for $j = 2, \ldots, k-1$, where $\alpha_0^{j+1} + \alpha_1^{j+1} g_1(Z) + \cdots + \alpha_k^{j+1} g_k(Z) + \gamma_1^{j+1} W$ are the predicted values of the *jth* regression in the first step of Algorithm 4. We can now define recursively $error.iv_j$ as:

$$error.iv_j \coloneqq residuals(error.iv_jprime \sim error.iv_1 + \cdots + error.iv_{j-1}). \quad (4.27)$$

Therefore, $error.iv_j$ is uncorrelated of $error.iv_{j'}$ for $j \neq j'$. In the *jth* regression in the first stage of algorithm 4,

$$f_j(X) \sim g_1(Z) + \cdots + g_k(Z) + error.iv_1 + \cdots + error.iv_{k-1}, \quad (4.28)$$

for $j = 2$, because of equation 4.25 and the independence of $error.iv_{j-1}$ with all other predictors, the coefficients of $error.iv_2, \ldots, error.iv_{k-1}$ are 0. Analogously, for all the other j, in equation 4.28 one can show (as in Guo and Small (2016)) that the coefficients of $error.iv_j, \ldots, error.iv_{k-1}$ are 0.

Now we can put everything together. Both regressions 4.26 and 4.28 have the same predicted values. Thus, they are also equal to $\widetilde{f_2(X)}$ up to a constant which by Theorem 4.8, we have that the regression, in the second stage of algorithm 4, provides the same estimate for the corresponding coefficient as the control function algorithm.

One can also prove, that if the assumptions 4 and 5 are satisfied, then the set $\{g_1(z), \ldots, g_k(Z), error.iv_1, \ldots, error.iv_{k-1}\}$ is a valid set of instrumental variables.

Therefore, the control function algorithm is equivalent to a two-stage least squares algorithm with an augmented set of random variables $\{g_1(Z), \ldots, g_k(Z), error.iv_1, \ldots, error.iv_{k-1}\}$ as shown in algorithm 4.

In the 2SLS algorithm with the augmented set of instrumental variables, we can interpret the loss of information as the projection of $f_1(X), \ldots, f_k(X)$ on $span(g_1(Z), \ldots, g_k(Z), W, error.iv_1, \ldots, error.iv_{k-1})$. Since we have proved that both algorithms provide the same estimate, it must be $span(e_1)^T = span(g_1(Z), \ldots, g_k(Z), W, error.iv_1, \ldots, error.iv_{k-1})$.

To summarise, the fact that the estimation with the control function is much better is because the regressions in the first stage of the least squares algorithm are poorly estimated. However, an augmented set of instrumental variables can be found to increase the performance of the 2SLS algorithm.

4.2 Approximating Smooth Functions With Splines

In practice, if the basis of functions is not known, we can modify the 2SLS and CF algorithms to be able to learn arbitrary smooth functions. We can use the smoothing spline regression, which is a piecewise polynomial of degree d. The splines have a basis representation of functions, which can then be linearly combined to give a flexible and smooth approximation of any smooth curve.

This section is based on Eubank (1988) to describe natural cubic smoothing splines. They are defined as the function that minimises the following penalised regression over the set of smooth functions where $\int_0^1 (f''(t))^2 dt < \infty$, that is, the Sobolev space $W_2^2[0,1]$:

$$\frac{1}{k}\sum_{i=1}^{k}(y_i - f(\xi_i)) + \lambda \int_0^1 (f''(t))^2 dt. \tag{4.29}$$

Here, $\lambda > 0$ is the smoothing parameter and $\xi_1 < \cdots < \xi_k$ are k knots that are k chosen values. It can be proven (for example, in Eubank (1988)) that the unique minimiser is a natural cubic smoothing spline. If we have k knots, we fit $k-1$ polynomials of degree 3. Since each polynomial is of the form $\sum_{i=0}^{3} \alpha_i x^i$,

- There are $4(k-1)$ parameters to estimate.

We want a smooth curve, so that at each inner knot $(\xi_2, \ldots, \xi_{k-1})$, the values, the first and second derivatives must match.

- Which gives $3(k-2)$ conditions.

To reduce the variance at the outer knot ξ_1, ξ_k, we add the two "natural" conditions, of which the second derivative must be 0:

$$p''(\xi_1) = p''(\xi_k) = 0. \tag{4.30}$$

This gives,

- $3(k-2)$ conditions.
- 2 natural conditions.

The smoothing spline then has $df = 4(k-1) - [3(k-2)+2] = k$ degrees of freedom. It can also be proven that we can write the smoothing spline as a linear combination of a basis consisting of n functions $f_1, \ldots, f_k$:

$$s(x) = \sum_{i=1}^{k} \beta_i f_i(x). \tag{4.31}$$

In practice, the basis function $f_1, \ldots, f_k$ can, for example (see Hastie et al. (2009), p.145), be given by:

$$\begin{aligned} &f_1(x) = 1, f_2(x) = x, f_{j+2}(x) = d_j(x) - d_{k-1}(x), \\ &d_j(x) = \frac{(x-\xi_j)_+^3 - (x-\xi_k)_+^3}{\xi_j - \xi_k}, \quad j = 0, \ldots, k-2. \end{aligned} \tag{4.32}$$

The knots $\xi_1, \ldots, \xi_k$ can, for example, be the k-quantiles of x.

4.2.1 Splines-Based Two-Stage Least Squares

We can use the spline representation for both stages for the 2SLS algorithm. We can assume a similar SCM as the one we had for the nonlinear 2SLS in equation 3.20, but the difference is that we can assume a smooth function of Z instead of knowing the basis of functions. The SCM can be described as:

$$\begin{aligned} X &= g(Z, W, H, N_X), \\ Y &= f(X) + \gamma_2 W + h_2(H, N_Y), \end{aligned} \tag{4.33}$$

with smooth functions g, f and we assume that $\mathbb{E}[h_2(H, N_Y)] = 0$. Furthermore, we still assume Assumption $1b$ and $2b$.

To estimate $g(Z)$ and $f(X)$, we take the basis of a natural cubic splines with k degrees of freedom. $g_1, \ldots, g_k$ are for the first stage and $f_1, \ldots, f_k$ for the second.

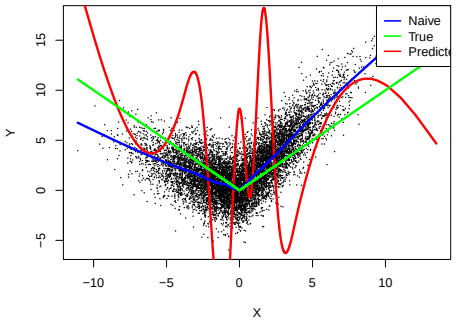
A problem that arises when we implement the 2SLS algorithm with splines is that Assumption $1b$ is no longer satisfied and as shown in Remark 3.11, it is needed otherwise we have multicollinearity between the predictors in the second stage. A potential reason why the Assumption 2 is no longer satisfied is that there is not enough variability between the functions in the spline basis.

A method to solve this issue is to increase the number of degrees of freedom in the first stage without changing it for the second stage so that the matrix is more likely to have full row rank. In the first stage, we can add one degree of freedom to each regression.

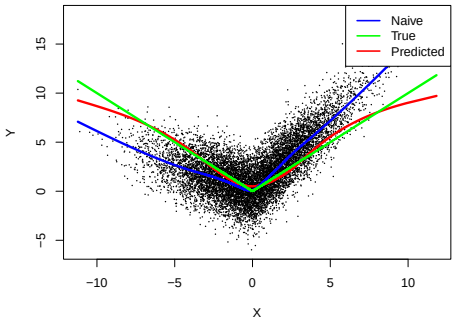
To illustrate this, consider the following example. Here, we have the following SCM:

$$\begin{aligned} X &= Z + H + N_X, \\ Y &= |X| + H + N_Y, \end{aligned} \tag{4.34}$$

where $H \sim \mathcal{N}(0, 2), Z \sim \mathcal{N}(0, 2)$ and $N_X, N_Y \sim \mathcal{N}(0, 1)$ and the number of observations $n = 10000$.



(a) 2SLS estimator when the degrees of freedom for the first and second stage are equal to 10. The estimator is very unstable and inaccurate.



(b) 2SLS estimator when the degrees of freedom is increased for the first stage and 10 for the second stage. It becomes more stable and can learn the true causal effect.

As we can see in Table 4.1, the Variance Inflation Factor (VIF) of the second stage, which is a measure of collinearity between the predictors in the second stage, is much higher without regularisation than with it. Although both VIFs are very high, the one for the 2SLS without regularisation does not seem to be a problem.

	$\widehat{X}$	$\widehat{f_2(X)}$	$\widehat{f_3(X)}$	$\widehat{f_4(X)}$	$\widehat{f_5(X)}$
Without regularization	6.06e+05	1.96e+07	1.48e+09	9.94e+09	1.42e+10
With regularization	2203	13447	20332	21217	41654
	$\widehat{f_6(X)}$	$\widehat{f_7(X)}$	$\widehat{f_8(X)}$	$\widehat{f_9(X)}$	$\widehat{f_{10}(X)}$
Without regularisation	8.77e+09	4.84e+09	1.92e+09	3.43e+08	1.04e+07
With regularization	28880	19205	32757	29345	8509

Table 4.1: VIF factor for the 2SLS with regularisation and 2SLS without regularization.

4.2.2 Splines-Based Fontrol Function Algorithm

We can apply the same principle to the control function algorithm by simply using splines for the first and second stages. This allows us to use the following generalization of the SCM 4.1:

$$\begin{aligned} X &= g(Z) + \gamma_1 W + h_1(H, N_X), \\ Y &= f(X) + \gamma_2 W + h_2(H, N_Y), \end{aligned} \tag{4.35}$$

where we assume that g and f are smooth and also that $\mathbb{E}[h_1(H, N_X)] = \mathbb{E}[h_2(H, N_Y)] = 0$.

The fact that we do not need Assumption 1*b* makes the algorithm with splines very flexible, that is, we can increase the number of degrees of freedom without harming the regression and without any problem with multicollinearity, which allows, if the assumptions regarding the structure of H are met, a very flexible and accurate estimation of the causal effect. More details can be found in Chapter 6.

Remark 4.10. *A generalisation of the control function algorithm is to expand the control function, which is the residuals of the first stage, used in the second stage as predictor with splines. In this case, the second stage is:*

$$Y \sim f_1(X) + \cdots + f_k(X) + W + h_1(res) + \cdots + h_k(res), \tag{4.36}$$

where $f_1, \ldots, f_k$ and $h_1, \ldots, h_k$ are two splines basis functions for X and res respectively. This allows us to avoid assuming Assumption 4.

Analogously, we can use a spline for W and weaken Assumption 5 by only assuming the independence between Z and H.

4.3 General Control Function

In a recent paper by Puli and Ranganath (2021), a generalisation of the control function method and principles has been elaborated. They also implemented an algorithm to learn the confounding effect in the first stage without assuming an additive form.

This paper clarifies the main idea of the control function, which is to find a variable that controls the treatment so that once we condition the control function on the treatment, the treatment is independent of the outcome. Not to be confused with interpreting a control function as a variable that is used to reconstruct the outcome together with the treatment. This view is often true and if available works well, but this is less general in the sense that we find the control function with the first stage. The main idea of the general control function is given in Theorem 1 in Puli and Ranganath (2021), which characterises a control function that can be used, with the treatment X, to reconstruct the outcome Y. Consider first the following SCM:

$$\begin{aligned} X &= g(Z, H, N_X), \\ Y &= f(X, H) + N_Y, \end{aligned} \tag{4.37}$$

where N_X and N_Y are independent noise. For simplicity, we do not consider an observed confounder W. However, it can be added.

We recall here that $p(a|b)$ denotes the density of the random variable $A = a$ given $B = b$. The random variable is always in uppercase and the value it takes is denoted in lowercase.

The following theorem is due to Puli and Ranganath (2021).

Theorem 4.11. *(Meta-Identification Result for Control Function)*

Assume that the data have been generated according to the SCM 4.37. Assume the following:

1. *(A1) The control function $\widehat{H}$ satsfies the **reconstruction** property: $\exists d : X = d(\widehat{H}, Z)$.*
2. *(A2) The IV is **jointly independent** of the control function $\widehat{H}$, confounder H and noises N_X, N_Y: $Z \perp\!\!\!\perp (H, \widehat{H}, N_X, N_Y)$.*
3. *(A3) **Strong IV**. For any compact set $B \subseteq supp(X)$, $\exists c_B > 0$ such that for almost every $x \in B, p(x|h, n_X) \geq c_B$.*

*Then, the control function $\hat{H}$ satisfies **ignorability** and **positivity**, respectively:*

$$\begin{aligned} &p(y|x, \hat{h}) = p(y|do(X = x), \hat{h}) \text{ and} \\ &\text{a.e. in } supp(X) : p(\hat{h}) > 0 \Rightarrow p(x|\hat{h}) > 0. \end{aligned} \tag{4.38}$$

Therefore, the true causal effect is uniquely determined by $\widehat{H}, X, Y$ for almost every $x \in supp(X)$:

$$\mathbb{E}_{\widehat{H}}[\mathbb{E}[Y|X = x, \widehat{H}]] = \mathbb{E}_{\widehat{H}}[\mathbb{E}[Y|do(X = x), \widehat{H}]] = \mathbb{E}[Y|do(X = x)], \tag{4.39}$$

where $\mathbb{E}_{\widehat{H}}[\cdot]$ denotes the expectation over $\widehat{H}$.

Remark 4.12. *The second assumption (A2) in Theorem 4.11 and, more specifically, the assumption that $Z \perp\!\!\!\perp \widehat{X}$ is essential for the CF algorithm. If we take the residuals as the CF, as we have done so far, then the first stage must be consistently estimated so that the CF is independent with the IV Z and uncorrelation between Z and $\hat{H}$ is not enough as in Example 4.5.*

Remark 4.13. *Theorem 4.11 is called meta-theorem because we it does not tell us how to guarantee joint independence in assumption $(A2)$.*
In Puli and Ranganath (2021), some special cases, such as when the confounder is additive in the treatment, where the joint independence is satisfied given the independence between the IV and the confounder, which holds by definition of an IV and the independence between the IV and the CF, which can be guaranteed since both are observable.

Lemma 4.14. *Ignorability, as in the Theorem 4.11, is equivalent to $Y_x \perp\!\!\!\perp X|\widehat{H}$.*

Proof. We first denote $Y_x = [Y|do(X = x)]$ as the potential outcome given a treatment $X = x$. We then have:

$$\begin{aligned} Y_x \perp\!\!\!\perp X|\widehat{H} &\Leftrightarrow p(y|x, \hat{h}) = p(y|x, \hat{h}) \\ &\Leftrightarrow p(y, x, \hat{h}) = p(y|do(X = x), \hat{h}). \end{aligned} \tag{4.40}$$

The first equivalence is the definition of conditional independence 2.8. The second equivalence is because of the fact that $[Y|X = x] = [Y_x|X = x]$ and $[Y_x|\widehat{H}] = [Y|do(X = x), \widehat{H}]$ respectively. □

Proof. (Theorem 4.11)

The proof is in two steps. The first one proves that reconstruction (A1) and joint independence (A2) together imply ignorability. The second step shows that strong IV (A3) and joint independence (A2) together imply positivity.

Ignorability. As in Lemma 4.14, we denote by Y_x the potential outcome. To show ignorability, due to Lemma 4.14 we need to show that $Y_x \perp\!\!\!\perp X|\widehat{H}$.

By the joint independence assumption (A2), we have joint independence $Z \perp\!\!\!\perp (H, \widehat{H})$, which implies, since joint independence implies conditional independence,

$$Z \perp\!\!\!\perp (H, \widehat{H}) \Rightarrow Z \perp\!\!\!\perp H|\widehat{H} = \hat{h} \quad \forall \hat{h} \in supp(\widehat{H}). \tag{4.41}$$

By the reconstruction assumption (A1), we have $X = d(\widehat{H}, Z)$. Hence, given $\widehat{H}$, X is purely a function of Z. Thus, we have:

$$Z \perp\!\!\!\perp H|\widehat{H} \Rightarrow d(\widehat{H}, Z) \perp\!\!\!\perp H|\widehat{H} \Rightarrow X \perp\!\!\!\perp H|\widehat{H}. \tag{4.42}$$

Given the treatment x, the potential outcome Y_x depends only on the confounding H and some noise N_Y, which is independent of all other variables. Therefore, we can write it as $Y_x = m_x(H, N_Y)$ and we have:

$$X \perp\!\!\!\perp H|\widehat{H} \Rightarrow X \perp\!\!\!\perp m_x(H, N_Y)|\widehat{H} \Rightarrow X \perp\!\!\!\perp Y_x|\widehat{H}. \tag{4.43}$$

Due to the symmetry of conditional independence, we have $Y_x \perp\!\!\!\perp X|\widehat{H}$, which shows ignorability.

Positivity. To show positivity, we expand $p(x|\hat{h})$ and show that $p(x|\hat{h}) > 0$ when $p(\hat{h}) > 0$.

By the law of total probability, we have:

$$p(x|\hat{h}) = \int (x, h, z, n_x|\hat{h}) dh dz dn_x. \tag{4.44}$$

We can then rewrite the term inside the integral, using the conditional density formula 2 times: $p(x, y) = p(x, y)p(y)$.

$$p(x, h, z, n_x|\hat{h}) = p(x|h, \hat{h}, z, n_x)p(z|h, \hat{h}, n_x)p(h, n_x|\hat{h}) \tag{4.45}$$

Plugging in the integral gives the following:

$$p(x|\hat{h}) = \int p(x|h,\hat{h},z,n_x)p(z|h,\hat{h},n_x)p(h,n_x|\hat{h})dhdzdn_x. \tag{4.46}$$

Since we can write $X = g(Z,H,N_X)$ and by assumption (A2) $Z \perp\!\!\!\perp (H,\hat{H},N_X)$, we have

$$p(x|\hat{h}) = \int p(x|h,z,n_x)p(z|h,n_x)p(h,n_x|\hat{h})dhdzdn_x. \tag{4.47}$$

We can now write it as:

$$\begin{aligned} p(x|\hat{h}) &= \int \left[\int p(x|h,z,n_x)p(z|h,n_x)dz\right] p(h,n_x|\hat{h})dhdn_x \\ &= \int p(x|h,n_x)p(h,n_x|\hat{h})dhdn_x \end{aligned} \tag{4.48}$$

By assumption (A3), for any compact set $B \subseteq supp(X)$ and for almost every $x \in B$ and for almost every $x \in B$, we have $p(x|h,n_x) \geq c_B$. Thus, we can write

$$\begin{aligned} p(x|\hat{h}) &\geq c_B \int p(h,n_x|\hat{h})dhdn_x \\ &= c_B \int \frac{p(h,n_x,\hat{h})}{p(\hat{h})}dhdn_x = c_B > 0. \end{aligned} \tag{4.49}$$

The last integral integrates to one, if $p(\hat{h}) > 0$. Thus, we have proved positivity.

Finally, we can compute the causal effect, as follows:

$$\mathbb{E}[\mathbb{E}[Y|\widehat{H},X=x]] = \mathbb{E}[\mathbb{E}[Y|\widehat{H},do(X=x)]] = \mathbb{E}[\mathbb{E}[Y|do(X=x)]], \tag{4.50}$$

where we used ignorability in the first equation and positivity in the second. □

This theorem gives an interesting view on the necessary assumptions for CF methods. In particular, it sheds light that the CF is used to control for the treatment, that is, the treatment X becomes independent of Y when conditioned on the CF $\widehat{H}$. All this only by assuming a relation between the treatment X and the CF, but nothing between the CF and the confounder in the second stage.

In opposition to the idea of the CF algorithm that we saw in the beginning of the chapter. The CF was used to control the confounder in the second stage as in Assumption 4, making the new error independent of the treatment.

Chapter 5

Deep Neural Networks Based Methods

A more flexible approach to the nonlinear IV problem can be achieved using Deep Neural Networks (DNN). A first method, proposed by Hartford et al. (2017), has been implemented and had a strong impact on the direction of research in the field. Later, many methods have been proposed to learn a complex causal effect, such as DeepGMM in Bennett et al. (2020), which also had a considerable impact. These methods take advantage of advances in deep learning that come with many advantages. One of them is the universality of Deep Neural Networks in Lu et al. (2017), which is the ability to learn any continuous functions in low, but also in higher-dimensional settings. The advantages of such methods come with some drawbacks, such as more instability in the estimation.

5.1 Deep Instrumental Variables

In the following, we describe the method proposed by Hartford et al. (2017). It uses a slightly more general setting than 2SLS or CF algorithms. We do not make any assumption for the first stage, that is, we allow for nonadditive error. Moreover, in the paper from Hartford et al. (2017), the instrumental variable satisfies Assumptions 1, 2 and 3 that we described at the beginning and assume the following SCM:

$$\begin{aligned} X &= g(Z, W, H, N_X), \\ Y &= f_{\theta_0}(X, W) + h_2(H, N_Y), \end{aligned} \tag{5.1}$$

where the function $f_{\theta_0}(\cdot)$ is from a parameterised family of functions $\{f_\theta(\cdot) : \theta \in \Theta\}$.

We propose to assume, instead of Assumptions 1, 2 and 3, the following assumptions and the SCM 5.1.

Assumption 1*c*. *θ_0 is the unique $\theta \in \Theta$ that satisfies*

$$\mathbb{E}[Y|Z,W] - \mathbb{E}[h_\theta(X,W)|Z,W] = 0 \tag{5.2}$$

almost surely.

Assumption 1*c* is much stronger than Assumption 1, but is necessary for DeepIV. Later in this chapter, we will discuss this assumption in more depth.

A slightly weaker assumption than assumption 2 is given by:

Assumption 2*c*. $\mathbb{E}[h_2(H, N_Y)|Z] = 0$ *a.s.*

The Assumption 2*c* implies that the mean of $h_2(H, N_Y)$ is 0.

The Assumption 3 is already accounted for in the SCM 5.1.

The idea of the method to obtain the causal effect is to first write the conditional expectation of Y given $Z = z$, using the counterfactual function h defined in 3.2.

$$\begin{aligned}\mathbb{E}[Y|Z=z,W=w] &= \mathbb{E}[f_{\theta_0}(x,w)|z,w] + \mathbb{E}[h_2(H,N_Y)|W=w] \\ &= \int h(x,w)F(dx|z,w) \\ &= \int h(x,w)p(x|z,w)dx.\end{aligned} \tag{5.3}$$

The last equation assumes that the measure F has a density p. This is analogue to the 2SLS algorithm, see Remark 5.1 for more details.

To estimate the counterfactual function $h(x,w)$, we can use a two-stage procedure. The first stage estimates the conditional pdf of X given Z and W, denoted $p_\varphi(x|z,w)$. This is done with a mixture density network described in Bishop (2006), which is a DNN where a Gaussian mixture model is applied to the last layer. The second stage solves the inverse problem 5.3 for h with a second neural network h_θ, which is the counterfactual function defined in equation 3.2. We assume that Θ is the set of parameters of the neural network which are the weights of the network. The parameters θ of the second network are optimised using the following loss function:

$$\mathcal{L}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \int h(x,w)p_\varphi(x|z_i,w_i)dx)^2, \tag{5.4}$$

where the sum is taken on the entire data set indexed with i.

Since we are integrating over the density p_φ, it needs to be completely specified in the first stage, which is done with the mixture density network. The loss is used in the following minimisation problem:

$$\arg\min_{\theta\in\Theta} \mathcal{L}(h_\theta). \tag{5.5}$$

Intuitively, the idea of the second stage is to integrate over the entire distribution of X, while keeping Z and W constant. By doing so, we reconstruct X for every value of the confounder H, and therefore, H in Y no longer has any influence on $h(x)$. To think of why it corrects for the confounding, it may help to consider the naive case, where we regress y on x directly. The loss would be:

$$\mathcal{L}'(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i, w_i))^2. \tag{5.6}$$

Here, the confounder H influences both the treatment x_i and the outcome y_i, in contrast to 5.4, where the confounder only influences y_i but not the integral.

We use Monte Carlo integration to obtain an unbiased estimate of the integral in 5.4. The estimate is given by:

$$\int h_\theta(x, w_i) p_\varphi(x|z_i, w_i)dx \approx \frac{1}{B}\sum_b h_\theta(x_b, w_i), \tag{5.7}$$

where $\{x_b\}_1^B \overset{i.i.d.}{\sim} p(\cdot|z_i, w_i)$. More details on the exact implementation can be found in Hartford et al. (2017).

In the loss function 5.4, we use y_i and not $\mathbb{E}[Y|Z = z_i]$ as Equation 5.3 suggests. The following claim shows that both are, in fact, equivalent. For simplicity, here we assume that Z and W are discrete and for notational purposes we ignore W.

Claim 2. *The minimisation problem 5.5 with the loss $\mathcal{L}(h)$ in equation 5.4 gives the same solution as with the loss:*

$$\mathcal{L}'(h) := \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{m_i}\sum_{j:z_i=z_j} y_j - \int h(x)p_\varphi(x|z_i)dx)^2, \tag{5.8}$$

where m_i is defined as the cardinality of the set $\{j : z_i = z_j\}$,
$\mathbb{E}_\varphi[Y|z_i] := \int h(x)p_\varphi(x|z_i)dx$ and $h(x)$ is the counterfactual function.

Since

$$\frac{1}{m_i}\sum_{j:z_i=z_j} y_j \tag{5.9}$$

is a consistent estimation of $\mathbb{E}[Y|Z = z_i]$ and $\int h(x)p_\varphi(x|z_i)dx$ is a consistent estimation of $\mathbb{E}[h(X)|Z = z_i]$, Claim 2 gives the reason why we use y_i and not $\mathbb{E}[Y|Z = z_i]$ in the loss function 5.4.

Proof. With the notation $\mathbb{E}_\varphi[h(X)|z_i] := \int h(x)p_\varphi(x|z_i)dx$, we write the loss in equation 5.4 as:

$$\begin{aligned}
\mathcal{L}(h) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}_\varphi[h(X)|z_i])^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j:z_i=z_j}(y_j - \mathbb{E}_\varphi[h(X)|z_i])^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j:z_i=z_j}(y_j^2 + \mathbb{E}_\varphi[h(X)|z_i]^2 - 2y_j\mathbb{E}_\varphi[h(X)|z_i]) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{m_i}\sum_{j:z_i=z_j}y_j^2 + \mathbb{E}_\varphi[h(X)|z_i]^2 \\
&\quad - 2(\frac{1}{m_i}\sum_{j:z_i=z_j}y_j)\mathbb{E}_\varphi[h(X)|z_i]).
\end{aligned} \tag{5.10}$$

Then, we can write $\mathcal{L}'(h)$ as:

$$\begin{aligned}
\mathcal{L}'(h) &= \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{m_i}\sum_{j:z_i=z_j}y_j - \mathbb{E}_\varphi[h(X)|z_i])^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((\frac{1}{m_i}\sum_{j:z_i=z_j}y_j)^2 + \mathbb{E}_\varphi[h(X)|z_i]^2 \\
&\quad - 2(\frac{1}{m_i}\sum_{j:z_i=z_j}y_j[Y|z_i])\mathbb{E}_\varphi[h(X)|z_i]).
\end{aligned} \tag{5.11}$$

Thus, by comparing the term inside the sum in equations 5.10 and 5.11, we see that only the first term differs. However, since, in the minimisation problem 5.5, we minimise over h and the first term does not depend on h we can write:

$$\mathcal{L}'(h) = \mathcal{L}(h) + C, \tag{5.12}$$

where C is equal to:

$$C := \frac{1}{m_i}\sum_{j:z_i=z_j}y_j^2 - (\frac{1}{m_i}\sum_{j:z_i=z_j}y_j)^2 = \widehat{var}(Y|z_i) < \infty, \tag{5.13}$$

where $\widehat{var}(Y|z_i)$ denotes the conditional sample variance of Y given z_i and we assume that $\widehat{var}(Y|z_i) < \infty$. C is a constant in h and therefore the h we obtain in the minimisation problem stays the same. □

Remark 5.1. *(DeepIV as a generalisation of 2SLS)*

We can see that DeepIV is a generalisation of 2SLS. To see this, let $h(x,w) = \beta_0 + \beta_1 f_1(x) + \cdots + \beta_k f_k(x) + \gamma w + \mathbb{E}[h_2(H)|W = w]$ *as in Section 3.3. The last integral in equation 5.3 is:*

$$\int h(x,w)p(x|z,w)dx = \beta_0 + \sum_{i=1}^{k} \beta_i \mathbb{E}[f_i(x)|z,w] + \gamma w + \mathbb{E}[h_2(H)|w], \quad (5.14)$$

where $p(x|z,w)$ *denote the conditional density of* X *given* Z, W*.*

In 2sls, $\mathbb{E}[f_i(x)|z,w]$ *corresponds to the predicted values* $\widehat{f_i(X)}$ *calculated in the first stage. In the second stage in 2SLS, we perform the regression:*

$$Y \sim \widehat{f_1(X)} + \cdots + \widehat{f_k(X)} + W \quad (5.15)$$

and get the coefficient with OLS, which is analog to DeepIV.

5.1.1 Fredholm Integral Equations

Next, we use functional analysis to identify the limitations of this method. The problem 5.5 can be seen as a Fredholm integral equations of the first kind, such as in KRESS (1989), which can be written as,

$$g(z) = \int h(x)p(x,z)dx, \quad (5.16)$$

where h is an unknown function and g, q are known functions. The goal is then to recover h.

In the context of initial and boundary value problems for partial differential equations, Hadamard postulated three conditions for a problem to be well-posed in Hadamard (1923).

1. Existence of a solution.
2. Uniqueness of the solution.
3. Continuous dependence of the solution on the data.

The third postulate ensures that small errors in the data result in only small errors in the solution. Any violation in one of the three postulates makes the problem ill-posed.

According to the book by KRESS (1989), the problem 5.16 is well-posed if the operator A given by $Ah := \int h(x)p(x,z)dx$ is bijective and the inverse A^{-1} is continuous.

Example 5.2. *For example, if* $p(x,z) = c$ *and* $g(z) = d$ *are both constant functions, we simplify the problem in 5.16 as* $d = c\int h(x)dx$ *and* $h(x)$ *can be any function that has an integral equal to* $\frac{d}{c}$*. Hence, in this case the problem is ill-posed.*

5.1.2 Stability of DeepIV

The first postulate in Section 5.1.1 is satisfied by assuming that the data are generated according to the SCM 5.1 and that there is a $\theta \in \Theta$, such that $f = f_\theta$, where f is the f in SCM 5.1. The second postulate is satisfied by the following assumption 1c.

The Assumption 1c helps to satisfy the second postulate, the uniqueness of solution. As we shall see in Section 6 and in Example 5.4, if the Assumption 1c almost breaks, then DeepIV might be biased.

The third postulate might not be satisfied. As shown later in Example 5.4, if the Assumption 1c is close to not being satisfied, and more generally, by solving the minimisation problem 5.5, the loss function might get stuck in a local minimum. Another consideration for the third postulate to be satisfied is that the weights of the neural networks are randomly initialised, which can play a role if the sample size is too small.

Intuitively, the Assumption 1c is an analogous condition to the Assumption 1b but with independence instead of uncorrelation since it used linear regression. We can see this with the following claim.

Claim 3. *For the setting in Section 3.3 and the 2SLS in Algorithm 2, Assumption* 1c *holds if and only if Assumption* 1b *holds.*

Proof. We first define a matrix $A \in \mathbb{R}^{k \times k+1}$ and a matrix $g(Z) \in \mathbb{R}^{k+1 \times n}$.

$$A = \begin{pmatrix} 1 & \hat{\alpha}_1^1 & \dots & \hat{\alpha}_k^1 \\ 1 & \ddots & & \vdots \\ \vdots & & & \\ 1 & \dots & & \hat{\alpha}_k^k \end{pmatrix}, g(Z) = \begin{pmatrix} 1 & \dots & 1 \\ g_1(Z^1) & \dots & g_1(Z^n) \\ \vdots & & \\ g_k(Z^1) & \dots & g_k(Z^n) \end{pmatrix}. \tag{5.17}$$

We then assume that there exists (at least) one $\beta = (\beta_0, \dots, \beta_k)$ such that

$$\begin{aligned} \mathbb{E}[Y|Z] &= \beta_0 + \sum_{i=1}^{k} \beta_i \mathbb{E}[f_i(X)|Z] \\ &= \beta_0 + \sum_{i=1}^{k} \beta_i \sum_{j=1}^{k} \alpha_i^j g_j(Z) \\ &= \beta^T A g(Z). \end{aligned} \tag{5.18}$$

We can now write:

$$Y = \beta^T A g(Z) + N_Y, \tag{5.19}$$

where N_Y is an independent noise.

By equation 5.18, we find that Assumption 1*c* holds if and only if there is a unique β that solves the regression 5.20.

We can then use the known result that OLS has a unique solution if and only if the matrix XX^T has full rank, where X is the predictor matrix. This can be applied to the following regression:

$$Y \sim Ag(Z). \tag{5.20}$$

If $g(Z)$ has rank $k+1$, then, according to a known property of the rank, the rank of the matrix $Ag(Z)^T$ is:

$$rank(Ag(Z)) = rank(A). \tag{5.21}$$

We find that the regression has a unique solution β if and only if the matrix A has full rank. The Claim 1 implies that the matrix A has full rank if and only if Assumption 1*b* holds.

□

We show now that Assumption 1*c* is stronger than Assumption 1.

Claim 4. *Assuming Assumption* 2*c, Assumption* 1*c implies Assumption 1.*

Proof. We show the claim by contraposition, that is, if $p(x|z,w)$ is constant in z, then there are $\theta_0, \theta_1 \in \Theta$ such that $\mathbb{E}[Y|Z,W] - \mathbb{E}[h_{\theta_i}(X)|Z,W] = 0$ a.s. for $i = 1, 2$.

First, $p(x|z,w) = p(x|w)$ implies:

$$\begin{aligned} \mathbb{E}[Y|Z,W] &= \mathbb{E}[f_{\theta_0}(X,W) + h_2(X,N_Y)|Z,W] \\ &= \mathbb{E}[f_{\theta_0}(X,W)|Z,W] + \mathbb{E}[h_2(X,N_Y)|W] \\ &= \mathbb{E}[f_{\theta_0}(X,W)|W] + \mathbb{E}[h_2(X,N_Y)|W] \\ &= \mathbb{E}[Y|W]. \end{aligned} \tag{5.22}$$

Analogously, $p(x|z,w) = p(x|w)$ implies, for every $\theta \in \Theta$:

$$\mathbb{E}[h_\theta(X,W)|Z,W] = \mathbb{E}[h_\theta(X,W)|W]. \tag{5.23}$$

Then, we define $h_{\theta_1}(X,W) = \mathbb{E}[Y|W]$ as a function of W which is constant in X and we have:

$$\mathbb{E}[h_{\theta_1}(X,W)|Z,W] = \mathbb{E}[Y|W]. \tag{5.24}$$

Thus, it holds:

$$\mathbb{E}[Y|Z,W] - \mathbb{E}[h_{\theta_1}(X)|Z,W] = 0. \tag{5.25}$$

Together with $h_{\theta_0}(X,W) := f_{\theta_0}(X,W) + \mathbb{E}[h_2(X,N_Y)|W]$ as in SCM 5.1, we have found $\theta_0, \theta_1 \in \Theta$ such that $\mathbb{E}[Y|Z,W] - \mathbb{E}[h_{\theta_i}(X)|Z,W] = 0$ a.s. for $i = 0, 1$, which ends the proof. □

Example 5.3. *The second assumption is not implied by the first one. To see this, consider the following example.*

$$\begin{aligned} X &= H + sgn(Z) + N_X, \\ Y &= X^2 + H + N_Y, \end{aligned} \tag{5.26}$$

where $sgn(Z)$ *is the sign function* $sgn(Z) := \begin{cases} -1, & \text{if } Z \leqslant 0, \\ 1, & \text{if } Z > 0 \end{cases}$ *and* $H, Z, N_x, N_y \overset{i.i.d.}{\sim} \mathcal{U}[-1, 1]$.

We clearly see that X *is dependent on* Z*, because we gain information on* X *knowing* Z*. However, we have*

$$\begin{aligned} \mathbb{E}[Y|Z] &= \mathbb{E}[X^2|Z] \\ &= \mathbb{E}[H^2 + sgn(Z)^2 + N_x^2 + 2Hsgn(Z) + 2HN_x + 2sgn(Z)N_x|Z] \\ &= \mathbb{E}[H^2] + 1 + \mathbb{E}[N_X^2]. \end{aligned} \tag{5.27}$$

Thus, $\mathbb{E}[Y|Z]$ *does not depend on* Z*.*

To see why Assumption 1c is necessary, let us see an example.

Assume that $\mathbb{E}[Y|Z] = c$, with a constant $c \in \mathbb{R}$. We can write equation 5.4 as:

$$c = \int h(x)p(x|z)dx. \tag{5.28}$$

Since $\int p(x|z)dx = 1$, h can be equal to c or $c + g(x)$ where g can be any function orthogonal to $p(x|z)$, that is, when the integral of the product is 0. Thus, the problem has many solutions and is therefore ill-posed.

Example 5.4. *In practice, if the conditional expectation of* Y *given* Z *is almost constant everywhere, then it does not estimate* h *well. An example is given by the SCM:*

$$\begin{aligned} X &= Z + H, \\ Y &= cos(X) + H, \end{aligned} \tag{5.29}$$

where $H \sim \mathcal{U}[-\alpha, \alpha]$ *and* $Z \sim \mathcal{U}[-5, 5]$.

The conditional expectation of Y *given* Z *is given by*

$$\mathbb{E}[Y|Z] = \frac{1}{2\alpha}(sin(Z + \alpha) - sin(Z - \alpha)). \tag{5.30}$$

This is not constant, however, in practice if α is larger than approximately 1.5 it makes the conditional expectation too close to 0 and therefore the problem becomes ill-posed.

In the following plot, α takes the value of $1, 4$ and 7 respectively. As we can see, giving the theoretical conditional mean of Y given Z improves the performance of DeepIV, but the performance still decreases as we increase α and therefore the conditional mean is closer to being constant 0. In CF implemented with splines, the problem does not depend on the value of α. In this case, it suggests that the problem with DeepIV is not due to the first stage, but rather to the second stage, where we take Y instead of $\mathbb{E}[Y|Z]$.

More importantly, it suggests that indeed when the conditional of Y given Z is almost constant, then a 2SLS type method is not appropriate and CF methods handle this case without problem. We can see in Figure 5.1 that the DeepIV estimator still captures the linear causal effect but not the exact details of the cosine function, that is, the naive estimator that regresses only Y on X would have a positive slope.

We will see in Remark 5.8 that the DeepGMM algorithm has the same problem as DeepIV and is also inconsistent for this example.

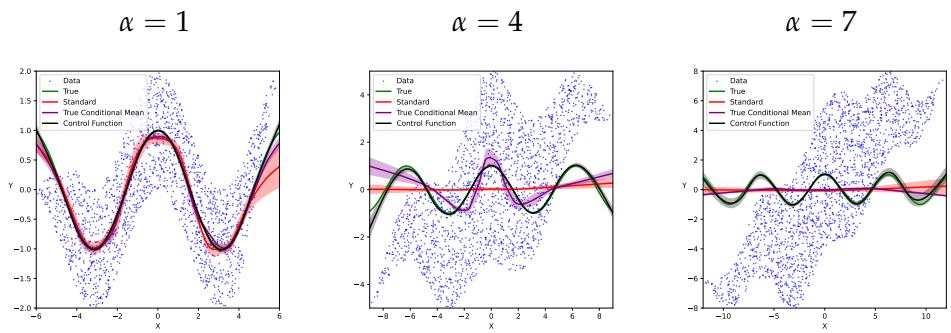


Figure 5.1: Plots of DeepIV, that is implemented for the SCM 5.29.

5.2 Deep Control Function

A generalisation of the control function algorithm is to use two neural networks instead of splines, one in the first stage to learn the residuals and one in the second stage, which allows us to weaken the Assumption 5 and 4. It is, however, not as general as in the general control function Section 4.3 since we need that the confounder enters the first stage additively. The idea of using nonparametric regression for the first and second stages of the control function algorithm can be traced back to Blundell and Powell (2004).

We assume the following SCM:

$$\begin{aligned} X &= g(Z, W) + h_1(H, N_X), \\ Y &= f(X, W, H) + N_Y. \end{aligned} \tag{5.31}$$

More precisely, the algorithm is implemented in Algorithm 5.

Algorithm 5 Deep Control Function algorithm

Stage 1
Fit a neural network for X on Z, W: $X \sim Z, W$.
Obtain the residuals *res*.
Stage 2
Fit a model for Y on X, W, res: $Y \sim X, W, res$.
Compute the counterfactual function: $\mathbb{E}_{res}[\mathbb{E}[Y|X = x, W = w, res]]$.
Smooth the counterfactual function using smoothing splines.

Since we do not have additive errors and the confounder in the outcome variable Y, we calculate a generalised version of the counterfactual function $h(x, w) = f(x, w) + \mathbb{E}[h_2(H)|W = w]$ defined in the equation 3.2. The counterfactual function, in this case is:

$$\begin{aligned} h(x, w) &= \mathbb{E}[Y|do(X = x), W = w] \\ &= \mathbb{E}_{res}[\mathbb{E}[Y|X = x, W = w, res]], \end{aligned} \tag{5.32}$$

where we will prove in the next subsection that the last equality holds true.

In practice, to compute the counterfactual function $\mathbb{E}_{res}[\mathbb{E}[Y|X = x, W = w, res]]$. We take m samples of size n of the residuals computed in the first stage, denoted by $B_1, \ldots, B_m$. Then, we apply the model of the second stage to X_{test}, W_{test} and B_i and obtain the predicted outcome $\widehat{Y}_i$. Repeating this for $i = 1 \ldots, m$ gives $\widehat{Y}_1, \ldots, \widehat{Y}_m$. Taking the mean gives the estimated counterfactual function $\hat{h}(X_{test}, W_{test})$:

$$\hat{h}(X_{test}, W_{test}) = \frac{1}{m} \sum_{i=1}^{m} \widehat{Y}_i. \tag{5.33}$$

To have a smooth estimate, we can increase m. We also use a smoothing spline that fits $\hat{h}(X_{test}, W_{test})$ on X_{test}, W_{test}. We select the smoothing parameter through cross-validation.

5.2.1 Identification

Identification of the causal effect using DeepCF can be shown using Theorem 4.11 with the residuals of the first stage as a control function. To apply the

result, we assume the SCM 5.31, Assumption (A3) in Theorem 4.11 and we assume conditional independence $Z \perp\!\!\!\perp H|W$ (Assumption 2).

Following Puli and Ranganath (2021), we show that if the SCM 4.37 holds, then we can relax the Assumption (A2) in Theorem 4.11 by only assuming marginal independence, that is, $Z \perp\!\!\!\perp H$, $Z \perp\!\!\!\perp \widehat{H}$. For simplicity, we ignore the observed confounder W.

We first state and proof lemma 1 in Puli and Ranganath (2021) that will be useful for applying the theorem 4.11 to the DeepCF algorithm.

Lemma 5.5. *Assuming a treatment of the form $X = g(Z, h_1(H, N_X))$ and assuming that the joint independence $(\widehat{H}, h_1(H, N_X)) \perp\!\!\!\perp Z$ holds. Then, if $\widehat{H} = e(X, Z)$ for a function e, the joint independence $(\widehat{H}, H, N_X) \perp\!\!\!\perp Z$ holds.*

We recall here that $p(a|b)$ denotes the density of the random variable $A = a$ given $B = b$. The random variable is always in uppercase and the value it takes is denoted in lowercase.

Proof. We first show $p(\hat{h}|h, z, n_x) = p(\hat{h}|h_1(h, n_x))$. We first use the conditional density formula:

$$p(\hat{h}, x|h, z, n_x) = p(\hat{h}|x, h, z, n_x)p(x|h, z, n_x). \tag{5.34}$$

By integrating over X we have:

$$p(\hat{h}|h, z, n_x) = \int p(\hat{h}|h, z, x, n_x)p(x|h, z, n_x)dx. \tag{5.35}$$

Since we can write $\widehat{H} = e(X, Z)$, we have the conditional independence $\widehat{H} \perp\!\!\!\perp (H, N_X)|Z, X$.

$$p(\hat{h}|h, z, n_x) = \int p(\hat{h}|z, x)p(x|h, z, n_x)dx. \tag{5.36}$$

Since $X = g(z, h_1(H, N_X))$, we have:

$$p(\hat{h}|h, z, n_x) = \int p(\hat{h}|z, x)p(x|z, h_1(h, n_x))dx. \tag{5.37}$$

Since we can write $\widehat{H} = e(X, Z)$, we also have the conditional independence $\widehat{H} \perp\!\!\!\perp h_1(H, N_X)|Z, X$.

$$p(\hat{h}|h, z, n_x) = \int p(\hat{h}|z, x, h_1(h, n_x))p(x|z, h_1(h, n_x))dx. \tag{5.38}$$

By using the conditional independence formula and integrating out, we have:

$$p(\hat{h}|h, z, n_x) = p(\hat{h}|z, h_1(h, n_x)) = p(\hat{h}|h_1(h, n_x)), \tag{5.39}$$

where we used the joint independence $(\widehat{H}, h_1(H, N_X)) \perp\!\!\!\perp Z$ in the last equality.

Now, we integrate both sides of Equation 5.39 with respect to $p(z|h, n_x)$ gives:

$$\begin{aligned}\int p(\hat{h}|h_1(h, n_x))p(z|h, n_x)dz &= \int p(\hat{h}|h, z, n_x)p(z|h, n_x)dz \\ &= p(\hat{h}|h, n_x),\end{aligned} \tag{5.40}$$

where we again used the conditional density formula and integrating out in the last equality. Since $p(\hat{h}|h_1(h, n_x))$ does not depend on Z, integrating the LHS of equation 5.40 gives:

$$\begin{aligned}\int p(\hat{h}|h_1(h, n_x))p(z|h, n_x)dz &= p(\hat{h}|h_1(h, n_x)) \Rightarrow p(\hat{h}|h_1(h, n_x)) \\ &= p(\hat{h}|h, n_x)\end{aligned} \tag{5.41}$$

Thus, we have:

$$p(\hat{h}|h, z, n_x) = p(\hat{h}|h_1(h, n_x)) = p(\hat{h}|h, n_x), \tag{5.42}$$

which implies the joint independence $(\widehat{H}, H, N_X) \perp\!\!\!\perp Z$. □

Then, we have the following claim, which has been proven in Puli and Ranganath (2021).

Proposition 5.6. *Assuming SCM 5.31, given the marginal independences $Z \perp\!\!\!\perp H$ and $Z \perp\!\!\!\perp \widehat{H}$, we have the joint independence $Z \perp\!\!\!\perp (\widehat{H}, H)$.*

Proof. Given the SCM 5.31, we can rewrite the random variables $X - \mathbb{E}[X|Z]$ as follows:

$$\begin{aligned}X - \mathbb{E}[X|Z] &= g(Z) + h_1(H, N_X) - \mathbb{E}[g(Z) + h_1(H, N_X)|Z] \\ &= h_1(H, N_X) - \mathbb{E}[h_1(H, N_X)].\end{aligned} \tag{5.43}$$

Next, due to the first stage of DeepCF, we can write the treatment X as the sum of the residuals $\widehat{H}$ and a function $g'(Z)$:

$$X = g'(Z) + \widehat{H}. \tag{5.44}$$

We show that $\widehat{H}$ determines $h_1(H, N_X)$:

$$\widehat{H} - \mathbb{E}[\widehat{H}] = X - \mathbb{E}[X|Z] = h_1(H, N_X) - \mathbb{E}[h_1(H, N_X)], \tag{5.45}$$

where we used Equation 5.44 in the first equality and SCM 5.31 in the second. We can then write:

$$\widehat{H} = h_1(H, N_X) + c, \tag{5.46}$$

where c is a constant. By the independence, $\widehat{H} \perp\!\!\!\perp Z$, we have:

$$p(\hat{h}, h_1(h, n_x)|z) = p(\hat{h}, \hat{h} - c|z) = p(\hat{h}, \hat{h} - c) = p(\hat{h}, h_1(h, n_x)). \tag{5.47}$$

Thus, we have $(\widehat{H}, h_1(H, N_x)) \perp\!\!\!\perp Z$ and therefore lemma 5.5 implies the joint independence $(\widehat{H}, H, N_x) \perp\!\!\!\perp Z$.

□

Therefore, if the first stage of DeepCF is consistently estimated, then we have $Z \perp\!\!\!\perp \widehat{H}$, where the control function $\widehat{H}$ is the residuals $h_1(H, N_X)$. Thus, we have that Z is independent of $\widehat{H}$ and thus the joint independence $Z \perp\!\!\!\perp (\widehat{H}, H)$ holds and we can apply Theorem 4.11 and this proves that DeepCF estimates the counterfactual function:

$$\mathbb{E}_{\widehat{H}}[\mathbb{E}[Y|X = x, \widehat{H}]] = \mathbb{E}[Y|do(X = x)], \tag{5.48}$$

where $\mathbb{E}_{\widehat{H}}[\cdot]$ denotes the expectation over $\widehat{H}$.

5.3 General Control Function Algorithm

In this section, we describe the algorithm that has been given by Puli and Ranganath (2021). The algorithm is constructed in such a way that it meets the requirements of Theorem 4.11.

The first step is to construct a control function $\widehat{H}$ that satisfies the reconstruction property for a function d:

$$X = d(Z, \widehat{H}), \tag{5.49}$$

and joint independence:

$$Z \perp\!\!\!\perp (H, \widehat{H}). \tag{5.50}$$

The idea is to take an autoencoder consisting of two neural networks. One is called an encoder, with parameters $\theta \in \Theta$ and the second one is a decoder with parameters $\varphi \in \Phi$.

The encoder takes Z and X as input and output one variable that aims to be the control function $\widehat{H}$, we write $e_\theta(x_i, z_i) = \hat{h}_\theta^i$. This network generates a sample of the control function $(\hat{h}_\theta^1, \ldots, \hat{h}_\theta^n)$.

The decoder takes $\widehat{H}_\theta$ and Z as input and output X, with the aim of reproducing the reconstruction function d. We denote the decoder by $d_\varphi(\hat{h}, z)$.

To guarantee joint independence, we only impose independence between $\widehat{H}_\theta$ and Z and assume that the reconstruction function d has a form that guarantees joint independence only by assuming independence. The reconstruction

function d guarantees joint independence, for example, with an additive treatment process $d(\widehat{H}_\theta, Z) = f'(\widehat{H}_\theta) + g'(Z)$ as proven in Proposition 5.6. To have independence between $\widehat{H}_\theta$ and Z, we impose that the mutual information I between $\widehat{H}_\theta$ and Z is 0. We define the following constrained maximisation problem called Variational Decoupling (VDE):

$$\max_{\theta\in\Theta,\varphi\in\Phi} \mathbb{E}_{(X,Z)}[\mathbb{E}_{\widehat{H}_\theta|X=x,Z=z}[log(p_\varphi(x|\hat{h}_\theta, z))]] \text{ s.t. } I(\widehat{H}_\theta, Z) = 0, \tag{5.51}$$

where $p_\varphi(x|\hat{h}_\theta, z)$ denotes the density of $x = d_\varphi(\hat{h}_\theta, z) + N'_X$ for a noise N'_X.

The maximisation problem can be turned into a Lagrange multiplier maximisation problem with $\lambda > 0$:

$$\max_{\theta\in\Theta,\varphi\in\Phi} \mathbb{E}_{(X,Z)}[\mathbb{E}_{\widehat{H}_\theta|X=x,Z=z}[log(p_\varphi(x|\hat{h}_\theta, z))]] - \lambda I(\widehat{H}_\theta, Z). \tag{5.52}$$

In Puli and Ranganath (2021), it is proven that the maximisation problem is equivalent to the following tractable maximisation problem.

$$\max_{\theta,\varphi,\nu} \mathbb{E}_{(X,Z)}[(1+\lambda)\mathbb{E}_{\widehat{H}_\theta|X=x,Z=z}[log(p_\varphi(x|\hat{h}_\theta, z)) - \lambda \mathbf{KL}(p(\hat{h}_\theta|x, z)||p_\nu(\hat{h}))]], \tag{5.53}$$

where $\mathbf{KL}(P||Q)$ denotes the Kullback-Leibler divergence and $p_\nu(\hat{h})$ denotes an auxiliary distribution parameterised by ν. The auxiliary distribution should be as close as possible to $p(\hat{h})$. The reason we take $p_\nu(\hat{h})$ and not directly $p(\hat{h})$ is that we only have a sample of $\widehat{H}|X, Z$ due to the construction of the control function $\widehat{H}$.

The second stage is given as a maximum-likelihood problem, where the parameters β of a neural network, denoted $f_\beta(\hat{h}, x)$, are learned under the true data distribution $p(y, x, z)$ and the control function distribution $p(\hat{h}|x, z)$:

$$\arg\max_\beta \mathbb{E}_{(Y,X,Z)}[\mathbb{E}_{\widehat{H}|X=x,Z=z}[log(p_\beta(y|\hat{h}, x)]], \tag{5.54}$$

where $p_\beta(y|\hat{h}, x)$ denotes the density of $y = f_\beta(\hat{h}, x) + N'_Y$ for a noise N'_Y.

In summary, we can express the GCFN algorithm as in Algorithm 6.

5.4 Deep Generalised Method of Moments

A different approach from the two-stage Least Squares and control function, which is currently seen as one of the best methods to learn complex causal effects, is called the Deep Generalised Method of Moment (DeepGMM) from Bennett et al. (2020).

Algorithm 6 General Control Function algorithm

Stage 1

Fit an autoencoder, with encoder $\widehat{H} \sim X, Z$ and decoder $X, Z \sim \widehat{H}$.

From the encoding function, obtain a sample of the control function given treatment and IV : $\widehat{H}|X, Z$.

Stage 2

Fit a neural network for Y on $X, \widehat{H}$: $Y \sim X, \widehat{H}$.

Compute the counterfactual function: $\mathbb{E}_{\widehat{H}}[\mathbb{E}[Y|X = x, \widehat{H}]]$.

It is based, as the name suggests, on the Generalised Method of Moments (GMM). As we will later show, it uses the same kind of assumptions as 2SLS and DeepIV.

In this section, we assume exactly the same as in Section 5.1 and Assumptions $1d$, which is similar as Assumption $1d$ as it will be shown later.

We assume the Assumption $2c$ instead of Assumption 2.

Given functions $g_1, \dots, g_m$, it uses the fact that $\mathbb{E}[g_j(Z)h_2(H, N_Y)] = 0$, giving us the condition,

$$\psi(g_1, \theta) = \cdots = \psi(g_m, \theta) = 0, \quad \text{where } \psi(g, \theta) = \mathbb{E}[g(Z)(Y - f_\theta(X))]. \tag{5.55}$$

For identification purposes we assume the following assumption that can be found in assumption 1 in Bennett et al. (2020):

Assumption $1d$. *θ_0 is the unique $\theta \in \Theta$ satisfying $\psi(g, \theta) = 0$ for all $g \in \mathcal{G}$.*

Assumption $1d$ restricts the shape of f_θ. We will see in Section 5.4.1 that this assumption is equivalent to Assumption $1c$.

The goal is to find θ:

$$\hat{\theta}^{GMM} \in \arg\min_{\theta \in \Theta} ||(\psi_n(g_1, \theta), \dots, \psi_n(g_m, \theta))||^2, \tag{5.56}$$

where $\psi_n(g, \theta) := \frac{1}{n}\sum_{i=1}^n g(Z_i)(Y_i - f_\theta(X_i))$ is the empirical counterpart of $\psi(g, \theta)$.

If the number of moments becomes too large, the estimator $\hat{\theta}^{GMM}$ can be inefficient since all moments influence the minimisation problem the same and, therefore, it can give too much importance to a moment that is not important. For example, if two moments are the same. This can be improved using another norm than the 2-norm $||\cdot||$, namely a norm that weight the moments $g_1, \dots, g_m$ by the inverse of the estimated covariance, for any

consistent estimate $\tilde{\theta}$ with θ_0:

$$||v||_{\tilde{\theta}}^2 = v^T C_{\tilde{\theta}}^{-1} v, \text{ where } [C_{\tilde{\theta}}]_{jk} = \frac{1}{n}\sum_{i=1}^{n} g_j(Z_i)g_k(Z_i)(Y_i - f_\theta(X_i))^2. \quad (5.57)$$

In Hansen (1982), it has been shown that using the norm $||\cdot||_{\tilde{\theta}}$ instead of the 2-norm leads to minimal asymptotic variance.

In the paper of Bennett et al. (2020), an equivalent reformulation of problem 5.56 is given by Lemma 5.7, which uses a variational approach.

Lemma 5.7. *Let $||v||_{\tilde{\theta}}$ be the optimally weighted norm as in equation 5.57 and let $\mathcal{G} = span(g_1, \ldots, g_m)$. Then*

$$||(\psi_n(g_1,\theta), \ldots, \psi_n(g_m,\theta))||_{\tilde{\theta}}^2 = \sup_{g\in\mathcal{G}} \psi_n(g,\theta) - \frac{1}{4n}\sum_{i=1}^{n} g^2(Z_i)(Y_i - f_{\tilde{\theta}}(X_i))^2. \quad (5.58)$$

This allows us to rewrite the problem 5.56 in a more general form, allowing $\mathcal{G}$ to be a more flexible set than $span(g_1, \ldots, g_m)$. We define
$\mathcal{G} := \{g_\tau(z) : \tau \in \mathcal{T}\}$ as the set of all neural networks given a specific architecture with varying weights τ. Similarly, we define $\mathcal{F} := \{f_\theta(x) : \theta \in \Theta\}$ a neural network with varying weights θ.

$$\hat{\theta}^{DeepGMM} = \arg\min_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} U_{\tilde{\theta}}(\theta,\tau),$$
$$\text{where } U_{\tilde{\theta}}(\theta,\tau) = \frac{1}{n}\sum_{i=1}^{n} g_\tau(Z_i)(Y_i - f_\theta(X_i)) - \frac{1}{4n}\sum_{i=1}^{n} g_\tau^2(Z_i)(Y_i - f_{\tilde{\theta}}(X_i))^2, \quad (5.59)$$

where g_τ and f_θ are both approximated with neural networks with $\tau \in \mathcal{T}$ and $\theta \in \Theta$ as parameters for the networks. The technical details on the optimisation can be found in Bennett et al. (2020).

The consistency of $\hat{\theta}^{DeepGMM}$ assuming Assumption 1*d*, 2*c* and SCM 5.1, as well as 4 other technical assumptions on $\mathcal{F}$ and $\mathcal{G}$ has been proven in Bennett et al. (2020).

5.4.1 Similarities between DeepIV and DeepGMM

In this section, we show a similarity in the assumptions between DeepIV and DeepGMM, which suggests that if DeepIV converge to a wrong causal effect, then DeepGMM might also converge to a wrong causal effect and vice versa.

Claim 5. *Assumption* 1*d is satisfied if and only if Assumption* 1*c is satisfied.*

Proof. We can write $\psi(g, \theta)$ as follows:

$$\begin{aligned}\psi(g,\theta) &= \mathbb{E}[g(Z)(Y - f_\theta(X))] \\ &= \mathbb{E}[g(Z)\mathbb{E}[Y - f_\theta(X)|Z]] \\ &= \mathbb{E}[g(Z)(\mathbb{E}[Y|Z] - \mathbb{E}[f_\theta(X)|Z])],\end{aligned} \tag{5.60}$$

where we used the tower property of the conditional expectation in the second equality.

Assumption 1*d* ⇒ **Assumption** 1*c*: We prove it by contraposition. Assume that there exist two unequal θ_1, θ_2, such that $\mathbb{E}[Y|Z] - \mathbb{E}[f_{\theta_i}(X)|Z] = 0$ a.s. for $i = 1, 2$. Then for all functions g, we have $\psi(g, \theta_1) = \psi(g, \theta_2) = 0$ by Equation 5.60.

Assumption 1*c* ⇒ **Assumption** 1*d*: We prove this again by contraposition. Assume that there exist two unequal θ_1, θ_2, such that $\psi(g, \theta_i) = 0$ a.s. for $i = 1, 2$ and for all functions g. Let $i \in \{1, 2\}$ be arbitrary, we can write $\mathbb{E}[Y|Z] - \mathbb{E}[f_{\theta_i}(X)|Z]$ as a function of Z, say $f_i(Z)$. Then, for all functions g we have:

$$\mathbb{E}[g(Z)f_i(Z)] = 0. \tag{5.61}$$

Thus, if we let $g(Z) = f_i(Z)$, then we have:

$$\mathbb{E}[g(Z)f_i(Z)] = \mathbb{E}[f_i(Z)^2], \tag{5.62}$$

which is equal to 0 if and only if $f_i(Z)$ is equal to 0 a.s. Hence, we have $\mathbb{E}[Y|Z] - \mathbb{E}[f_{\theta_i}(X)|Z] = 0$ a.s. which ends the proof.

□

With Claims 2 and 5, we can show that DeepIV has more than one solution if and only if DeepGMM has more than one solution. Therefore DeepGMM is as ill-posed as DeepIV even though DeepGMM might converge faster or might be more robust in some cases. The reasoning is as follows:

Let us assume that the minimisation problem 5.5 for DeepIV has two solutions h_{θ_0} and h_{θ_1}. This is equivalent to $\mathcal{L}(h_{\theta_0}) = \mathcal{L}(h_{\theta_1})$, where $\mathcal{L}(h)$ is the loss function of DeepIV defined in equation 5.4.

By Claim 2 this is equivalent to $\mathcal{L}'(h_{\theta_0}) = \mathcal{L}'(h_{\theta_1})$.

This is in turn equivalent to the fact that Assumption 1*c* is not satisfied.

Lastly, by Claim 5 this is equivalent to the Assumption 1*d* not being satisfied, which might lead to inconsistencies in DeepGMM.

Remark 5.8. *Following Example 5.4, the GMM estimator $\hat{\theta}^{GMM}$ is also ill-posed when the conditional mean of Y given Z is constant or almost constant, that is,*

$\mathbb{E}[Y|Z] = \mathbb{E}[Y]$. *The reason can be seen as follows for all functions g:*

$$\begin{aligned}\mathbb{E}[g(Z)(Y - f_\theta(X))] &= \mathbb{E}[g(Z)\mathbb{E}[Y - f_\theta(X)|Z]] \\ &= \mathbb{E}[g(Z)(\mathbb{E}[Y|Z] - \mathbb{E}[f_\theta(X)|Z])] \\ &= \mathbb{E}[g(Z)(\mathbb{E}[Y] - \mathbb{E}[f_\theta(X)|Z])].\end{aligned} \tag{5.63}$$

This is equal to 0 *if* $\mathbb{E}[f_\theta(X)|Z] = \mathbb{E}[Y]$*, which is satisfied if, for example, we have* $f_\theta(x) = \mathbb{E}[Y]$ *for all x regardless of the true* f_{θ_0}*. Thus, we have that* $\psi_n(g, \theta) = 0$ *for all g for more than one value of* θ*. Therefore, there are more than one minimisers for the minimisation problem 5.56. The GMM estimator* $\hat{\theta}^{GMM}$ *is therefore not unique.*

This also holds for the norm $||\cdot||_{\hat{\theta}}$ *and by lemma 5.7, the DeepGMM estimator is therefore also not unique.*

The DeepGMM estimators will most likely not be consistent or would have difficulties to converge, for settings, such as, in Example 5.3, 5.4 and Example 5.3. Simulations for example 5.4 are shown in Section 6.

Chapter 6

Simulations

In this chapter, we conduct a simulation study on the algorithms we previously described. The goal is to get a sense of which method performs well in which setting and to see what it does if some assumptions are broken. We also consider cases where both the treatment and the instrument have more than one dimension and how the estimators increase in accuracy when we increase the number of observations.

6.1 Implementation of the Algorithms

In this section, we give some details on the implementation of each method. The implementation of all the codes used for the plots in this thesis can be found on my GitHub [1].

- **Naive**: The naive estimator is a spline with 10 degrees of freedom that fits Y on X without taking care of the confounder and instrumental variable.
- **2SLS**: 2SLS is implemented with splines with 10 degrees of freedom in the second stage and in the first stage it is 10 for the first regression, 11 for the second and so on.
- **CF**: The CF algorithm is also implemented with splines with 10 degrees of freedom for both stages.
- **DeepCF**: DeepCF is implemented with two neural networks, both of which have 3 layers with, respectively, 64, 32, and 16 neurones per layer. The first stage uses 300 epochs, which is the number of times the networks pass through the entire dataset. The second stage uses 300 epochs. The time that DeepCF uses is much longer than to fit every other estimator, except DeepIV.

[1] https://github.com/SamuelJoray

- **DeepIV**: DeepIV is implemented with the same neural network structure as DeepCF. The first stage uses 100 epochs and the second uses 300 epochs. It uses ReLU activations. Since in most of our setups the first stage will be of the form $X = Z + H + N_X$ and H, N_X are normally distributed, we used only 2 number of components in the mixture of gaussian network in the first stage. Originally, it was set to 10, by setting it to 2, we get better performances. DeepIV uses the latest implementation of the EconML package[2].
- **DeepGMM**: DeepGMM also uses two neural networks of the same size as the ones in DeepCF and DeepIV. However, the number of epochs is much higher. It is 6000 and 3000 in the first and second stage, respectively. It uses ReLU activations. We encountered errors when trying to adjust the number of epochs. DeepGMM uses the latest implementation of the CausalML package[3].
- **GCFN**: GCFN uses an autoencoder with 100 neurones in the 2 hidden layers for both encoder and decode with 2000 epochs. The neural network in the second stage also has 2 hidden layers of 100 neurones each and uses 4000 epochs. It uses ReLU activations. The number of epochs is much larger than DeepIV and DeepCF. The original method in Puli and Ranganath (2021) uses 100 epochs for all the networks, and we tried to increase this number to obtain a better precision. After having contacted the A. Puli, we obtain the code of the gcfn method available publicly[4].

6.2 Estimated Mean Function

In this section, we look at different settings to have a general idea of how the algorithms that we described behave. More precisely, we estimate the counterfactual function with the different algorithms.

In the following, the variables Z, H, N_X, N_Y are sampled with the following distributions:

$$\begin{aligned} H &\sim \mathcal{N}(0,2), \\ Z &\sim \mathcal{N}(0,4), \\ N_X &\sim \mathcal{N}(0,1), \\ N_Y &\sim \mathcal{N}(0,1). \end{aligned} \tag{6.1}$$

In each plot, the true causal effect is shown in green and the grey region is the 95% confidence region, which has been calculated on 5 runs. The black

[2] https://github.com/py-why/EconML
[3] https://github.com/CausalML/DeepGMM
[4] https://github.com/rajesh-lab/gcfn-code

line represents the mean over the 5 runs of each algorithm. In the left plot, the data are shown in blue for only the last run. The number of data points used for training is 2000.

6.2.1 General Overview

First, we look at three situations to see how the algorithms perform in standard cases. The data is generated according to the following SCM:

$$\begin{aligned} X &= Z + H + N_X, \\ Y &= f(X) + H + N_Y, \end{aligned} \tag{6.2}$$

where the function f is, respectively,
$f_1(x) = \frac{1}{4}x, f_2(x) = |x|$ and $f_3(x) = sign(x)$. The plots are shown, respectively, in Figures 6.1, 6.2 and 6.3. For these standard cases, all the algorithms seem to perform similarly in terms of both precision and stability.

The true causal effect is shown in each plot in green.

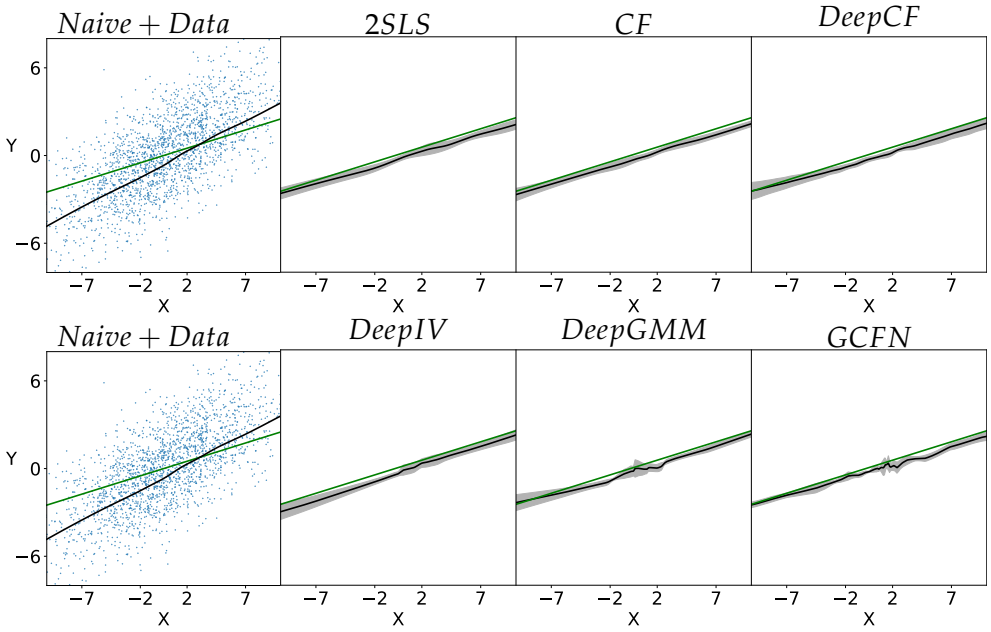


Figure 6.1: Linear effect. All estimators learn the causal effect correctly. The stability, described by the confidence region in grey, is similar for all estimators.

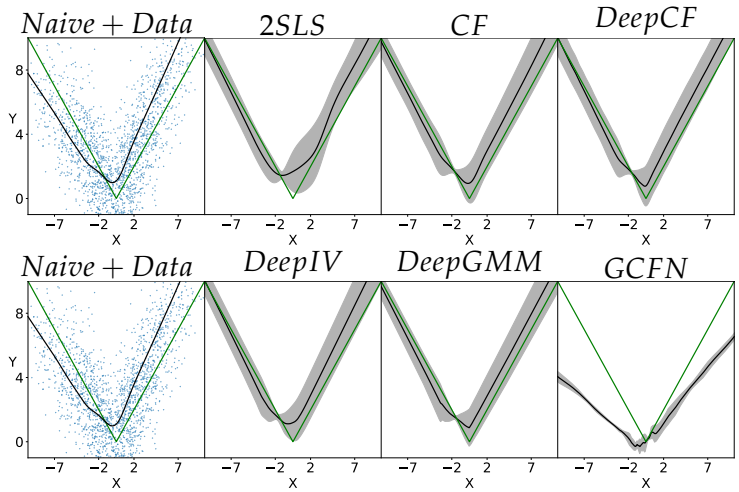


Figure 6.2: The function in the outcome is the absolute value function. All estimators except GCFN estimate the causal effect well. They mitigate the confounder by having a symmetry by the vertical axis $X = 0$. The naive estimator does not have this axis of symmetry due to the confounder. GCFN gets the shape of the absolute value but neither as the true effect nor as the naive estimator. It is not symmetric, though very stable.

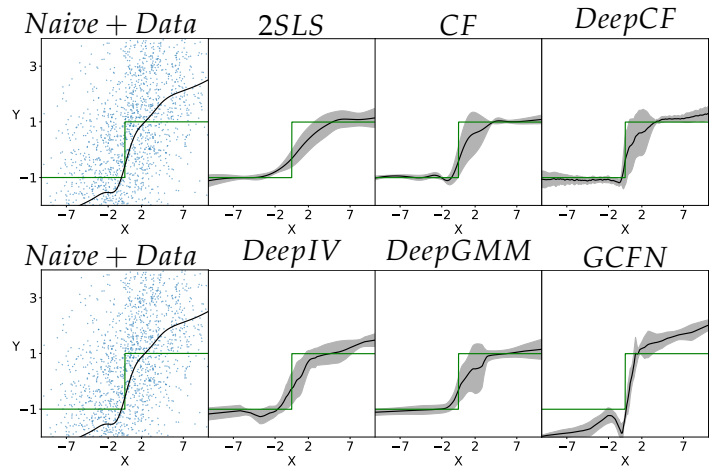


Figure 6.3: The function in the outcome is the sign function. 2SLS and DeepIV correct for the confounder, but do not get the step in the function well. The same holds for CF, DeepCF and DeepGMM with a slight improvement for those estimators. GCFN captures the shape of the sign function but increases instead of remaining horizontal, resembling the behavior of the naive estimator, even if to a lesser extent.

6.2.2 2SLS Algorithms are Ill-Posed

Next, we review Example 5.4 and show that for the CF type of estimators it performs well, but for the 2SLS type estimators, they are all ill-posed when Assumption 1*c* and their respective equivalent assumptions are not met. We

recall the SCM that is used to generate the data:

$$\begin{aligned} X &= Z + H, \\ Y &= cos(X) + H, \end{aligned} \tag{6.3}$$

where $H \sim \mathcal{U}[-\alpha, \alpha]$ and $Z \sim \mathcal{U}[-5, 5]$.

If α is bigger than 3, the conditional expectation $\mathbb{E}[Y|Z]$ is almost constant in Z and therefore Assumptions 1c and 1d are almost not satisfied. Since Z has a compact support, the Strong IV assumption in Theorem 4.11 is not met.

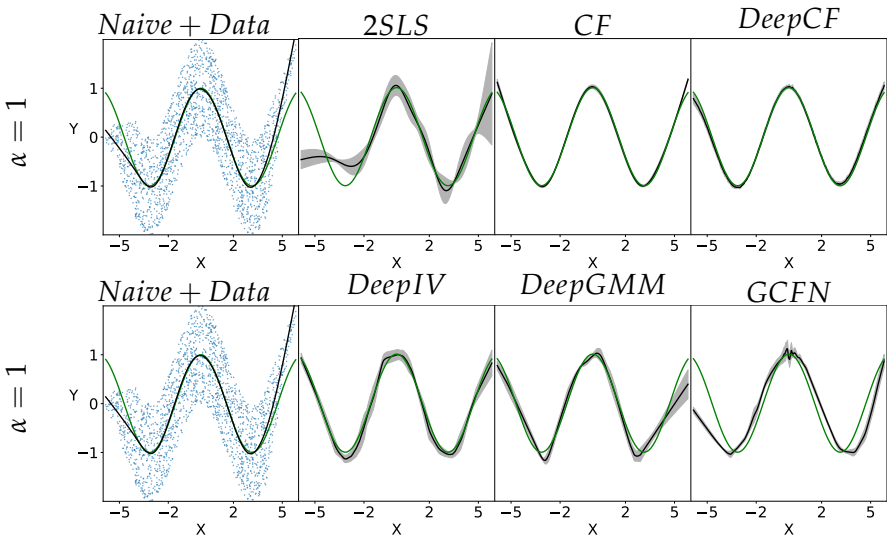


Figure 6.4: Sine curve with a small confounder. CF and DeepCF fit the shape of the sine curve and the confounder perfectly and are very stable. DeepIV and DeepGMM fit the shape very well and the stability is good. 2SLS is more unstable than the other, but still captures the shape and GCFN almost captures the shape and is very stable.

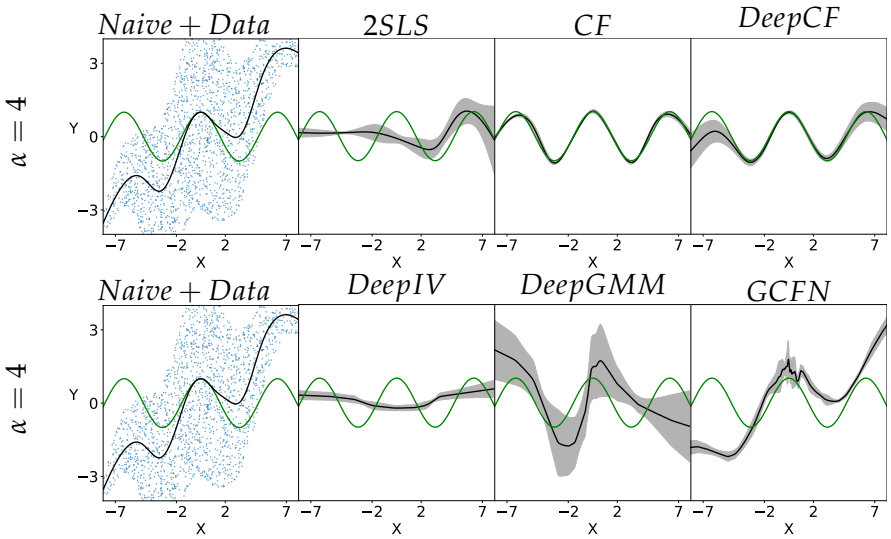


Figure 6.5: Sine curve with medium confounding effect. CF is again very precise and stable. DeepCF is very precise but has a very light positive linear trend. DeepIV does not get the shape of the sine curve due to the ill-posedness of the problem, but it is stable. DeepGMM is very unstable but gets the horizontal trend and almost the shape of the sine. Finally, GCFN does not mitigate the confounder well. The shape of the curve is similar to that of the naive estimator. This shows that DeepGMM is, in some sense, less affected than DeepIV if the assumption $1d$ is not satisfied. In the sense that it still seems to get the shape and might be consistent, which is better than DeepIV.

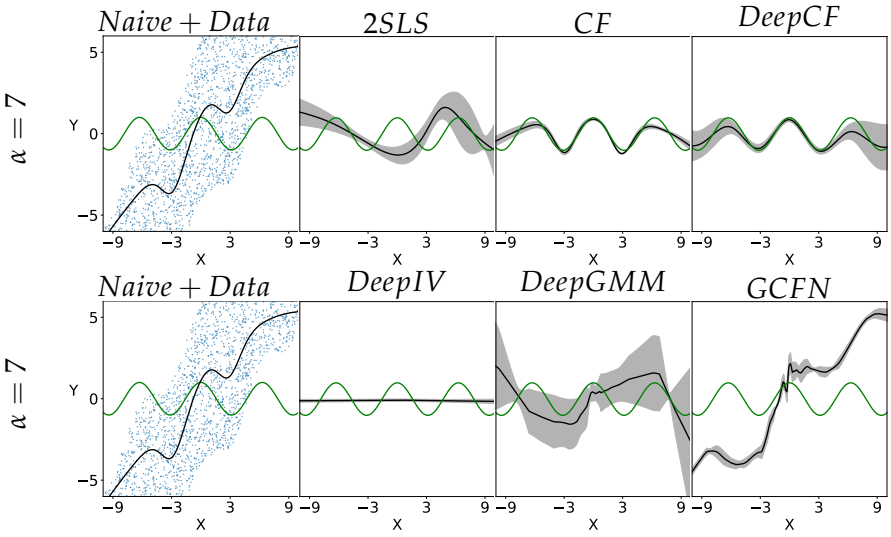


Figure 6.6: Sine curve with a strong confounding effect. CF and DeepCF perform very well, only slightly worse than in the previous cases. 2SLS and DeepGMM are unstable but get the horizontal trend. They reproduce a curve similar to the sine function, but very imprecisely. DeepIV is stable, but it gets only the horizontal trend, which is the other solution to the minimisation problem of deepIV.

6.2.3 Multiplicative Confounder in the Second Stage

The following setting demonstrates another strength of DeepCF compared to the other estimators. The data is generated according to the following SCM:

$$\begin{aligned} X &= \frac{1}{4}Z + \frac{1}{2}H + N_X, \\ Y &= X + |X|H + N_Y, \end{aligned} \tag{6.4}$$

where Z, H, N_X, N_Y are sampled as in 6.1.

The residuals of the first stage are denoted by $res = H + N_X$. The counterfactual function is:

$$\begin{aligned} h(x) &= \mathbb{E}_{res}[\mathbb{E}[Y|X, res]] \\ &= \mathbb{E}_{res}[\mathbb{E}[X + |X|H + N_Y|X, res]] \\ &= \mathbb{E}_{res}[\mathbb{E}[X + |X| \cdot H|X, res]] \\ &= \mathbb{E}_{res}[X + |X| \cdot \mathbb{E}[H|res]] \\ &= X + |X| \cdot \mathbb{E}_{res}[\mathbb{E}[H|res]] \\ &= X + |X| \cdot \mathbb{E}_{res}[H] \\ &= X, \end{aligned} \tag{6.5}$$

where we used tower property of conditional expectation in the second last equality and $\mathbb{E}[H] = 0$ in the last one.

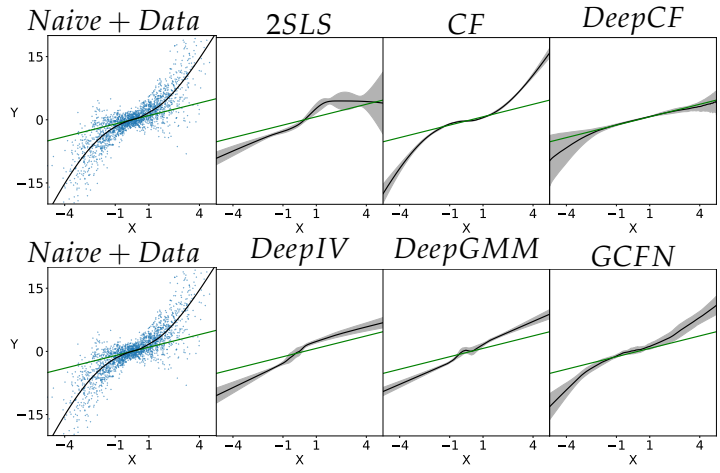


Figure 6.7: The plots suggest that only DeepCF is unbiased. It is very stable for the points in the middle and more unstable for the points (x, y) that we have not observed, but still in the range of X. The 2SLS, DeepIV and DeepGMM are close to the true effect, very stable but still biased. This is not concerning, as it typically assumes that the error and confounder are additive in the second stage. CF is similar as the naive estimator because the residuals of the first stage are uncorrelated with those of the second stage, which are $|X|H + N_Y$. GCFN is similar to CF and the naive estimator, but slightly better.

6.2.4 Multiplicative Confounder in the First Stage

This example shows a setting where the confounder H is multiplied by Z in the first stage and therefore breaks the assumptions of CF and DeepCF. We consider the following SCM.

$$\begin{aligned} X &= Z + ZH + N_X, \\ Y &= sin(X) + H + N_Y, \end{aligned} \tag{6.6}$$

where $Z, H \overset{i.i.d.}{\sim} \mathcal{N}(0,2)$ and $N_X, N_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

Note that the residuals of the first stage, given by $ZH + N_X$, are uncorrelated with Z but are dependent on Z. This is another example, where $\mathbb{E}[Y|Z]$ is almost but not exactly equal to 0. Hence, the Assumption $1c$ almost does not hold. The results are given in Figure 6.8.

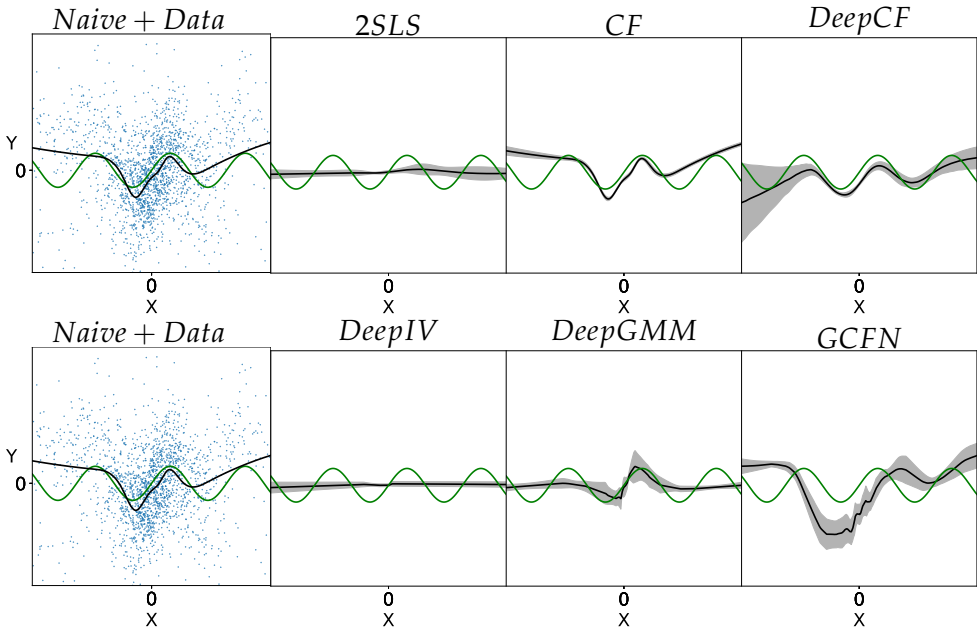


Figure 6.8: One assumption of DeepCF and CF is broken. CF is very similar to the naive case, since the residuals and IV are uncorrelated. DeepCF performs similarly to the naive case, which suggests that CF and DeepCF cannot learn the causal effect. DeepIV and 2SLS are almost constant 0 due to the fact that $\mathbb{E}[Y|Z]$ is almost constant and therefore Assumption $1c$ is almost not satisfied. DeepGMM captures well the causal effect, which again suggests that it handles much better than DeepIV the cases, where Assumption $1c$ almost but not completely breaks. GCFN captures some of the sine shape but with high imprecision.

6.3 Mean Squared Error

To study the convergence of the different algorithms, we consider another type of plot, the mean squared error (MSE) on the Y-axis and the number of

observations on the X-axis. We test two different settings, which satisfy the most restrictive assumptions.

The data of the first experiment is generated according to the SCM :

$$\begin{aligned} X &= Z + H + N_X, \\ Y &= |X| + H + N_Y, \end{aligned} \tag{6.7}$$

where $Z, H \overset{i.i.d.}{\sim} \mathcal{N}(0,2)$ and $N_X, N_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,0.5)$.

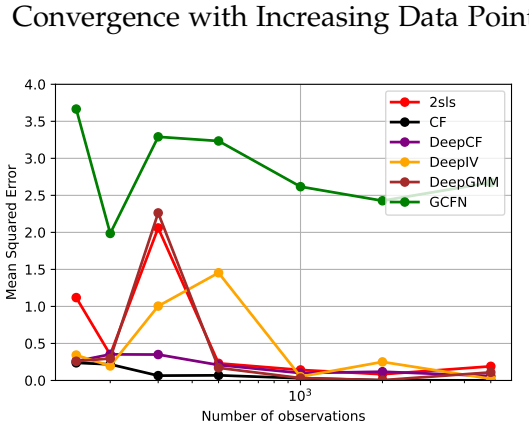


Figure 6.9: The mean squared error is shown on the Y-axis and the number of observations on the X-axis with a logarithmic scale. The MSE of all estimators but GCFN converge to 0. GCFN does not get significantly better as the size of the data increases.

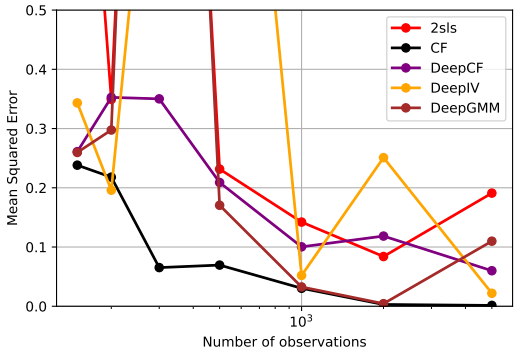


Figure 6.10: Same plot as in Figure 6.10, but zoomed in on the Y-axis. We see that the CF algorithm outperform every other algorithm. CF and DeepCF shows a constant decrease. DeepIV, DeepGMM and 2SLS are less stable.

6.4 Higher Dimensions

Now, we consider two higher-dimensional settings for the DeepCF, DeepIV and DeepGMM estimators. We do not include GCFN since it performed strictly worse in the last parts as the other estimators. We also do not include CF and 2SLS, though technically possible, the splines are not well generalisable to higher dimensions. To see how the estimators perform with respect to higher dimensions, we consider a plot that shows the mean squared error on the Y-axis and the number of dimensions on the X-axis. We only consider higher dimensions for Z and X. The confounder H and the outcome variable Y remain of dimension 1.

The setting of the first simulation, where we only consider linearity in the predictors. The first stage is:

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix} + \begin{bmatrix} H \\ \vdots \\ H \end{bmatrix} + \begin{bmatrix} N_{X_1} \\ \vdots \\ N_{X_k} \end{bmatrix}. \tag{6.8}$$

The second stage is given by:

$$Y = X_1 + \cdots + X_k + H + N_Y, \tag{6.9}$$

where $k = 2, 3, 5, 8, 12, 20, 30, 50, 75, 100$ are the dimensions that X and Z can take. All variables are sampled as follows: $Z_1, \ldots, Z_k, H, N_{X_1}, \ldots, N_{X_k}, N_Y, \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. The results are shown in Figures 6.11 and 6.12.

The second example introduces factors between the predictors creating non-linearity in the first stage. The first stage is given as follows:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} X_k \end{bmatrix} = \begin{bmatrix} Z_1 Z_2 \\ Z_2 Z_3 \\ \vdots \\ Z_{k-1} Z_k \\ Z_k \end{bmatrix} + \begin{bmatrix} H \\ \vdots \\ H \end{bmatrix} + \begin{bmatrix} N_{X_1} \\ \vdots \\ N_{X_k} \end{bmatrix}. \tag{6.10}$$

The second stage is given by:

$$Y = X_1 + \cdots + X_k + H + N_Y. \tag{6.11}$$

The results are shown in Figures 6.13 and 6.14.

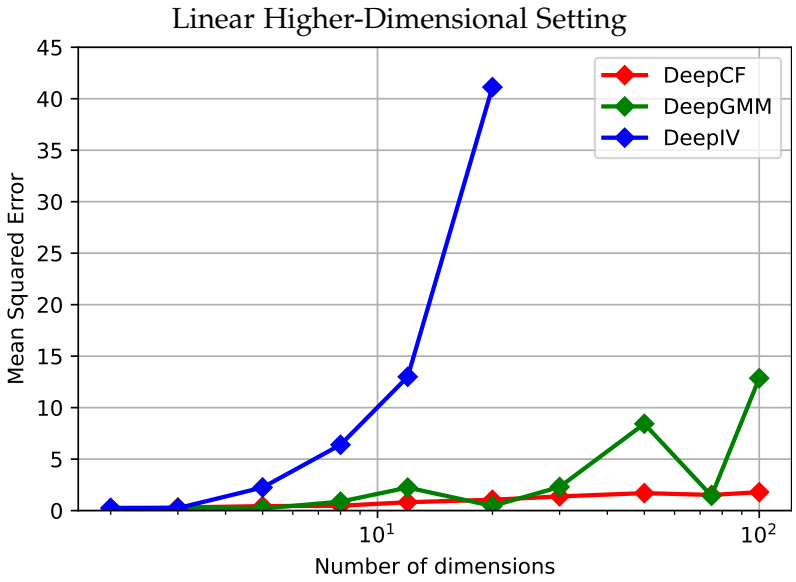


Figure 6.11: The linear setting with higher dimensions is shown. The X-axis is on a logarithmic scale. DeepIV produced values only for dimensions below 20 and the MSE grows very rapidly as the dimensions increase, much faster than DeepCF and DeepGMM. DeepGMM is unstable for higher dimensions and the MSE of DeepCF increases but remains very low and stable.

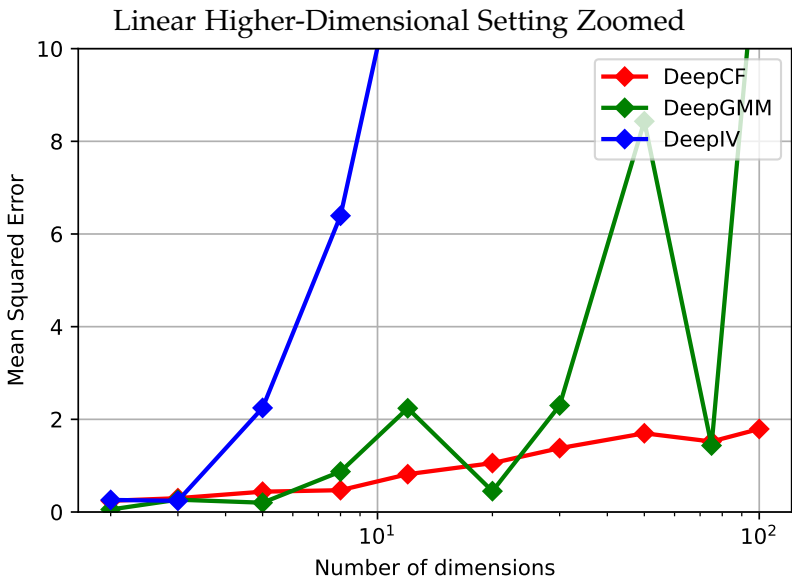


Figure 6.12: The linear setting with higher dimensions is shown with a zoom on the Y-axis. The X-axis is on a logarithmic scale. For dimensions 1 to 2, the MSE of DeepCF, DeepGMM and DeepIV are comparable. This is inline to what we have seen in Figure 6.1, then DeepCF outperforms DeepGMM.

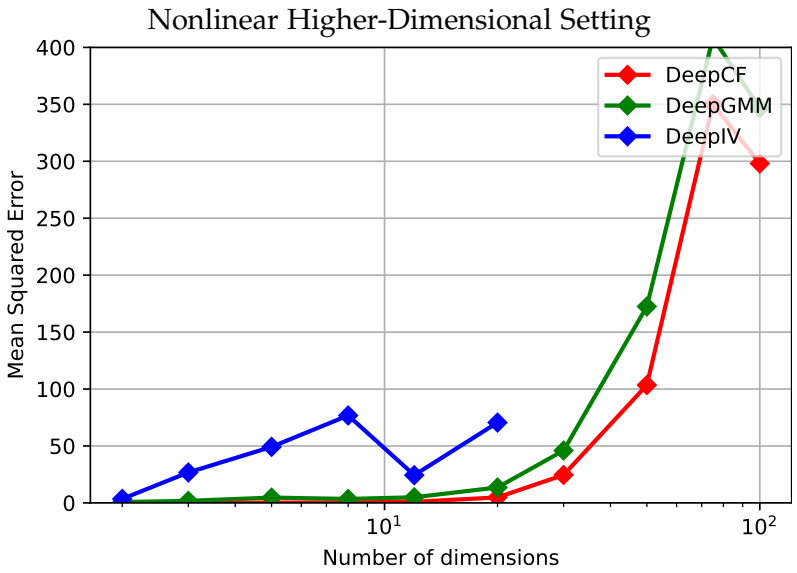


Figure 6.13: The nonlinear setting with higher dimensions is shown. Note that the scale of the Y-axis is much larger than in the previous example 6.11. The X-axis is on a logarithmic scale. DeepIV produced values only for dimensions below 20 and are much higher than the one for the others estimators. The MSE of both DeepGMM and DeepCF explodes for dimensions bigger than 20. DeepCF slightly outperforms DeepGMM for high dimensions.

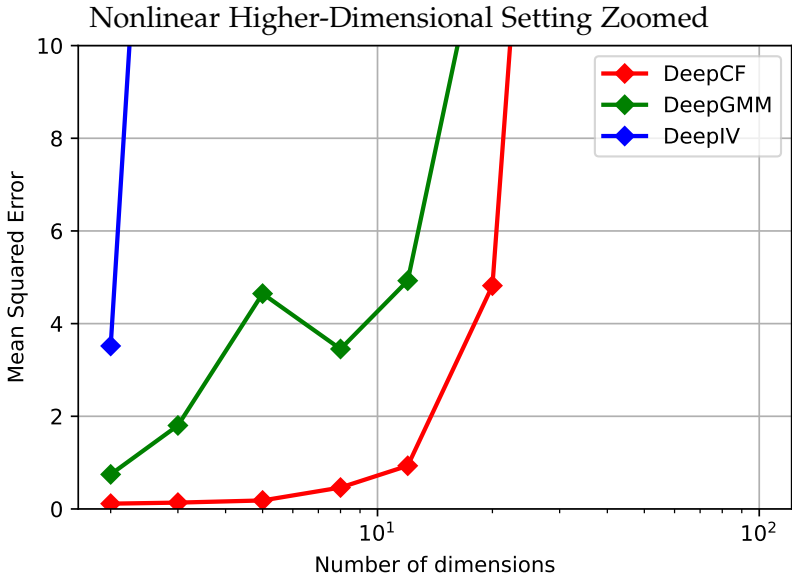


Figure 6.14: The nonlinear setting with higher dimensions with a zoom on the Y-axis is shown. The X-axis is on a logarithmic scale. DeepCF significantly outperforms DeepGMM for the dimensions below 30.

Chapter 7

Conclusion

Methods that utilise instrumental variables are incredibly diverse, employing various assumptions, each with its own set of advantages and inconveniences. This work focusses on elucidating the estimation of causal effects using instrumental variables. We delve into the two main branches of IV methods: two-stage least squares and control functions.

From theoretical perspectives, the primary advantage of 2SLS methods (such as 2SLS, DeepIV, DeepGMM) lies in their independence from requiring a specific form for the first stage, which proves convenient since the relationship between the IV and the treatment is often not the primary focus in experimental science. Indeed, the IV is a tool and the focus lies in the relationship between the treatment and the outcome. Consequently, the exact form of the treatment may be unknown, whether the IV is additive in the treatment or multiplicatively related to the confounder. However, a major drawback is the stringent requirement of the strong assumption 1*c*; even slight deviations from this assumption can render 2SLS methods unstable or inconsistent. We show the necessity of Assumption 1*c* for all 2SLS methods.

Control function methods (such as CF, DeepCF, GCFN), on the other hand, offer great flexibility in the second stage but demand that the confounder be additive in the treatment. We illustrate that the CF algorithm outperforms the 2SLS algorithm when the confounder is additive in the first stage.

In our simulations, we observed that 2SLS methods exhibited slightly greater instability compared to CF methods. Specifically, an example showcasing a scenario where the confounder and treatment are multiplicatively related in the second stage revealed that only DeepCF accurately learned the causal effect. Furthermore, we present two instances where Assumption 1*c* is nearly violated, indicating that although unstable, DeepGMM can still estimate the causal effect, unlike DeepIV, which converges to a constant function. This finding complements the equivalence of an assumption utilized in both

methods, as discussed in Section 5.4.1.

This thesis introduces DeepCF, which is intriguing both theoretically and in simulations. Despite its similarities with CF, where splines replace neural networks, it gives a nice generalisation of Assumption 4 and Assumption 5. DeepCF demonstrates superior performance in practice, outperforming DeepGMM and DeepIV across various settings, including low and higher dimensions. Additionally, our simulation results highlight the importance of the assumption in CF and DeepCF that the residuals of the first stage must be independent of H for the methods to function effectively. Otherwise, they might behave similarly to naive estimator. However, if this assumption holds, then CF and DeepCF outperform all other methods.

While the concept of utilising an autoencoder to generate general control functions is theoretically compelling, GCFN's practical performance is lacking. Although it often identifies the shape of the causal effect, it struggles to effectively mitigate the confounder.

Prospective research directions may include investigating the consistency of DeepIV under strong assumptions, potentially with an even stronger Assumption 1c to ensure that this assumption is not close of being broken. While proving consistency for DeepIV would be theoretically significant, its practical utility may be limited, as DeepGMM consistently outperforms DeepIV in various scenarios, despite exhibiting similar behaviours in many cases.

Another promising direction involves improving or better tuning the algorithm for GCFN, aiming to effectively implement the concepts of general control function given in the identification theorem. Such advances could be significant due to weak assumptions outlined in Theorem 4.11. Finally, may or may not be a trivial question, but elucidating the relationship between the reconstruction assumption, which suggests that the control function controls for the treatment, and the idea of having a control function that controls for the confounder in the second stage, as in Assumption 5, merits further exploration.

Appendix A

Appendix

Lemma A.1. *If a multiple linear regression is given by* $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \varepsilon$ *where* $x_1, \ldots, x_k$ *are uncorrelated with* x_{k+1} *and* ε *and where* x_{k+1} *and* ε *are correlated. Then we can estimate the coefficients* $\beta_1, \ldots, \beta_k$ *consistently but* β_{k+1} *is inconsistent.*

Proof. We first assume wlog that all covariates $x_1, \ldots, x_{k+1}$ have mean 0. We first use the solution of the least squares estimator where X denote the matrix of observations with $k+1$ predictors and n observations:

$$X = \begin{pmatrix} 1 & x_1^1 & \ldots & x_{k+1}^1 \\ 1 & x_1^2 & \ldots & x_{k+1}^2 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_1^n & \ldots & x_{k+1}^n \end{pmatrix} \tag{A.1}$$

We use Slutsky's Theorem in the second equation:

$$plim(\hat{\beta}) = plim((X^T X)^{-1} X^T y) = \beta + (plim(\frac{1}{n} X^T X))^{-1} plim(\frac{1}{n} X^T \varepsilon) \tag{A.2}$$

We first calculate now the term $(plim(X^T X))^{-1}$:

$$X^T X = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_1^i & \ldots & \sum_{i=1}^n x_{k+1}^i \\ \sum_{i=1}^n x_1^i & \sum_{i=1}^n (x_1^i)^2 & \ldots & \sum_{i=1}^n x_1^i x_{k+1}^i \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{k+1}^i & \sum_{i=1}^n x_{k+1}^i & \ldots & \sum_{i=1}^n (x_{k+1}^i)^2 \end{pmatrix} \tag{A.3}$$

$$plim(\frac{1}{n} X^T X) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & var(x_1) & \ldots & cov(x_1, x_{k+1}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & cov(x_1, x_{k+1}) & \ldots & var(x_{k+1}) \end{pmatrix} \tag{A.4}$$

Since we assume that x_j and x_{k+1} for $j = 1, \dots, k$ are uncorrelated, we have $cov(x_j, x_{k+1}) = 0$ and we can calculate the inverse of this matrix, knowing the inverse of the matrix A:

$$A := \begin{pmatrix} var(x_1) & \dots & cov(x_1, x_k) \\ \vdots & \ddots & \vdots \\ cov(x_1, x_k) & \dots & var(x_k) \end{pmatrix} \tag{A.5}$$

The matrix $plim(\frac{1}{n}X^TX)$ is now a 2×2 Block matrix and we can use a formula to compute its inverse:

$$(plim(\frac{1}{n}X^TX))^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \boldsymbol{A^{-1}} & \mathbf{0} \\ 0 & \mathbf{0} & (var(x_{k+1}))^{-1} \end{pmatrix} \tag{A.6}$$

We can then calculate the term $plim(\frac{1}{n}X^T\varepsilon)$.

$$plim(\frac{1}{n}X^T\varepsilon) = plim \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \\ \frac{1}{n}\sum_{i=1}^n x_1^i\varepsilon_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n x_{k+1}^i\varepsilon_i \end{pmatrix} = \begin{pmatrix} cov(1, \varepsilon) \\ \vdots \\ cov(x_k, \varepsilon) \\ cov(x_{k+1}, \varepsilon) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ cov(x_{k+1}, \varepsilon) \end{pmatrix} \tag{A.7}$$

$cov(x_j, \varepsilon) = 0$ since x_j is uncorrelated with ε. Hence, the term $(plim(\frac{1}{n}X^TX))^{-1}plim(\frac{1}{n}X^T\varepsilon)$ vanishes and we get that $\hat{\beta}_j$ is consistent with β_j for $j = 0, \dots, k$ but inconsistent for $j = k + 1$.

$$\begin{aligned} &plim(\hat{\beta}_j) = \beta_j + (plim(\frac{1}{n}X^TX))^{-1}plim(\frac{1}{n}X^T\varepsilon)_j = \beta_j \qquad j = 0, \dots, k \\ &\hat{\beta}_{k+1} = \beta_{k+1} + \frac{cov(x_{k+1}, \varepsilon)}{var(x_{k+1})} \end{aligned} \tag{A.8}$$

□

The following theorem can be found in Murphy and Topel (2002).

Theorem A.2. *Consider the model:*

$$y = x_2\beta + f(\theta, x_1)\gamma + u, \tag{A.9}$$

where θ is a vector and is estimated in a first stage by $\hat{\theta}$, x_2 is an exogenous vector to the error u and x_1 is also an exogenous vector and used in the first stage to get $\hat{\theta}$ and f is a vector of functions.
Let F be the matrix of predictions $f(\hat{\theta}, x_1)$. Let $A = (X_2, F)$ the matrix of observations for the second stage. Let $F^* = \frac{\partial f}{\partial \theta}$ be the matrix of derivatives of F with respect

to θ.
If the following assumptions are satisfied, then the least squares estimates of β and γ are consistent:

1. *$lim_{n\to\infty} A^T A = Q_0$, where Q_0 is a positive definite matrix.*
2. *$f(\theta, x_1)$ is twice continuously differentiable with respect to θ for every x_1 and $\frac{\partial f}{\partial \theta}$ is bounded in the sense: $lim_{n\to\infty} \frac{1}{n} A^T F^* = Q_1$, where Q_1 is a matrix with entry in $\mathbb{R}$ and $|\frac{\partial^2 f(\hat{\theta}, x_1)}{\partial \theta_i \partial \theta_j}| \leq M_2(x_1)$, where $lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} M_2(x_{1i}) < \infty$*
3. *$\hat{\theta}$ is estimated consistently with θ in the first stage.*

Proof. An outline of the proof can be found in Murphy and Topel (2002). □

Lemma A.3. *Let X be an $m \times n$ observation matrix with m predictors and n observations and Y the outcome. The residuals $(Y - \hat{Y})$ of the multiple regression $Y \sim X$ are uncorrelated with the column of X, i.e. the predictors. As a special case, we get that the predicted values are uncorrelated with the residuals.*

Proof. The coefficients $\beta = (\beta_0, \ldots, \beta_m)$ are estimated by the least squares as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{A.10}$$

We can write

$$X^T X\hat{\beta} = X^T Y \Leftrightarrow X^T (X\hat{\beta} - X) = 0 \tag{A.11}$$

Thus, every columns in X are uncorrelated with $(X\hat{\beta} - X)$ which are the residuals of the multiple regression $Y \sim X$. □

Bibliography

I. Andrews, J. Stock, and L. Sun. Weak instruments in iv regression: Theory and practice. 2018.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009. URL https://doi.org/10.1515/9781400829828.

A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis, 2020.

C. M. Bishop. *Pattern recognition and machine learning*. New York : Springer, 2006.

R. W. Blundell and J. L. Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/3700739.

J. Bound, D. A. Jaeger, and R. M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 1995.

D. Card. Chapter 30 - the causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pages 1801–1863. Elsevier, 1999. doi: https://doi.org/10.1016/S1573-4463(99)03011-4. URL https://www.sciencedirect.com/science/article/pii/S1573446399030114.

R. L. Eubank. *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker, 1988.

W. N. Evans and J. S. Ringel. Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics*, 72(1):135–154, 1999. ISSN 0047-2727. doi: https://doi.org/10.1016/S0047-2727(98)00090-5. URL https://www.sciencedirect.com/science/article/pii/S0047272798000905.

Z. Guo and D. S. Small. Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 2016.

J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Yale University Press, New Haven, 1923.

L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–54, 1982. URL https://EconPapers.repec.org/RePEc:ecm:emetrp:v:50:y:1982:i:4:p:1029-54.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/hartford17a.html.

T. Hastie, R. Tibshirani, and R. Friedman. *The Elements of Statistical Learning.* Springer New York, NY, 2009.

M. I. U. Husnain, A. Subramanian, and A. Haider. Robustness of geography as an instrument to assess impact of climate change on agriculture. *International Journal of Climate Change Strategies and Management*, 2018.

R. KRESS. *Linear Integral Equations.* New York: Springer-Verlag, 1989.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: a view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6232–6240, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

K. Murphy and R. Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97, 2002. URL https://EconPapers.repec.org/RePEc:bes:jnlbes:v:20:y:2002:i:1:p:88-97.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. . *Econometrica*, 75(5):1565–1578, 2003.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2000.

J. Peters, D. Janzig, and B. Schölkopf. *Element of Causal Inference*. The MIT Press, 2017.

A. M. Puli and R. Ranganath. General control functions for causal effect estimation from instrumental variables, 2021.

P. Wright. *The Tariff on Animal and Vegetable Oils*. Investigations in international commercial policies. Macmillan, 1928. URL https://books.google.ch/books?id=zJBBAAAAIAAJ.

A. Wu, K. Kuang, R. Xiong, and F. Wu. Instrumental variables in causal inference and machine learning: A survey, 2022.