

Mémoire

Auteur : Girkes, Théo

Promoteur(s) : Van Messem, Arnout

Faculté : Faculté des Sciences

Diplôme : Master en sciences mathématiques, à finalité approfondie

Année académique : 2024-2025

URI/URL : <http://hdl.handle.net/2268.2/22956>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

UNIVERSITY OF LIÈGE



FACULTY OF SCIENCES

DEPARTMENT OF MATHEMATICS

Causal inference and robustness with a focus on the inverse probability weighted mean

*A thesis submitted in fulfillment of the requirements for the
Master's degree in mathematical sciences, research focus*

AUTHOR: THÉO GIRKES

SUPERVISOR: PR. ARNOUT VAN MESSEM

ACADEMIC YEAR 2024–2025


Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Arnout Van Messem, for generously dedicating a great deal of his time, not only to proofreading my thesis, but also to guiding me throughout the two years of my master. I am deeply grateful for all the valuable advice he provided and for the confidence he placed in me. His contribution played a key role in shaping the direction and quality of this thesis.


Secondly, I would like to mention my father, Daniel, whose support and genuine interest in my work have been deeply encouraging.

Finally, I would like to dedicate my last words to my girlfriend, Soline, for her unwavering encouragement and patience throughout the countless hours of work, always at my side.

Abstract

This thesis provides an overview of causal inference and statistical robustness, with a focus on integrating classical robustness analysis into the causal framework. We introduce the basic definitions in causality, and discuss the modeling of causal situations with causal graphs. After that, we detail the process of inverse probability weighting, and focus on the estimator of the inverse probability weighted mean (IPWM). We study its properties in terms of bias, asymptotic variance, and convergence. Then, we explore the fundamentals of statistical robustness through the breakdown point and the influence function, both empirically and theoretically. Two possible applications of the influence function, namely, approximating the asymptotic distribution of an estimator and M-estimators, are discussed. Next, these robustness tools are applied to the IPWM estimator, revealing its lack of robustness in the presence of outliers. To confirm these findings, simulations are performed.  implementations of both the generating algorithm and the simulations are given. This work highlights the need for more robust alternatives to the IPWM in causal estimation contexts.

Résumé

Ce mémoire propose une vue d'ensemble de l'inférence causale et de la robustesse statistique, avec un accent particulier sur l'intégration de l'analyse classique de la robustesse dans un cadre causal. Nous introduisons les définitions fondamentales de la causalité et abordons la modélisation des situations causales à l'aide de graphes causaux. Ensuite, nous détaillons le procédé de pondération par la probabilité inverse (IPW), en nous concentrant sur l'estimateur de la moyenne pondérée par la probabilité inverse (IPWM). Nous étudions ses propriétés en termes de biais, de variance asymptotique et de convergence. Par la suite, nous explorons les fondements de la robustesse statistique à travers le point de rupture et la fonction d'influence, tant leur version empirique que théorique. Deux applications possibles de la fonction d'influence, à savoir l'approximation de la distribution asymptotique d'un estimateur et les M-estimateurs, sont également discutées. Ces outils de robustesse sont ensuite appliqués à l'estimateur IPWM, mettant en évidence son manque de robustesse en présence de valeurs aberrantes. Afin de confirmer ces résultats, des simulations ont été réalisées. Des implémentations  de l'algorithme de génération de données et des simulations sont fournies. Ce travail souligne ainsi la nécessité de développer des alternatives plus robustes à l'IPWM dans le cadre de l'estimation causale.

Contents

Acknowledgments	I
Abstract	II
Résumé	III
Contents	IV
1 Introduction	1
2 Causal inference	4
2.1 Introduction to causal inference	4
2.1.1 Causality	4
2.1.2 Causation versus association	7
2.1.3 Exchangeability	8
2.2 Causal graphs	9
2.2.1 Definition of a causal graph	9
2.2.2 Basic structures in causal graphs	14
2.3 Inference	18
2.3.1 Basic definitions	18
2.3.2 Methods to build an estimator	19
3 Inverse probability weighting	27
3.1 The process of inverse probability weighting	27
3.2 Inverse probability weighted mean	35
3.3 Empirical estimator of the inverse probability weighted mean	38
3.4 Properties of the IPWM estimator when the propensity score is known	45
4 Statistical Robustness	50
4.1 Concept of robustness	50
4.2 Breakdown point	52
4.2.1 Empirical breakdown point	53
4.2.2 Theoretical breakdown point	54
4.3 Influence function	59

4.3.1	Empirical influence function	59
4.3.2	The notion of derivation	61
4.3.3	Theoretical influence function	71
4.3.4	Link between the empirical and theoretical influence function	78
4.3.5	Desired properties of the influence function of an estimator	79
4.3.6	Asymptotic distribution approximation	80
4.3.7	M-estimators	86
5	Robustness of the inverse probability weighted mean estimator	91
5.1	State of the art	91
5.2	Breakdown point of the IPWM	92
5.3	Influence function of IPWM	93
5.3.1	Computation of the influence function	93
5.3.2	Properties of the influence function of the IPWM	95
6	Simulations and empirical analysis	98
6.1	Causal data	98
6.2	Generation of causal data	100
6.3	Simulations	101
7	Conclusion	108
A	Mathematical prerequisites	110
A.1	Functional analysis	110
A.1.1	Vector and metric spaces	110
A.1.2	Hilbert spaces	112
A.2	Measure theory and probability	113
A.2.1	σ -algebra	113
A.2.2	Measurability	114
A.2.3	Measure	115
A.2.4	Probability	117
A.2.5	Independence	119
A.2.6	Conditional expectation and conditional independence	120
A.2.7	Integration	122
A.2.8	Bochner's integration	127
B	Scripts	136
B.1	Generation of causal data	136
B.2	Simulations	138
	List of Figures	142
	List of Tables	143

<i>CONTENTS</i>	VI
List of Algorithms	144
Bibliography	145

All models are wrong, but some are useful.

George Box

A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.

J.W. Tukey

Chapter 1

Introduction

In many fields, it is of interest to study the impact of a certain variable on a system (e.g., the health of a person or an ecological environment) or on another variable, called the outcome. A pharmaceutical company, for example, might want to investigate how a newly developed treatment impacts the sleep quality of people suffering from insomnia. A classical approach would be to perform a randomized observational study on a representative sample of the population of interest, in which all participants will be randomly assigned to the treatment or control group. Using the collected data, a statistical analysis is then performed to evaluate the impact of the treatment on the quality of sleep.

It is important to note that, in this context, we would like to quantify how the fact of undergoing the treatment compared to not having the treatment, while all other factors remain the same, influences the outcome under study. In other words, we would like to infer the causal effect of the treatment by comparing the outcome when receiving the treatment with the outcome when not receiving the treatment in such a way that this is the only difference between the two situations. This type of inference is called causal inference and differs from classical inference by the philosophy of what we try to evaluate, namely the causal effect as opposed to the association effect.

The concept of causality has been studied since the early twentieth century. In 1921, Wright [69] was one of the first to provide a theoretical foundation to distinguish between correlation and causation using what he calls ‘path coefficients.’ However, this technique received some criticism [43] and did not attract much interest in the scientific community until the 1950s, when concern about bias in the estimation of a treatment effect was raised due to the limitations of observational studies [24]. Research in this domain over the following decades culminated in Rubin [54] and Robins [49] providing a theoretical framework to properly define the causal effect using potential outcomes, which Neyman had previously defined informally in his thesis in 1923 [61]. In the late 1990s, Pearl [44] introduced multiple tools to analyze causality, such as causal graphs.

Research on the topic of causal inference is gaining in popularity, as many industries like to perform an analysis of the impact of a certain variable (e.g., a treatment, the composition of a chemical element, the weather, etc.) while everything else remains unchanged. Examples can be found in domains such as the pharmaceutical and medical industry [70],

economics [13, 14], psychology [51], physics [57] and biology [38]. The reason for the increased popularity of causal inference is that it gives more valuable information about what differences, implied by the change of the variable of interest, should be expected than merely studying the association effect, as the causal effect is not influenced by any other factors, while the association effect is.

As mentioned, and under the assumption that the sample is representative of the population under study, most of the time, a randomized study allows us to perform such an analysis of the influence of the treatment on the outcome. Since the attribution of the treatment is completely at random, this kind of study will effectively compare two situations that only differ from each other in the treatment, thus leading to the causal effect, or more precisely, the average causal effect (as opposed to the individual causal effect) of the treatment.

The major issue with this approach is that such a randomized study is not always feasible, due to technical constraints (e.g., the sample is not representative or too small), budget restrictions (e.g., the treatment is highly expensive and cannot be given to a sufficient number of patients) or ethical considerations (e.g., the treatment is developed to cure a serious disease for which it would not be ethical to randomly decide who receives the treatment and who the placebo). As such, we would like to estimate the average causal effect of a treatment through a non-randomized study. At first, this might seem intractable, but under certain conditions it is possible by using a specific estimator, the inverse probability weighted mean (IPWM), which is part of the family of inverse probability weighted estimators.

The IPWM is, as its name suggests, a variation of the traditional mean estimator used in classical inference, for which each observation is given a weight equal to the inverse probability that this individual received treatment. Since the classical estimator is known to be non-robust against outliers, i.e., it is highly influenced by the presence of outlying data points in the dataset, it is natural to study formally the robustness of the IPWM and potentially propose a more robust estimator.

However, to the best of our knowledge, there are only a few publications that investigate the robustness properties of the IPWM estimator, and when done, it is mainly from an empirical point of view [9, 71]. As such, rigorous mathematical foundations are usually lacking, or the proposed proofs are not detailed enough. As far as we know, nearly no formal theorems on the breakdown point or the influence function, whether these are empirical or theoretical, of this estimator are available in the literature. In fact, the major part of the theoretical research we have found on the influence function of the IPWM estimator was developed in the context of debiased machine learning (DML) and automatic debiased machine learning (auto-DML) [13, 14, 15]. In these fields, causal effects are estimated using machine learning techniques (such as a neural network or a regression model) combined with Neyman-orthogonalization [5] to avoid the regularization bias induced by the model. The influence function is utilized in the calculation of the Neyman-orthogonal moments and also appears in the asymptotic variance of the estimator, but not for the robustness properties of the estimator.

Nowadays, a variant of the IPWM, called the augmented inverse probability weighted mean

(AIPWM) or the doubly robust inverse probability weighted mean (DRIPWM), is mainly used, as it has the doubly robust property (which has nothing to do with robustness against outliers) [12, 35, 36, 59]. This estimator depends on a regression model of the outcome based on some covariates and a logistic model of the propensity score (which is the probability of receiving a treatment based on some covariates). Most of the research in this field concerns applications of the AIPWM in practical contexts (e.g., in medicine [15]), improving the estimator [10] with respect to its bias in specific cases (e.g., missing data) or reducing its (asymptotic) variance [33].

The main goals of this thesis are:

1. Introducing the concept of causality, which is a popular topic in the statistical community for its practical insights but is not developed at the University of Liège.
2. Providing the basics of statistical robustness theory and applying them to the inverse probability weighted mean estimator to obtain new results about its robustness.
3. Investigating the generating process of data with causal effects.

This thesis serves as a solid base for further research on the topic of robustness against outliers in the context of causal inference by providing solid mathematical foundations.

In this thesis, we first define the concept of causal inference and its aspects in Chapter 2. Next, we give the formal definition of the IPWM and demonstrate some of its properties in Chapter 3. The basic tools for verifying robustness are formally introduced in Chapter 4, after which, in Chapter 5, we apply these to the IPWM. In Chapter 6, the theoretical results are empirically confirmed by simulations. In addition, an algorithm for generating causal data is given.

Chapter 2

Causal inference

In this chapter, we will introduce the concept of causal inference and point out where causal inference differs from classical inference. We will introduce the core concepts of causality and give some measures to estimate a causal effect. After the introduction of this concept, we will explain how to use directed graphs to visualize causal relations and explain how these can be used to model certain situations. This will help us later in Section 6.3 to simulate causal data. Finally, some basic definitions and methods from classical inference are provided, accompanied by examples, as in Chapter 3 they will be studied for the inverse probability weighted mean.

Note that the content and structure of this chapter are mainly based on the following two books [31, 45].

2.1 Introduction to causal inference

2.1.1 Causality

Let us start by introducing the subject of causal inference and give some definitions that will form the basis for the rest of this thesis.

We will first define the concept of inference, as causal inference is a specific type of inference:

Definition 2.1.1 (Inference). Inference is the use of collected data, i.e., a sample, to make a statement about the underlying distribution of the entire population.

Note that this is a vague definition; this is due to the fact that there is no explicit definition of inference aside from the dictionary's one, which is not mathematically rigorous. However, this definition summarizes the idea behind this concept developed in [11].

Example 2.1.1. Suppose that we have a set of n observations, $\mathbf{x} = (x_1, \dots, x_n)$, i.i.d.

from an unknown distribution \mathcal{F}_θ , which depends on a parameter θ (e.g., $\mathcal{F}_\theta = \text{Exp}(\theta)$). Based on only our sample \mathbf{x} , we might then be interested in the following:

- Estimating θ .
- Obtaining a confidence interval of level $\alpha \in [0, 1]$ for θ , $CI(\theta, \alpha)$.
- Performing a hypothesis test $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta \neq \theta_0$, to see if the hypothesis $H_0 : \theta = \theta_0$ is plausible.

Definition 2.1.2 (Causal inference). Causal inference is a type of inference that considers and analyzes the influence and magnitude of a variable A (the action) on a, supposedly dependent, variable Y (the outcome).

In others words, in causal inference we are interested in the way a change in the value of the variable A will affect, directly or not, the value of the variable Y .

Remark 2.1.1 (Notation). Following [31], we will denote the action variable with A and the effect or outcome variable with Y . More specifically, we will consider, unless otherwise stated, the case where A is a random binary variable that represents the fact that an individual has been treated ($A = 1$) or not ($A = 0$) and Y is a random variable that represents the outcome of a patient.

Definition 2.1.3 (Potential/counterfactual outcomes). Denote $Y^{a=1}$ the outcome variable that would have been observed if the value of the variable A was $a = 1$ and $Y^{a=0}$ the outcome variable that would have been observed if the value of the variable A was $a = 0$. In that case, $Y^{a=1}$ and $Y^{a=0}$ are called potential outcomes or counterfactual outcomes.

For the sake of simplicity, we will consider that there is no interference between the potential outcomes of an individual and the value of the action variable for the other individuals, meaning that $Y_i^{A_i}$ is independent of the value of A_j when $i \neq j$. Furthermore, considering interference would lead to a not well-defined causal effect for an individual, as it could imply that the potential outcome of one individual would not only depend on the action relative to this individual, which should constitute the only difference between the potential outcomes.

Remark 2.1.2 (Notation). For a specific individual i , we write Y_i^a its outcome value under the assumption that $A_i = a$.

We can now formally define what we call the causal effect for an individual.

Definition 2.1.4 (Causal effect for an individual). The action A has a causal effect on an individual's outcome Y_i if $Y_i^{a=1} \neq Y_i^{a=0}$.

Remark 2.1.3 (Consistency). It is important to note that only one of the potential outcomes is the real, observed outcome, since when we perform an experiment to collect data, we will have to choose between treating the patient or not. Therefore, if we have a sample of size n , and $A_i = a$ for individual i , then $Y_i^a = Y_i^{A_i} = Y_i$, where Y_i denotes the observed outcome of the individual i , and we can write $Y^A = Y$, which is referred to as the consistency property.

We are interested in the value of the potential outcomes that are not the observed outcome (for example, the value of $Y^{a=0}$ if $A = 1$). However, it is not always feasible to retrieve this value, and as such it is considered a missing value because we did not observe it. So, rather than working on a causal effect for an individual, we will look at the average causal effect.

Definition 2.1.5 (Average causal effect). Let A be a binary random variable. We say that A has an average causal effect on a binary outcome Y if

$$\mathbb{P}(Y^{a=1} = 1) \neq \mathbb{P}(Y^{a=0} = 1).$$

We can generalize this definition for a non-binary outcome Y . In this case A has an average causal effect on Y if

$$\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[Y^{a=0}] \Leftrightarrow \mathbb{E}[Y^{a=1} - Y^{a=0}] \neq 0.$$

Remark 2.1.4. We could also define the average causal effect in the case where A is a discrete non-binary variable (and Y is non-binary). A has an average causal effect on Y if for at least two values of A , a and a' ,

$$\mathbb{E}[Y^a] \neq \mathbb{E}[Y^{a'}] \Leftrightarrow \mathbb{E}[Y^a - Y^{a'}] \neq 0.$$

However, in this thesis, we will only consider the binary case.

Remark 2.1.5. In practice, we are often interested on the average causal effect of a treatment. As a result, in this specific case, we call it the average treatment effect (ATE). Also, for the rest of this thesis, we will focus the discussion on the specific context of a binary variable A that represents a possible treatment affectation and Y a real outcome of a patient.

Remark 2.1.6. For $\mathbb{E}[Y^{a=1}] = \mathbb{E}[Y^{a=0}]$, we say that the null hypothesis of no average causal effect is true, and for $Y^{a=1} \neq Y^{a=0}$ we say that the sharp causal null hypothesis is true.

Remark 2.1.7. Following [31], we define the average causal effect of the population as the difference in terms of the mean of the potential outcomes, but this effect can be defined using any difference of a functional of the potential outcomes. For example, we could have used the median, variance, standard deviation, median absolute deviation (MAD), etc.

Let us now define three ways of measuring the causal effect.

Definition 2.1.6 (Measures of causal effect). For a binary action variable A and a binary outcome variable Y , we define the following measures of the causal effect:

1. Causal risk difference: $\mathbb{P}(Y^{a=1} = 1) - \mathbb{P}(Y^{a=0} = 1)$
2. Risk ratio: $\frac{\mathbb{P}(Y^{a=1} = 1)}{\mathbb{P}(Y^{a=0} = 1)}$
3. Odds ratio: $\frac{\mathbb{P}(Y^{a=1} = 1) / \mathbb{P}(Y^{a=1} = 0)}{\mathbb{P}(Y^{a=0} = 1) / \mathbb{P}(Y^{a=0} = 0)} = \frac{\mathbb{P}(Y^{a=1} = 1) / (1 - \mathbb{P}(Y^{a=1} = 1))}{\mathbb{P}(Y^{a=0} = 1) / (1 - \mathbb{P}(Y^{a=0} = 1))}$

Remark 2.1.8. These definitions could also be generalized to a non-binary action variable A . However, for the sake of simplicity, we will only consider the case where A is binary.

Each of these measures provides information on the magnitude of the causal effect and may be chosen depending on the context. The causal risk ratio is just the difference of the probabilities of being cured had the patient taken the treatment or not, and the sign of this difference tells which probability is greater. The risk ratio is the ratio of these probabilities; the higher it is, the greater $\mathbb{P}(Y^{a=1} = 1)$ compared to $\mathbb{P}(Y^{a=0} = 1)$. Finally, the odds ratio is the ratio of the odds of being cured had the patient taken the treatment, and of being cured had the patient not taken the treatment. Therefore, these three measures compare the same two probabilities in different ways.

Example 2.1.2. Suppose that we have 1,000,000 patients, of which 75 patients would be cured if treated and 25 patients would be cured if not treated. We then see that the causal risk difference is equal to $\frac{75 - 25}{1,000,000} = \frac{1}{20,000}$, while the risk ratio is $\frac{75/1,000,000}{25/1,000,000} = 3$. We see that the risk ratio shows that the probability of being cured of the disease had the patient taken the treatment is three times higher than the probability of being cured of the disease had the patient not taken the treatment, while the causal risk difference shows the difference in magnitude between these two probabilities. Finally, the odds ratio is $\frac{(75/1,000,000)/(1 - 75/1,000,000)}{(25/1,000,000)/(1 - 25/1,000,000)} = \frac{75 \times 99,9975}{25 \times 99,9925} = 3.00015$, which means that the odds of being cured are three times higher had the patient taken the treatment.

2.1.2 Causation versus association

Let us define the concept of association and discuss the difference with causation, as it is primordial to not confound them and to make a clear distinction between both concepts.

Definition 2.1.7 (Association). Let A be an action variable and Y the outcome variable. We say that there is an association between A and Y if $\mathbb{E}[Y|A = 1] \neq \mathbb{E}[Y|A = 0]$.

We see that this definition is quite similar to the one of the average causal effect, but there is a difference about what is considered and how we will design our inference methods. To understand the difference, note that Y^a is the outcome variable if we were in the universe in which $A = a$ (of course, one of the values $a \in \{0, 1\}$ is the real value of A in our universe). In this context, we therefore consider the entire population, while $Y|A = a$ is the outcome variable of the subgroup associated with the value a , therefore, we only consider the part of the total population that respects the specific property.

Example 2.1.3. Consider the case where A is a variable representing whether a patient is treated or not ($A = 1$ or $A = 0$) and Y the outcome variable that represents whether a patient has been cured of some disease or not ($Y = 1$ or $Y = 0$). In this context, $Y^{a=1}$ is the variable that represents whether a patient has been cured (or not) in the case that he was treated, but it is possible that he was not treated. On the other hand, $Y|A = 1$ is the variable that represents whether a patient has been cured (or not) among all patients who were treated.

This is a slight difference, but an important one, as association relates to two distinct subgroups of the population, whereas causation considers the entire population.

2.1.3 Exchangeability

Let us now define the different types of exchangeability following [31], as these are properties that we are interested in and which will allow us to do inference.

Definition 2.1.8 (Full exchangeability). Let $\mathcal{A} = \{a_1, \dots, a_n\}$ be the set of all possible values of the random variable A . Full exchangeability is defined as

$$\{Y^{a_1}, \dots, Y^{a_n}\} \perp\!\!\!\perp A.$$

Full exchangeability corresponds to the fact that the joint distribution of the potential outcomes is independent of the distribution of the action variable A . This situation can be achieved in fully randomized studies, as the treatment a patient receives is totally random and therefore does not depend on the potential outcomes of the patient had he received the treatment or not.

However, full randomization of the treatment is not always possible in practice due to ethical, economical, or technical constraints. This is why other types of exchangeability have been introduced in the literature.

Definition 2.1.9 (Marginal exchangeability). Let $\mathcal{A} = \{a_1, \dots, a_n\}$ be the set of all possible values of the random variable A . Marginal exchangeability is defined as

$$\forall a \in \mathcal{A}, \quad Y^a \perp\!\!\!\perp A.$$

Marginal exchangeability corresponds to the fact that all the distributions of the potential outcomes are independent of the distribution of the action variable A . Note that, marginal exchangeability is implied by full exchangeability, but the reciprocal is not necessarily true. As such, marginal exchangeability is a weaker property than full exchangeability.

Despite marginal exchangeability being a weaker property than full exchangeability, it is still not always possible to obtain it without a fully randomized study. Therefore, we sometimes need to rely on an even weaker property, namely conditional exchangeability.

Definition 2.1.10 (Conditional exchangeability). Let $\mathcal{A} = \{a_1, \dots, a_n\}$ be the set of all possible values of the random variable A . Conditional exchangeability is defined as the fact that there exists a set of variables L such that

$$\forall a \in \mathcal{A}, \quad (Y^a \perp\!\!\!\perp A) | L, \tag{2.1}$$

in which case, we say that we have conditional exchangeability given L .

Conditional exchangeability corresponds to the fact that all the distributions of the potential outcomes and the distribution of the action variable A are independent given some set L of variables. This last property is the one that is most often used, as it is easiest to obtain and can even be obtained in non-randomized studies.

2.2 Causal graphs

In this section, we introduce one of the most simple yet useful tools we have at our disposal in causal inference to determine causal relations, causal graphs¹.

2.2.1 Definition of a causal graph

Let us first define what a graph is, what we mean by a causal graph, and illustrate the concept with some examples.

Definition 2.2.1 (Graph). A graph is a pair $G = (V, E)$, where V is a set of elements called the vertices, and $E \subseteq V^2$ is a subset of V^2 . The elements of E are called edges.

¹The content of this section is mainly based on [23, 44, 45].

Remark 2.2.1. Note that $(v, v') \in E$ is not the same as $(v', v) \in E$, the order matters. However, when a graph $G = (V, E)$ is such that $(v, v') \in E \Leftrightarrow (v', v) \in E$, we say that G is undirected. In the case where G does not respect this property, we say that G is directed.

Definition 2.2.2 (Acyclic graph). A graph $G = (V, E)$ is acyclic if it does not contain any cycle, i.e., $\nexists v_1, \dots, v_n \in V, e_1, \dots, e_n \in E$ for some $n \in \mathbb{N}_0$ such that

$$e_i = (v_i, v_{i+1}), i \in \{1, \dots, n-1\} \text{ and } e_n = (v_n, v_1).$$

Definition 2.2.3 (Causal graph). A causal graph $G = (V, E)$ is a directed acyclic graph (DAG) where the vertices (elements of V) represent random variables and the directed edges (elements of E) represent the link between the variables. That is, a variable X has a directed edge towards a variable Y if X may influence Y .

Remark 2.2.2 (Notation). To express that a variable X may have an influence on a variable Y , we write $X \rightarrow Y$.

Example 2.2.1. Suppose we have four variables X, Y, Z , and W , where $X \rightarrow Y$, $X \rightarrow Z$, $X \rightarrow W$, $W \rightarrow Z$, and $Y \rightarrow W$. The resulting causal graph is represented in Figure 2.1.

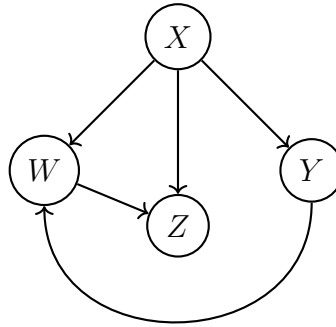


Figure 2.1: Example of a simple causal graph.

At this point, we have just represented graphically the links we know between our variables, and we might think that this is the extent of information that we obtain from the causal graph. However, the truth is that by using this graphical representation we can retrieve information that was intractable without it and that will be crucial to perform a correct causal analysis. Furthermore, at the same time, we get access to all the knowledge of graph theory and all DAG-associated theorems.

Remark 2.2.3. A causal graph needs to be acyclic which seems to limit the number of potential situations we can represent. For example, if we have a variable X that is the

study time of a student and Y his result on the exams, we could argue that the study time of the student influences his results ($X \longrightarrow Y$) but also that his results at the exams may influence his motivation, so the student could study less ($Y \longrightarrow X$). We end up with a cyclic relation and we might think that we cannot represent this situation with a causal graph, but the relation is not cyclic if we take into account the fact that the two events take place at different times, that is, the study time of the student will influence his exam results, but these results will influence his next study time for his next exam session. So, considering different variables for the study time and the results depending on the corresponding exam session, we can use a causal graph to model the situation.

As mentioned before, we can retrieve relevant information from a causal graph, but which information, other than the links that exist between the variables, can we obtain from this representation? To answer this question, let us consider the following situation:

Suppose that we want to estimate how the study time of a student, which we will represent by the variable X , influences his performance during his exam session, which we will represent by the variable Y . If we suppose that only X influences Y , meaning that no other factors have an influence on Y , then a simple way of estimating the influence of X on Y would be to consider the probability $\mathbb{P}(Y = y|X = x)$, which tells us the probability of Y to take a specific value y when X has the value x . With this we obtain the information we want to evaluate how X influences Y by changing its probability of taking the value y . However, this works only in the case where X is the only parameter that can influence Y , because if we consider another variable Z , which can be the motivation of the student, then the probability $\mathbb{P}(Y = y|X = x)$ may not give us the influence of X , but rather the influence of Z and X , since we may have the causal graph in Figure 2.2. So, even with the variable X set to x , we could argue that Y could still be influenced by Z and we could not really measure how X influences Y .

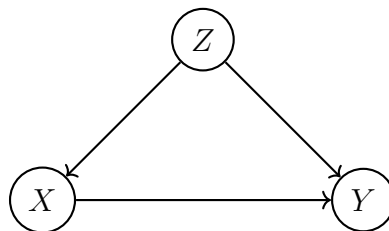


Figure 2.2: Causal graph of the influence of the study time.

Are we in a dead-end with no way of retrieving the causal link between X and Y ? Well, if we only consider using classical statistical tools, yes, we are. Luckily, we have causal graphs to help us. If we assume that no other variables than X and Z have a causal relationship with Y , that is, the graph contains all the possible variables that may have an effect on Y , then one way to estimate the influence of X and Y is to prevent Z from interfering with Y . By looking at the causal graph, we can see that Z is what we call a confounding variable, which will be explained in the following section, and that all we need

is to consider the influence of X on Y in each stratum created by Z . By doing so Z will no longer influence Y as the value of Z will be fixed in some way. If we consider Z to be a discrete variable, the influence of X on Y could then be written as the following sum:

$$\sum_z \mathbb{P}(Y = y | X = x, Z = z) \mathbb{P}(Z = z).$$

If we consider Z to be a continuous variable, this becomes:

$$\int_{\mathbb{R}} \mathbb{P}(Y = y | X = x, Z = z) f_Z(z) dz.$$

We have done what is called adjusting for Z : we have used our graph to understand the structure of the relationship we have between our variables and have used this knowledge to obtain the desired information about the influence of X on Y .

It is essential to understand that relying solely on data is insufficient to accurately estimate causal effects. This limitation arises because we either lack knowledge of, or fail to exploit, the underlying structure of the data, that is, the network of known or unknown relationships between variables. Without incorporating this structural information, our estimates remain incomplete or potentially misleading.

It is important to note that the causal graph associated with a given situation is not directly observable as we may not be aware of all the relationships between variables. However, causal graphs have specific properties that constrain the possible data they can represent. By leveraging our prior knowledge of the situation and systematically examining different graph structures, we can identify a graph that provides a reasonably accurate representation of the underlying reality.

Let us define the parents of a node X in a directed acyclic graph G .

Definition 2.2.4 (Parents of a node in a DAG). Let $G = (V, E)$ be a directed acyclic graph and $X \in V$ be a node of G . We call the parents of X the set of nodes $\text{Parents}(X) \subseteq V$, defined as

$$\text{Parents}(X) = \{Y \in V | (Y, X) \in E\}.$$

For any variable $Y \in \text{Parents}(X)$, we say that Y is a parent of X .

Also, we will need to know what a topological sort of a graph is and what property of a graph ensures that it possesses one.

Definition 2.2.5 (Topological sort of a graph). Let $G = (V, E)$ be a directed graph. A topological sort of G is an enumeration $v_1, \dots, v_{|V|}$ of the elements of V such that

$(v_i, v_j) \in E$ implies that $i < j$. In other terms, a topological sort of G can be seen as a bijection $\rho : V \rightarrow \{1, \dots, |V|\}$ such that

$$(v, v') \in E \Rightarrow \rho(v) < \rho(v').$$

Proposition 2.2.1. *Let $G = (V, E)$ be a directed graph. G is acyclic if and only if G possesses a topological sort.*

The demonstration of Proposition 2.2.1 can be found in [48].

Using what precedes, we can find another piece of information that the causal graph gives us about the joint distribution of the variables. However, in order to obtain it, we must make the assumption that all variables in the graph only depend on their parent variables, i.e, the variables only depend on the variables that have a direct effect on them. In this context, the joint distribution can be rewritten as in the following proposition.

Proposition 2.2.2. *Suppose that we have $n \in \mathbb{N}_0$ variables X_1, \dots, X_n and the causal graph G associated to them. Let us note $\text{Parents}(X_i)$ the set of parent variables of X_i in G and suppose that X_i is independent of all other variables conditionally on its parents (i.e., $X_i \perp\!\!\!\perp X_j | \text{Parents}(X_i)$ for $i \neq j$), then*

$$\mathbb{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbb{P}(X_i | \text{Parents}(X_i))$$

Proof. This is a direct consequence of the chain rule. We can assume that X_1, \dots, X_n are ordered such that the variable X_i has its parents among the variables X_1, \dots, X_{i-1} , i.e., $\text{Parents}(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$, which is always possible since the graph G is acyclic and, therefore, there exist at least one topological sort of G by Proposition 2.2.1. We have the following

$$\begin{aligned} \mathbb{P}(X_1, \dots, X_n) &= \prod_{i=1}^n \mathbb{P}(X_i | X_1, \dots, X_{i-1}) \text{ by the chain rule} \\ &= \prod_{i=1}^n \mathbb{P}(X_i | \text{Parents}(X_i)) \text{ by independence} \end{aligned}$$

□

By using this property, we can reduce the complex task of computing the joint probability to the computation of some simpler conditional distributions. Computationally, this is more advantageous since we need less memory to store some small conditional distributions instead of a more complex joint distribution.

Example 2.2.2. Let us consider a simple example to illustrate this. Suppose that we have n binary variables X_1, \dots, X_n . To store the entire joint distribution in memory we need to keep $2^n - 1$ values, namely all the possible combinations of values that X_1, \dots, X_n can take together minus the last one. This is because if we know every probability except for one we can obtain it by computing 1 minus the sum of all other probabilities. Now, suppose that we have the following causal graph:

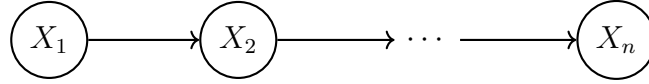


Figure 2.3: Example of a specific causal graph that will induced a simplified joint distribution.

We can rewrite the joint distribution as

$$\mathbb{P}(X_1, \dots, X_n) = \mathbb{P}(X_1) \prod_{i=2}^n \mathbb{P}(X_i | X_{i-1}).$$

In this case, using the fact that X_1, \dots, X_n are binary variables, the marginal distribution $\mathbb{P}(X_1)$ can be fully described with one value, $\mathbb{P}(X_1 = 1)$ (or $\mathbb{P}(X_1 = 0)$), since

$$\mathbb{P}(X_1 = 0) = 1 - \mathbb{P}(X_1 = 1).$$

Likewise, for all $i \in \{2, \dots, n\}$, the conditional distribution $\mathbb{P}(X_i | X_{i-1})$ can be fully described with the two values $\mathbb{P}(X_i = 1 | X_{i-1} = 1)$ and $\mathbb{P}(X_i = 1 | X_{i-1} = 0)$. This comes from the fact that

$$\mathbb{P}(X_i = 0 | X_{i-1} = 1) = 1 - \mathbb{P}(X_i = 1 | X_{i-1} = 1)$$

and

$$\mathbb{P}(X_i = 0 | X_{i-1} = 0) = 1 - \mathbb{P}(X_i = 1 | X_{i-1} = 0).$$

Therefore, we only needs to keep $1 + 2 + \dots + 2 = 1 + 2(n - 1) = 2n - 1$ values in memory in order to fully describe the joint distribution, which is linear in the number of variables instead of exponential.

Finally, a causal graph will also tell us if a variable Y might be influenced by a variable X . For this, we just look if there is an undirected path between X and Y , as the only way for X to have an effect on Y is to influence it directly ($X \rightarrow Y$) or to influence an ancestor (in the causal graph) of Y .

2.2.2 Basic structures in causal graphs

Let us now introduce three common structures we can find in general causal graphs and that have specific properties.

Chains

The first one is what we call a chain.

Definition 2.2.6 (Chain). A chain is a specific structure of a causal graph composed of three variables X, Y and Z such that

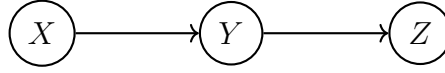


Figure 2.4: Chain structure.

If we want to analyze the causal relation we might have in a chain, we see that Y might influence Z , and that X might influence Y as well as Z .

Now, we may be interested not only in independence but also in conditional independence, and this is where the causal graph brings a new piece of information.

Proposition 2.2.3 (Conditional independence in chains). *If X, Y and Z form a chain, then we have the following properties*

- X and Z might be dependent,
- X and Y might be dependent,
- Y and Z might be dependent,
- $X \perp\!\!\!\perp Z|Y$.

We say that conditioning on Y blocks the influence of X on Z .

Proof. We only need to prove the last point, as the others are trivial. X has an effect on Y and since Y influences Z we see that X influences Z , but by conditioning on Y , a change in the value of X will not have any effect on Y , since its value is fixed, so X no longer has an effect on Y and neither does it have an effect on Z . \square

We will represent a variable we condition on by a gray filled circle as in the following graph.

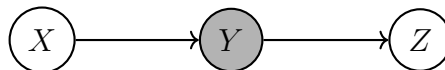


Figure 2.5: Causal graph with conditioning on Y .

Forks

The second structure is called a fork.

Definition 2.2.7 (Fork). A fork is a specific structure of a causal graph composed of three variables X, Y and Z such that

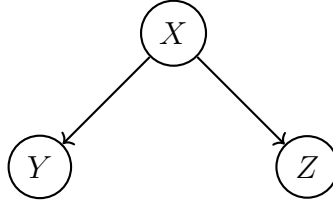


Figure 2.6: Fork structure.

Here, X is called the common cause.

Once again, let us look at the dependence we can find in a fork.

Proposition 2.2.4 (Conditional independence in forks). *If X, Y and Z form a fork where X is the common cause, then*

- X and Z might be dependent,
- X and Y might be dependent,
- Y and Z might be dependent,
- $Y \perp\!\!\!\perp Z | X$.

We say that conditioning on X blocks the influence of Y on Z .

Proof. If we do not condition on X then Y might have an effect on Z because if the value of Y changes then due to the fact that X influences both Y and Z , it could have an effect on the value of Z . The value of X might be related to the change in the value in Y and therefore also affect the value of Z . However, if we condition on X , this possible influence of Y on Z is blocked because X now has a fixed value. \square

Colliders

The third and final structure is called a collider.

Definition 2.2.8 (Collider). A collider is a specific structure of a causal graph composed of three variables X, Y and Z such that

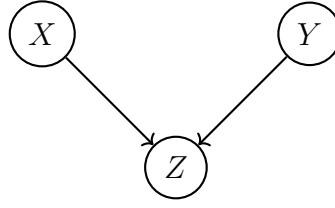


Figure 2.7: Collider structure.

Here, Z is called the common effect.

Let us look at the dependence we can find in a collider.

Proposition 2.2.5 (Conditional independence in colliders). *If X, Y and Z form a collider where Z is the common effect, then*

- X and Z might be dependent,
- Y and Z might be dependent,
- $X \perp\!\!\!\perp Y$ unconditionally,
- X and Y might be dependent conditionally on Z .

We say that conditioning on Z unblocks the influence of X on Y .

Proof. We have that X and Y are unconditionally independent because they do not have any relation between them aside from their common effect Z . A change in X (resp. Y) will not cause any change in Y (resp. X), only the value of Z might change accordingly to the value of X and Y .

If we condition on Z then the value of X might have an effect on the value of Y . To see this we just need to consider a simple situation, let X and Y be the result of two independent dices and let Z be the sum of these two results ($Z = X + Y$). If we condition on Z , i.e, taking $Z = z$, then a change in the value of X will result in a change in the value of Y because their sum has to be equal to z . They are thus dependent. \square

Causal graphs will be used at the end of this thesis in Section 6.3, to model a situation in which there are causal relations. Furthermore, an algorithm to generate data containing causal relations modelled by a causal graph will be introduced and discussed in Section 6.2.

2.3 Inference

Now that we have covered the basis for causality, we can dive into inference and develop its key aspects. Note that this section is mainly inspired by [11, 25, 27]. Let us start by giving some basic definitions.

2.3.1 Basic definitions

Definition 2.3.1 (Estimator and estimate). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample of size n , distributed according to some distribution F , i.e., $X_1, \dots, X_n \stackrel{iid}{\sim} F$. An estimator (also called statistic) $T(\mathbf{X})$ is any function T of \mathbf{X} . Furthermore, if $\mathbf{x} = (x_1, \dots, x_n)$ is a sample of observed values of \mathbf{X} , then $T(\mathbf{x})$ is called an estimate.

Remark 2.3.1. It is important to make a clear distinction between an estimator and an estimate, as the former is a random variable, while the latter is a fixed value. The estimate $T(\mathbf{x})$ is an observed realization of the estimator $T(\mathbf{X})$.

Most of the time, we use an estimator $T(\mathbf{X})$ in order to estimate a parameter θ depending on the distribution F . In this case, we usually write $T(\mathbf{X}) = \hat{\theta}_n(\mathbf{X})$ or simply $\hat{\theta}_n$. Of course, in this context, the goal of an estimator is to approximate as “good” as possible the unknown parameter θ . In order to compare how well different estimators approximate θ , we define some properties that are desired.

Definition 2.3.2 (Properties of an estimator). Let $\hat{\theta}_n$ be an estimator of some parameter θ . Then it is said that:

- $\hat{\theta}_n$ is unbiased (resp. biased) if $\text{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta = 0$ (resp. $\text{Bias}(\hat{\theta}_n) \neq 0$).
- $\hat{\theta}_n$ is asymptotically unbiased if $\lim_{n \rightarrow +\infty} \text{Bias}(\hat{\theta}_n) = 0$.
- $\hat{\theta}_n$ has an asymptotic zero variance if $\lim_{n \rightarrow +\infty} \mathbb{V}[\hat{\theta}_n] = 0$.
- $\hat{\theta}_n$ is weakly consistent if $\forall \varepsilon > 0, \lim_{n \rightarrow \infty} \mathbb{P}(\|\hat{\theta}_n - \theta\| \geq \varepsilon) = 0$ (i.e., $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$).
- $\hat{\theta}_n$ is strongly consistent if $\mathbb{P}(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta) = 1$ (i.e., $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta$).

Example 2.3.1. Let us consider $X_1, \dots, X_n \stackrel{iid}{\sim} F$, where F is a distribution of mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. We define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ and $S_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$.

We can prove that $\mathbb{E}[\bar{X}_n] = \mu$, meaning that \bar{X}_n is an unbiased estimator of μ , and

that $\mathbb{E}[S_n] = \frac{n-1}{n}\Sigma$, meaning that S_n is a biased estimator of Σ , but we have that $\lim_{n \rightarrow \infty} \mathbb{E}[S_n] = \lim_{n \rightarrow \infty} \frac{n-1}{n}\Sigma = \Sigma$ so S_n is asymptotically unbiased. As a result, a variant of the estimator S_n , that is $S'_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T = \frac{n}{n-1} S_n$, is preferred because $\mathbb{E}[S'_n] = \frac{n}{n-1} \mathbb{E}[S_n] = \Sigma$, and S'_n is therefore unbiased. Furthermore, we have that $\lim_{n \rightarrow +\infty} \mathbb{V}[\bar{X}_n] = \lim_{n \rightarrow +\infty} \frac{\Sigma}{n} = \mathbf{0}$ and \bar{X}_n has asymptotic zero variance.

A well-known theorem is the strong law of large numbers (SLLN), which states that under a condition about the existence of the first absolute moment of X_i , \bar{X}_n is strongly consistent (see [47]), which also implies that it is weakly consistent.

Theorem 2.3.1 (Strong law of large numbers). *Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F$ for some univariate distribution F , with $\mathbb{E}[X_i] = \mu$, then*

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

if and only if $\mathbb{E}[|X_i|] < +\infty$.

A common measure of the quality of an estimator is what we call the mean squared error (MSE).

Definition 2.3.3 (Mean squared error). Let $\hat{\theta}_n$ be an estimator of some parameter θ . The mean squared error is the quantity given by

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = (\mathbb{E}[\hat{\theta}_n] - \theta)^2 + \mathbb{V}[\hat{\theta}_n] = \text{Bias}(\hat{\theta}_n)^2 + \mathbb{V}[\hat{\theta}_n].$$

Example 2.3.2. If $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ are two estimators of θ and $\text{MSE}(\hat{\theta}_n^{(1)}) \leq \text{MSE}(\hat{\theta}_n^{(2)})$, then $\hat{\theta}_n^{(1)}$ could be considered more adapted, that is, $\hat{\theta}_n^{(1)}$ can be seen as an objectively better estimator in terms of tradeoff bias/variance, as its mean squared error is lower. Note that, we could prefer to have an estimator with the lowest bias regardless of its variance (or the opposite). In such case, the mean squared error is not what will help to make a decision between two estimators.

2.3.2 Methods to build an estimator

In a general context, it may not be straightforward to build an estimator. Nevertheless, there exist some methods that can give a possible solution used to obtain estimators of the parameters characterizing the underlying distribution. Let us quickly explain three of them.

The first is called the method of moments. It uses the fact that the theoretical moment of order $k \in \mathbb{N}_0$, that is, $\mu_k = \mathbb{E}[X^k]$, depends on the parameters of the underlying distribution.

Remark 2.3.2 (Notation). In the rest of this thesis, we will write “independently and identically distributed” as “iid”.

Definition 2.3.4 (Method of moments). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample of size n , drawn i.i.d. from a distribution F depending on a parameter vector $\theta = (\theta_1, \dots, \theta_p)$ with $p \in \mathbb{N}_0$. Let us write $\forall k \in \mathbb{N}_0$

$$\mu_k(\theta) = \mathbb{E}[X_i^k] \text{ and } m_k = \frac{1}{n} \sum_{i=1}^n X_i^k,$$

the theoretical and empirical moment of order k , respectively. The estimator of θ obtained by the method of moments, denoted by $\hat{\theta}_{MM}$, is the solution in terms of m_1, \dots, m_p of the following system of equations:

$$\begin{cases} m_1 &= \mu_1(\theta) \\ \vdots & \vdots \\ m_p &= \mu_p(\theta) \end{cases}.$$

Example 2.3.3. Suppose that we have $\mathbf{X} = (X_1, \dots, X_n)$ a sample of size n such that $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Gumbel}(\alpha, \beta)$ with $\alpha \in \mathbb{R}, \beta > 0$. This continuous distribution is characterised by its density function

$$f(x; \alpha, \beta) = \frac{1}{\beta} \exp\left(-\exp\left(-\frac{x - \alpha}{\beta}\right)\right) \exp\left(-\frac{x - \alpha}{\beta}\right).$$

Let us start by computing $\mu_1(\alpha, \beta)$. We have

$$\begin{aligned} \mu_1(\alpha, \beta) &= \mathbb{E}[X_i] \\ &= \int_{\mathbb{R}} x \frac{1}{\beta} \exp\left(-\exp\left(-\frac{x - \alpha}{\beta}\right)\right) \exp\left(-\frac{x - \alpha}{\beta}\right) dx, \end{aligned}$$

and using the change of variable

$$\varphi : \mathbb{R} \rightarrow [0, +\infty[: x \mapsto \exp\left(-\frac{x - \alpha}{\beta}\right) \tag{*}$$

with inverse $\varphi^{-1}(t) = \alpha - \beta \ln(t)$ and Jacobian $J(t) = \frac{\beta}{t}$. Therefore, we obtain

$$\mu_1(\alpha, \beta) = \int_0^{+\infty} (\alpha - \beta \ln(t)) \frac{1}{\beta} \exp(-t) t \frac{\beta}{t} dt$$

$$\begin{aligned}
 &= \alpha \int_0^{+\infty} \exp(-t) dt - \beta \int_0^{+\infty} \ln(t) \exp(-t) dt \\
 &= \alpha [-\exp(-t)]_0^{+\infty} - \beta \int_0^{+\infty} \ln(t) \exp(-t) dt \\
 &= \alpha + \beta\gamma \text{ where } \gamma = \lim_{n \rightarrow +\infty} \left(\sum_{k=1}^n \frac{1}{k} - \ln(n) \right).
 \end{aligned}$$

The last equality holds due to the fact that $\int_0^{+\infty} \ln(t) \exp(-t) dt = D_t \Gamma(t)|_{t=1} = -\gamma$ (see [68]), where

$$\Gamma :]0, +\infty[\rightarrow \mathbb{R} : n \mapsto \int_0^{+\infty} t^{n-1} \exp(-t) dt.$$

Furthermore, for $\mu_2(\alpha, \beta)$, we have

$$\begin{aligned}
 \mu_2(\alpha, \beta) &= \mathbb{E} [X_i^2] \\
 &= \int_{\mathbb{R}} x^2 \frac{1}{\beta} \exp \left(-\exp \left(-\frac{x - \alpha}{\beta} \right) \right) \exp \left(-\frac{x - \alpha}{\beta} \right) dx \\
 &= \int_0^{+\infty} (\alpha - \beta \ln(t))^2 \exp(-t) dt \text{ using the change of variable } (*) \\
 &= \alpha^2 \int_0^{+\infty} \exp(-t) dt - 2\alpha\beta \int_0^{+\infty} \ln(t) \exp(-t) dt + \beta^2 \int_0^{+\infty} (\ln(t))^2 \exp(-t) dt \\
 &= \alpha^2 + 2\alpha\beta\gamma + \beta^2 D_t^2 \Gamma(t)|_{t=1} \\
 &= \alpha^2 + 2\alpha\beta\gamma + \beta^2 \left(\gamma^2 + \frac{\pi^2}{6} \right).
 \end{aligned}$$

The last two equalities follow from the facts that $D_t^2 \Gamma(t)|_{t=1} = \int_0^{+\infty} (\ln(t))^2 \exp(-t) dt$ and $D_t^2 \Gamma(t)|_{t=1} = \gamma^2 + \frac{\pi^2}{6}$ (see [68]).

We can finally use the method of moments to obtain estimators for α and β . For that let us solve the following system in terms of m_1 and m_2 :

$$\begin{aligned}
 &\begin{cases} m_1 &= \mu_1(\alpha, \beta) \\ m_2 &= \mu_2(\alpha, \beta) \end{cases} \\
 &\Leftrightarrow \begin{cases} m_1 &= \alpha + \beta\gamma \\ m_2 &= \alpha^2 + 2\alpha\beta\gamma + \beta^2 \left(\gamma^2 + \frac{\pi^2}{6} \right) \end{cases} \\
 &\Leftrightarrow \begin{cases} m_1 - \beta\gamma &= \alpha \\ m_2 &= (m_1 - \beta\gamma)^2 + 2(m_1 - \beta\gamma)\beta\gamma + \beta^2 \left(\gamma^2 + \frac{\pi^2}{6} \right) \end{cases} \\
 &\Leftrightarrow \begin{cases} \alpha &= m_1 - \beta\gamma \\ \beta^2 \frac{\pi^2}{6} &= m_2 - m_1^2 \end{cases}
 \end{aligned}$$

$$\Leftrightarrow \begin{cases} \alpha &= \frac{m_1 - \gamma \frac{\sqrt{6(m_2 - m_1^2)}}{\pi}}{\pi} \\ \beta &= \frac{\sqrt{6(m_2 - m_1^2)}}{\pi} \end{cases},$$

where we always have that $m_2 - m_1^2 \geq 0$, because

$$\begin{aligned} S_n &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2 \frac{1}{n} \sum_{i=1}^n X_i \bar{X}_n + \frac{1}{n} \sum_{i=1}^n \bar{X}_n^2 \\ &= m_2 - 2 \bar{X}_n \frac{1}{n} \sum_{i=1}^n X_i + \bar{X}_n^2 \\ &= m_2 - 2m_1^2 + m_1^2 \\ &= m_2 - m_1^2, \end{aligned}$$

and $S_n \geq 0$. So, the obtained estimator is $\hat{\theta}_{MM} = (\hat{\alpha}_{MM}, \hat{\beta}_{MM})$ where

$$\hat{\alpha}_{MM} = \bar{X}_n - \gamma \frac{\sqrt{6s_n}}{\pi} \text{ and } \hat{\beta}_{MM} = \frac{\sqrt{6s_n}}{\pi},$$

with $s_n = \sqrt{S_n}$.

Another manner of obtaining estimators is to determine them using the maximum likelihood principle, that is, choosing the estimators in such a way that they maximize the likelihood function.

Definition 2.3.5 (Likelihood function and log-likelihood function). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample of size n drawn i.i.d. from a continuous (resp. discrete) distribution F depending on a parameter vector $\theta \in \Theta$, and let $f(\cdot; \theta)$ be the density (resp. probability mass) function of F . The likelihood function is defined as the function

$$L : \Theta \rightarrow [0, +\infty[: \theta \mapsto L(\theta; X_1, \dots, X_n) = \prod_{i=1}^n f(X_i; \theta).$$

The log-likelihood is defined as the natural logarithm of the likelihood, i.e., the function

$$\ell : \Theta \rightarrow \mathbb{R} : \theta \mapsto \ell(\theta; X_1, \dots, X_n) = \sum_{i=1}^n \ln(f(X_i; \theta)).$$

Definition 2.3.6 (Maximum likelihood estimators (MLE)). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample of size n , drawn i.i.d. from a distribution F , either discrete or continuous, depending on a parameter vector $\theta \in \Theta$. The maximum likelihood estimator of $\theta \in \Theta$ is given by

$$\hat{\theta}_{ML}(\mathbf{X}) = \arg \max_{\theta \in \Theta} L(\theta; X_1, \dots, X_n) = \arg \max_{\theta \in \Theta} \ell(\theta; X_1, \dots, X_n).$$

The idea of this method is to take the estimator that is the most likely based on the sample that is available.

Example 2.3.4. Suppose that we have $\mathbf{X} = (X_1, \dots, X_n)$ a sample of size n such that $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Exp}(\lambda)$ with $\lambda > 0$. This continuous distribution has the following density function

$$f(x; \lambda) = \lambda \exp(-\lambda x) \mathbb{1}_{[0, +\infty[}(x).$$

The maximum likelihood estimator for λ is

$$\hat{\lambda}_{ML} = \arg \max_{\lambda > 0} \ell(\lambda; X_1, \dots, X_n).$$

In order to find the value of $\hat{\lambda}_{ML}$, we can take the partial derivative of the log-likelihood function with respect to λ and equal it to 0, that is

$$\begin{aligned} \frac{\partial}{\partial \lambda} \ell(\lambda; X_1, \dots, X_n) &= \frac{\partial}{\partial \lambda} \sum_{i=1}^n \ln(f(X_i; \lambda)) \\ &= \frac{\partial}{\partial \lambda} \sum_{i=1}^n (\ln(\lambda) - \lambda X_i) \\ &= \frac{\partial}{\partial \lambda} (n \ln(\lambda) - n \lambda \bar{X}_n) \\ &= \frac{n}{\lambda} - n \bar{X}_n, \end{aligned}$$

and we have

$$\frac{\partial}{\partial \lambda} \ell(\lambda; X_1, \dots, X_n) = 0 \Leftrightarrow \frac{n}{\lambda} - n \bar{X}_n = 0 \Leftrightarrow \lambda = \frac{1}{\bar{X}_n}.$$

Furthermore, the second partial derivative of the log-likelihood with respect to λ is

$$\frac{\partial^2}{\partial \lambda^2} \ell(\lambda; X_1, \dots, X_n) = -\frac{n}{\lambda^2},$$

which is strictly negative $\forall \lambda > 0$. This implies that the log-likelihood has a unique global maximum exactly at $\lambda = \frac{1}{\bar{X}_n}$, therefore, we can conclude that

$$\hat{\lambda}_{ML} = \frac{1}{\bar{X}_n}.$$

One important property of the maximum likelihood estimator is that for any function of the parameter that is of interest to estimate, it only requires to apply this function on the MLE of the parameter to obtain the MLE of the transformed parameter (see [11]).

Proposition 2.3.1 (Invariance of MLEs). *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample of size n , drawn independently and identically from a distribution F depending on a parameter vector $\theta \in \Theta$, and let $\hat{\theta}_{ML}$ be MLE of θ . Then for any bijection $g : \Theta \rightarrow \mathcal{E}$, where \mathcal{E} is a space, the MLE of $g(\theta)$ is $g(\hat{\theta}_{ML})$.*

Proof. Suppose that $\hat{\theta}_{ML}$ is the MLE of θ , that is

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} L(\theta; X_1, \dots, X_n).$$

If we write $\eta = g(\theta) \Leftrightarrow \theta = g^{-1}(\eta)$, then to obtain the MLE of η , we need to maximize the likelihood function in terms of $g^{-1}(\eta)$, that is

$$\hat{\eta}_{ML} = \arg \max_{\eta \in \mathcal{E}} L(g^{-1}(\eta); X_1, \dots, X_n).$$

Since $\hat{\theta}_{ML}$ is the MLE of θ , we have that the maximum of the likelihood is attained at $\hat{\theta}_{ML}$, therefore $\hat{\theta}_{ML} = g^{-1}(\hat{\eta}_{ML})$, and thus, $\hat{\eta}_{ML} = g(\hat{\theta}_{ML})$. \square

The last method to build estimators uses an approach fundamentally different from the previous ones. The parameter of interest θ is considered as a random variable (as opposed to being a fixed value) and thus follows some distribution, called the prior distribution. The prior, as its name suggests, is chosen before obtaining the sample², which will serve as an “update” by applying the Bayes formula to the prior distribution to obtain the posterior distribution. The latter can finally be used to build estimators, which are called Bayes estimators.

Definition 2.3.7 (Bayes estimator). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample of size n , drawn i.i.d. from a distribution F depending on some parameter $\theta \in \Theta$, and let $f(\cdot; \theta)$ be the density (resp. probability mass) function of F . Moreover, let $\pi(\theta)$ be the prior distribution of θ . The posterior distribution of θ , $\pi(\cdot | \mathbf{X})$, then satisfies

$$\pi(\theta | \mathbf{X}) = \frac{f(\mathbf{X}; \theta) \pi(\theta)}{\int_{\Theta} f(\mathbf{X}; \theta) \pi(\theta) d\theta} \propto f(\mathbf{X}; \theta) \pi(\theta).$$

A Bayes estimator of θ is any estimator $\hat{\theta}$ obtained using the posterior distribution.

²In fact, the prior distribution serves as to quantify our knowledge on the possible values the parameter can take. In some cases, we could want to use what we call an uninformative prior which express the fact that we have no prior knowledge on the parameter. Also, the prior distribution can be chosen to obtain a specific posterior distribution.

Example 2.3.5. Let $\mathbf{X} = (X_1, \dots, X_n)$ a sample of size n such that $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Exp}(\lambda)$ and let the prior distribution of λ , $\pi(\lambda)$, be $\text{Exp}(\tau)$ with $\tau > 0$. We know the following about the posterior distribution

$$\begin{aligned}\pi(\lambda|\mathbf{X}) &\propto \left[\prod_{i=1}^n \lambda \exp(-\lambda X_i) \right] \times \tau \exp(-\lambda \tau) \mathbb{1}_{[0,+\infty[}(\lambda) \\ &\propto \tau \lambda^n \exp \left(-\lambda \left(\sum_{i=1}^n X_i + \tau \right) \right) \mathbb{1}_{[0,+\infty[}(\lambda),\end{aligned}$$

and the multiplicative constant is

$$\begin{aligned}\left(\int_{\mathbb{R}} f(\mathbf{X}; \lambda) \pi(\lambda) d\lambda \right)^{-1} &= \left(\int_{\mathbb{R}} \tau \lambda^n \exp \left(-\lambda \left(\sum_{i=1}^n X_i + \tau \right) \right) \mathbb{1}_{[0,+\infty[}(\lambda) d\lambda \right)^{-1} \\ &= \left(\tau \int_0^{+\infty} \lambda^n \exp \left(-\lambda \left(\sum_{i=1}^n X_i + \tau \right) \right) d\lambda \right)^{-1} \\ &= \left(\frac{\tau}{\left(\sum_{i=1}^n X_i + \tau \right)^{n+1}} \int_0^{+\infty} t^n \exp(-t) dt \right)^{-1} \\ &= \left(\frac{\tau}{\left(\sum_{i=1}^n X_i + \tau \right)^{n+1}} \Gamma(n+1) \right)^{-1} \\ &= \frac{\left(\sum_{i=1}^n X_i + \tau \right)^{n+1}}{\tau \Gamma(n+1)}.\end{aligned}$$

Therefore we obtain that

$$\pi(\lambda|\mathbf{X}) = \frac{\alpha^{n+1}}{\Gamma(n+1)} \lambda^n \exp(-\lambda \alpha) \mathbb{1}_{[0,+\infty[}(\lambda)$$

where $\alpha = \sum_{i=1}^n X_i + \tau$, which is the density function of the distribution $\Gamma \left(n+1, \sum_{i=1}^n X_i + \tau \right)$.

A possible Bayes estimator, $\hat{\lambda}$, for λ can be built by taking the mean of the posterior distribution, that is the average value of λ under its updated distribution based on the sample \mathbf{X} . For the mean of $\Gamma(n+1, \alpha)$, we have

$$\int_0^{+\infty} \lambda \frac{\alpha^{n+1}}{\Gamma(n+1)} \lambda^n \exp(-\lambda \alpha) d\lambda = \frac{\alpha^{n+1}}{\Gamma(n+1)} \int_0^{+\infty} \lambda^{n+1} \exp(-\lambda \alpha) d\lambda$$

$$\begin{aligned}
&= \frac{\alpha^{n+1}}{\Gamma(n+1)} \int_0^{+\infty} \left(\frac{t}{\alpha}\right)^{n+1} \exp(-t) \frac{1}{\alpha} dt \\
&= \frac{1}{\alpha \Gamma(n+1)} \int_0^{+\infty} t^{n+1} \exp(-t) dt \\
&= \frac{\Gamma(n+2)}{\alpha \Gamma(n+1)} \text{ by definition of } \Gamma \\
&= \frac{(n+1)\Gamma(n+1)}{\alpha \Gamma(n+1)} \\
&= \frac{n+1}{\alpha}.
\end{aligned}$$

As a result, our Bayes estimator for λ is given by

$$\hat{\lambda} = \frac{n+1}{\sum_{i=1}^n X_i + \tau}.$$

Remark 2.3.3. In order to use these three methods to build estimators, a hypothesis on the underlying distribution must be made. This is of course not an easy hypothesis to make, but a statistical analysis of the data and the use of statistical tests might be helpful.

Chapter 3

Inverse probability weighting

We can finally introduce the main subject of this thesis, the inverse probability weighting, although we will only describe the concept and give some properties, as we will go into further details regarding its robustness later. Note that this chapter is mainly based [25, 28, 31, 34, 37, 42, 53, 63].

3.1 The process of inverse probability weighting

In causal inference, our goal is to estimate the average causal effect of a treatment A on an outcome Y , which is given by $\mathbb{E}[Y^{a=1} - Y^{a=0}]$. In order to do that, a study is performed and data are collected on a sample of patients. However, as mentioned in the introduction, a randomized study is not always feasible (see Chapter 1), and therefore we might not have a representative sample of patients, treated or not. Due to this problem, the resulting estimated average treatment effect might be biased.

Now, let us consider the case where we have the conditional exchangeability of the potential outcomes and the treatment given a vector of covariates. This means that we assume the outcome of a patient in a world where he did (resp. he did not) receive the treatment is independent of his actual treatment given some vector of covariates. In a sense, this vector of covariates gives us enough information to infer on the potential outcome of the patient regardless of its actual treatment (due to conditional independence). Then, in this specific case, we can simulate, based on the sample of patients we have, a representative sample of the population. To obtain it, we consider our sample of patients in which each patient will be assigned a weight proportional to the inverse of the probability that this patient received the treatment he actually received; this probability is called the propensity score. Let us formally define it.

Definition 3.1.1 (Propensity score). For A the treatment variable and L a vector of covariates that take values in a space \mathcal{M} (e.g., \mathbb{R}^p). The propensity score is defined as

the function

$$e : \mathcal{M} \rightarrow [0, 1] : l \mapsto e(l) = \mathbb{P}(A = 1 | L = l).$$

Remark 3.1.1. In what follows, we will suppose that the treatment of a patient is independent of the treatment of other patients and independent of the covariates of other patients, which, if we write our sample as $\{(L_i, A_i, Y_i)\}_{i=1}^n$, implies that

$$\mathbb{P}(A_1, \dots, A_n | L_1, \dots, L_n) = \prod_{i=1}^n e(L_i)^{A_i} (1 - e(L_i))^{1-A_i}.$$

If we write L the vector of covariates, the fact that conditional exchangeability given L holds means that, within each stratum created by L , we have a homogeneous subsample in that each patient can be exchanged with another one regardless of their actual treatment. Therefore, by attributing a weight corresponding to the inverse of the propensity score, i.e., $1/\mathbb{P}(A = a | L = l)$, to each observation, we construct a sample that represents a population, called a pseudo-population, where each patient has both the outcome where he has taken the treatment and the outcome where he has not taken the treatment. Note that we do not obtain a sample where for each original observation, we have two observations with the two potential outcomes, we only obtain a representative sample of the distribution of these potential outcomes. Figure 3.1 illustrates this process, where the vector of covariates L is just a binary variable, A is the treatment variable and $P_{A=a}$ is the pseudo-population corresponding to the treatment a .

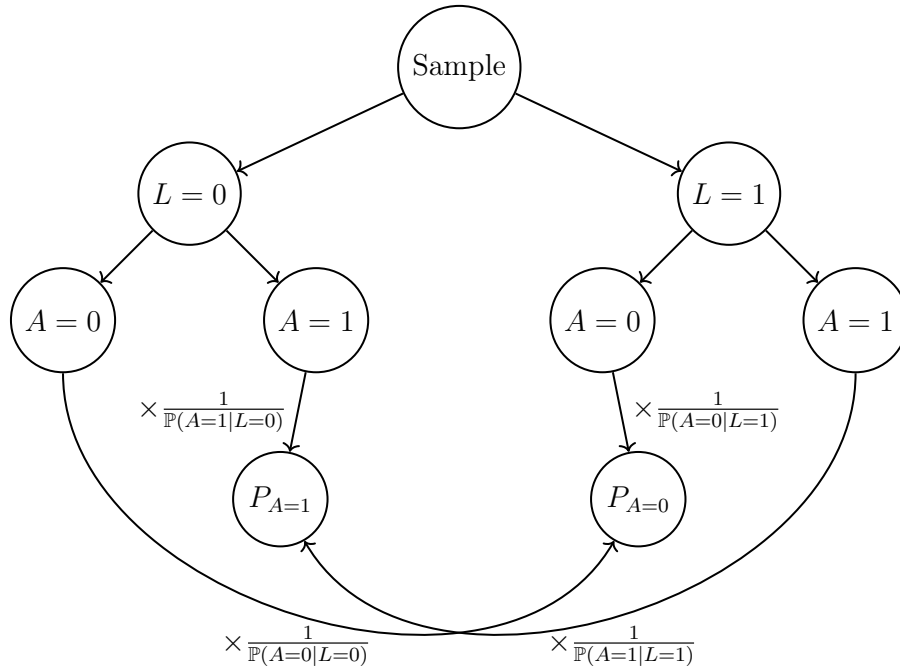


Figure 3.1: Graphical representation of the pseudo-population.

Remark 3.1.2. In the rest of this thesis, for the sake of simplicity, we will consider the following case:

- L , the vector of covariates, is a real random vector, that is, L takes values in \mathbb{R}^p ($p \in \mathbb{N}_0$).
- A , the treatment variable, is a binary variable, that is, A takes values in $0, 1$.
- Y , the outcome variable, is a real variable, that is, Y takes values in \mathbb{R} .

This weighting using the inverse of the propensity score is the key idea behind the inverse probability weighting method, which relies on the conditional exchangeability and can be formally defined as follows.

Definition 3.1.2 (Inverse probability weighting). Let A be the treatment variable, Y be the outcome variable, L be the vector of covariates such that conditional exchangeability ($Y^a \perp\!\!\!\perp A | L$) holds, then inverse probability weighting (IPW) is the process of attributing a weight

$$w_{a|l} = \frac{1}{\mathbb{P}(A = a | L = l)}$$

to each observation with the corresponding values for A and L . Furthermore, write

$$W^{A|L} = \frac{1}{\mathbb{P}(A|L)}$$

the random variable of the weighting.

Remark 3.1.3. Note that the event $\{L = l\}$ could be such that $\mathbb{P}(L = l) = 0$, in which case $\mathbb{P}(A = a | L = l)$ does not make sense using the usual definition (as in Definition A.2.20). However, it is still possible to define $\mathbb{P}(A = a | L = l)$ properly using

$$\mathbb{P}(A = a | L = l) = \mathbb{E} [\mathbb{1}_{\{A=a\}} | L] \Big|_{L=l}.$$

To simplify notation, we will write this as $\mathbb{E} [\mathbb{1}_{\{A=a\}} | L = l]$.

In Definition 3.1.2, we mentioned that $W^{A|L}$ is a random variable, however, we need to prove it. Let us use the following formal definition of the conditional probability as random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

$$\mathbb{P}(A = a | L(\cdot)) : \Omega \rightarrow [0, 1] : \omega \mapsto \mathbb{P}(A = a | L(\omega)) = \mathbb{E} [\mathbb{1}_{\{A=a\}} | L(\omega)],$$

which is a well-defined random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ (see Definition A.2.26). Using this definition, we obtain that $W^{A|L}$ is, indeed, a random variable as shown in the proposition below.

Proposition 3.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Moreover, let L and A be, respectively, a random vector and a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then the function*

$$W^{A|L} : \Omega \rightarrow [1, +\infty[: \omega \mapsto \frac{1}{\mathbb{P}(A = A(\omega) | L = L(\omega))}$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.

Proof. In order to prove that $W^{A|L}$ is a random variable, we must show that it is measurable, i.e., $\forall B \in \mathcal{B}([1, +\infty[), \{\omega \in \Omega | W^{A|L}(\omega) \in B\} \in \mathcal{F}$, where $\mathcal{B}(\mathcal{X})$ is the Borel σ -algebra of the set \mathcal{X} and is defined as

$$\sigma(\{X \subseteq \mathcal{X} : X \text{ is an open set}\}).$$

For $B \in \mathcal{B}([1, +\infty[)$ we have

$$\begin{aligned} \{\omega \in \Omega | W^{A|L}(\omega) \in B\} &= \left\{ \omega \in \Omega \left| \frac{1}{\mathbb{P}(A = A(\omega) | L = L(\omega))} \in B \right. \right\} \\ &= \left\{ \omega \in \Omega \left| \frac{1}{\mathbb{E}[\mathbb{1}_{\{A=A(\omega)\}} | L = L(\omega)]} \in B \right. \right\} \\ &= S_0 \cup S_1, \end{aligned}$$

where, for $a \in \{0, 1\}$, we define S_a as

$$S_a = \left\{ \omega \in \Omega \left| \frac{1}{\mathbb{E}[\mathbb{1}_{\{A=a\}} | L = L(\omega)]} \in B, A(\omega) = a \right. \right\}.$$

Moreover, we have

$$\begin{aligned} S_a &= \left\{ \omega \in \Omega \left| \frac{1}{\mathbb{E}[\mathbb{1}_{\{A=a\}} | L = L(\omega)]} \in B, A(\omega) = a \right. \right\} \\ &= \left\{ \omega \in \Omega \left| \frac{1}{\mathbb{E}[\mathbb{1}_{\{A=a\}} | L = L(\omega)]} \in B \right. \right\} \cap \{\omega \in \Omega | A(\omega) = a\} \\ &= \left\{ \omega \in \Omega \left| \mathbb{E}[\mathbb{1}_{\{A=a\}} | L = L(\omega)] \in \left(\frac{1}{\cdot}\right)^{-1}(B) \right. \right\} \cap \{\omega \in \Omega | A(\omega) = a\} \end{aligned}$$

with $\left(\frac{1}{\cdot}\right)^{-1}(B) = \left\{ b \left| \frac{1}{b} \in B \right. \right\}$. Now, for $a = 0$ or 1 , the function

$$\mathbb{E}[\mathbb{1}_{\{A=a\}} | L] : \Omega \rightarrow]0, 1] : \omega \mapsto \mathbb{E}[\mathbb{1}_{\{A=a\}} | L(\omega)]$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ (as explain before). Note that we have that $\mathbb{E}[\mathbb{1}_{\{A=a\}}|L]$ take values in $[0, 1]$ due to the positivity assumption, so we avoid any division by zero. Furthermore A is a random variable by hypothesis and we have that $(\frac{1}{\cdot})^{-1}(B) \in \mathcal{B}([0, 1])$ due to the continuity of the function $(\frac{1}{\cdot})$ on $[1, +\infty[$. Also, since \mathcal{F} is a σ -algebra, we have that \mathcal{F} contains finite intersections of its elements, as well as countable unions of them. Therefore, $S_a \in \mathcal{F}$, and we can conclude that $\forall B \in \mathcal{B}([1, +\infty[), \{\omega \in \Omega | W^{A|L}(\omega) \in B\} \in \mathcal{F}$. \square

In the following proofs, we will need to use some properties of conditional expectations (see [34]).

Proposition 3.1.2 (Properties of conditional expectations). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ two random variables, Z a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ two σ -algebras. We have the following properties:*

1. $\mathbb{E}[X|\mathcal{G}] \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.
2. If $X \perp\!\!\!\perp Y$, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
3. $\forall \alpha, \beta \in \mathbb{R}, \mathbb{E}[\alpha X + \beta Y|\mathcal{G}] = \alpha \mathbb{E}[X|\mathcal{G}] + \beta \mathbb{E}[Y|\mathcal{G}]$.
4. If Y is \mathcal{G} -measurable, then $\mathbb{E}[XY|\mathcal{G}] = Y \mathbb{E}[X|\mathcal{G}]$.
5. $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]$.
6. If $(X_n)_{n \in \mathbb{N}_0}$ is a sequence of positive random variables of $L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n \xrightarrow{a.s.} X$, then $\lim_{n \rightarrow +\infty} \mathbb{E}[X_n|Z] \stackrel{a.s.}{=} \mathbb{E}[X|Z]$.
7. $X \perp\!\!\!\perp Y|Z$ if and only if $\mathbb{E}[h(Y)|X, Z] \stackrel{a.s.}{=} \mathbb{E}[h(Y)|Z]$ for any Borel-measurable function $h : \mathbb{R} \rightarrow \mathbb{R}$ such that $h(Y) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.
8. If $X \perp\!\!\!\perp Y|Z$, then $\mathbb{E}[XY|Z] \stackrel{a.s.}{=} \mathbb{E}[X|Z] \mathbb{E}[Y|Z]$.

Proof. The first 6 properties are demonstrated in [25]. Concerning the second-to-last property, we have that if we suppose that $\mathbb{E}[h(Y)|X, Z] = \mathbb{E}[h(Y)|Z]$ almost surely for any Borel-measurable function $h : \mathbb{R} \rightarrow \mathbb{R}$, then we have $\forall A, B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
 \mathbb{P}(X \in A, Y \in B|Z) &= \mathbb{E}[\mathbb{1}_{\{X \in A, Y \in B\}}|Z] \text{ by definition} \\
 &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{X \in A\}} \mathbb{1}_{\{Y \in B\}}|X, Z]|Z] \text{ as } \sigma(Z) \subseteq \sigma(X, Z) \\
 &= \mathbb{E}[\mathbb{1}_{\{X \in A\}} \mathbb{E}[\mathbb{1}_{\{Y \in B\}}|X, Z]|Z] \text{ as } \mathbb{1}_{\{X \in A\}} \text{ is } \sigma(X, Z)\text{-measurable} \\
 &\stackrel{a.s.}{=} \mathbb{E}[\mathbb{1}_{\{X \in A\}} \mathbb{E}[\mathbb{1}_{\{Y \in B\}}|Z]|Z] \text{ by assumption with } h(Y) = \mathbb{1}_{\{Y \in B\}} \\
 &= \mathbb{E}[\mathbb{1}_{\{Y \in B\}}|Z] \mathbb{E}[\mathbb{1}_{\{X \in A\}}|Z] \text{ as } \mathbb{E}[\mathbb{1}_{\{Y \in B\}}|Z] \text{ is } \sigma(Z)\text{-measurable} \\
 &= \mathbb{P}(X \in A|Z) \mathbb{P}(Y \in B|Z),
 \end{aligned}$$

which signifies that $X \perp\!\!\!\perp Y|Z$. Conversely, if $X \perp\!\!\!\perp Y|Z$, then $\forall A, B, C \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_{\{X \in A, Z \in C\}} \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_{\{X \in A\}} \mathbb{1}_{\{Z \in C\}} \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_{\{X \in A\}} \mathbb{1}_{\{Z \in C\}} | Z \right] \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{E} \left[\mathbb{1}_{\{X \in A\}} | Z \right] \mathbb{1}_{\{Z \in C\}} \right] \text{ as } \mathbb{1}_{\{Z \in C\}}, \mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \text{ are } \sigma(Z)\text{-measurable} \\
&= \mathbb{E} \left[\mathbb{P}(X \in A | Z) \mathbb{P}(Y \in B | Z) \mathbb{1}_{\{Z \in C\}} \right] \text{ by definition} \\
&= \mathbb{E} \left[\mathbb{P}(X \in A, Y \in B | Z) \mathbb{1}_{\{Z \in C\}} \right] \text{ by assumption} \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{X \in A, Y \in B\}} | Z \right] \mathbb{1}_{\{Z \in C\}} \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{X \in A, Y \in B, Z \in C\}} | Z \right] \right] \text{ as } \mathbb{1}_{\{Z \in C\}} \text{ is } \sigma(Z)\text{-measurable} \\
&= \mathbb{P}(X \in A, Y \in B, Z \in C).
\end{aligned}$$

We just showed that

$$\mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_E \right] = \mathbb{P}(Y \in B, E),$$

for all $E \in \mathcal{C} = \{\{X \in A, Z \in C\} | A, C \in \mathcal{B}(\mathbb{R})\}$. Note that \mathcal{C} contains $\{X \in A\}$ and $\{Z \in C\}$ for all $A, C \in \mathcal{B}(\mathbb{R})$ as $\mathbb{R} \in \mathcal{B}(\mathbb{R})$. Furthermore, \mathcal{C} is a π -system and is such that $\sigma(\mathcal{C}) = \sigma(X, Z)$.

Now consider the following two measures on $(\Omega, \sigma(X, Z))$:

- $\mu : \sigma(X, Z) \rightarrow [0, 1] : D \mapsto \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_D \right]$.
- $\nu : \sigma(X, Z) \rightarrow [0, 1] : D \mapsto \mathbb{P}(Y \in B, D)$.

As shown above, they are equal on \mathcal{C} , and as a result, by Lemma A.2.1, they are equals on $\sigma(X, Z)$. So, we have that for any $D \in \sigma(X, Z)$,

$$\mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \mathbb{1}_D \right] = \mathbb{P}(Y \in B, D).$$

We also have that $\forall D \in \sigma(X, Z)$,

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | X, Z \right] \mathbb{1}_D \right] \\
&= \mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} \mathbb{1}_D \right] \text{ by definition of } \mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | X, Z \right] \\
&= \mathbb{P}(Y \in B, D).
\end{aligned}$$

Moreover, we have that $\mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right] \right]$ is $\sigma(X, Z)$ -measurable as it $\sigma(Z)$ -measurable and $\sigma(Z) \subseteq \sigma(X, Z)$. Therefore, by unicity (a.s.) of the conditional expectation, we have

$$\mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | X, Z \right] \stackrel{\text{a.s.}}{=} \mathbb{E} \left[\mathbb{1}_{\{Y \in B\}} | Z \right],$$

which implies that we have proved the claim for functions h of the form $\mathbb{1}_B$, with $B \in \mathcal{B}(\mathbb{R})$. Now, consider that $h : \mathbb{R} \rightarrow \mathbb{R}$ is a Borel-measurable simple function (see Definition A.2.29), that is, $\exists N \in \mathbb{N}, \alpha_1, \dots, \alpha_N \in \mathbb{R}, A_1, \dots, A_N \in \mathcal{B}(\mathbb{R})$ such that $i \neq j \Rightarrow A_i \cap A_j = \emptyset$ and

$$h : \mathbb{R} \rightarrow \mathbb{R} : y \mapsto \sum_{i=1}^N \alpha_i \mathbb{1}_{\{y \in A_i\}}.$$

We have that

$$\begin{aligned} \mathbb{E}[h(Y)|X, Z] &= \mathbb{E}\left[\sum_{i=1}^N \alpha_i \mathbb{1}_{\{Y \in A_i\}} \middle| X, Z\right] \\ &= \sum_{i=1}^N \alpha_i \mathbb{E}[\mathbb{1}_{\{Y \in A_i\}} | X, Z] \\ &\stackrel{\text{a.s.}}{=} \sum_{i=1}^N \alpha_i \mathbb{E}[\mathbb{1}_{\{Y \in A_i\}} | Z] \\ &= \mathbb{E}\left[\sum_{i=1}^N \alpha_i \mathbb{1}_{\{Y \in A_i\}} \middle| Z\right] \\ &= \mathbb{E}[h(Y) | Z]. \end{aligned}$$

In the case where h is a Borel-measurable function with positive values, there exists (see Proposition A.2.11) a sequence of positive simple function $(h_n)_{n \in \mathbb{N}_0}$ such that $h_n \rightarrow h$ pointwise and $0 \leq h_1 \leq h_2 \leq \dots \leq h$. Note that this last inequality implies that $h_n(Y) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ since $h(Y) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. We have that $\forall n \in \mathbb{N}_0$,

$$\mathbb{E}[h_n(Y)|X, Z] \stackrel{\text{a.s.}}{=} \mathbb{E}[h_n(Y)|Z],$$

and $h_n(Y) \xrightarrow{\text{a.s.}} h(Y)$. As a result,

$$\lim_{n \rightarrow +\infty} \mathbb{E}[h_n(Y)|X, Z] \stackrel{\text{a.s.}}{=} \mathbb{E}[h(Y)|X, Z] \stackrel{\text{a.s.}}{=} \mathbb{E}[h(Y)|Z] \stackrel{\text{a.s.}}{=} \lim_{n \rightarrow +\infty} \mathbb{E}[h_n(Y)|Z],$$

by unicity of the limit. The general case for a Borel-measurable function h is obtained by writing $h = h^+ - h^-$, where h^+ (resp. h^-) is the positive (resp. negative) part of h and using the precedent case.

Finally, for the last point of the proposition, using what precedes, we have

$$\begin{aligned} \mathbb{E}[XY|Z] &= \mathbb{E}[\mathbb{E}[XY|X, Z]|Z] \text{ as } \sigma(Z) \subseteq \sigma(X, Z) \\ &= \mathbb{E}[X\mathbb{E}[Y|X, Z]|Z] \text{ as } X \text{ is } \sigma(X, Z)\text{-measurable} \\ &= \mathbb{E}[X\mathbb{E}[Y|Z]|Z] \text{ by the property of Item 7} \\ &= \mathbb{E}[X|Z]\mathbb{E}[Y|Z] \text{ as } \mathbb{E}[Y|Z] \text{ is } \sigma(Z)\text{-measurable.} \end{aligned}$$

□

Since the estimation of the propensity score is the main problem we face in inverse probability weighting, it might be interesting to look at some of its properties. This propensity score is what we call a balancing score, as defined in the following proposition (see [53]).

Proposition 3.1.3. *Let A be the treatment variable and L be a vector of covariates. A Borel-measurable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ ($p \in \mathbb{N}_0$) is a balancing score, i.e.,*

$$A \perp\!\!\!\perp L|f(L)$$

if and only if there exists a measurable function $g : \mathbb{R} \rightarrow [0, 1]$ such that $e(L) \stackrel{\text{a.s.}}{=} g(f(L))$.

Proof. Suppose that there exists a measurable function $g : \mathbb{R} \rightarrow [0, 1]$ such that

$$e(L) \stackrel{\text{a.s.}}{=} g(f(L)),$$

and let us show that f is a balancing score, i.e.,

$$A \perp\!\!\!\perp L|f(L).$$

Let us start by obtaining that

$$e(L) \stackrel{\text{a.s.}}{=} \mathbb{P}(A = 1|f(L)). \quad (3.1)$$

In order to prove this equality, we need the following two properties:

1. $\sigma(f(L)) \subseteq \sigma(L)$.
2. $g(f(L))$ is $\sigma(f(L))$ -measurable.

First, an element of $\sigma(f(L))$ is of the form $\{\omega \in \Omega | f(L(\omega)) \in B\}$ for $B \in \mathcal{B}(\mathbb{R})$ and we have

$$\{\omega \in \Omega | f(L(\omega)) \in B\} = \{\omega \in \Omega | L(\omega) \in f^{-1}(B)\},$$

where $f^{-1}(B) \in \mathcal{B}(\mathbb{R}^p)$ since f is Borel-measurable. So, $\{\omega \in \Omega | f(L(\omega)) \in B\} = \{\omega \in \Omega | L(\omega) \in B'\}$ for some $B' \in \mathcal{B}(\mathbb{R}^p)$, which implies that it is an element of $\sigma(L)$. Therefore $\sigma(f(L)) \subseteq \sigma(L)$.

Secondly, since g is measurable, we have that if $B \in \mathcal{B}([0, 1])$, then $B' = g^{-1}(B) \in \mathcal{B}(\mathbb{R})$. This implies that

$$\begin{aligned} \{\omega \in \Omega | g(f(L(\omega))) \in B\} &= \{\omega \in \Omega | f(L(\omega)) \in g^{-1}(B)\} \\ &= \{\omega \in \Omega | f(L(\omega)) \in B'\} \in \sigma(f(L)). \end{aligned}$$

Therefore, we can conclude that $g(f(L))$ is $\sigma(f(L))$ -measurable.

Using these two properties, let us prove that Equation (3.1) holds. We have the following:

$$\begin{aligned}
\mathbb{P}(A = 1|f(L)) &= \mathbb{E}[\mathbb{1}_{\{A=1\}}|f(L)] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{A=1\}}|L] | f(L)] \text{ because } \sigma(f(L)) \subseteq \sigma(L) \\
&= \mathbb{E}[\mathbb{P}(A = 1|L) | f(L)] \\
&= \mathbb{E}[e(L)|f(L)] \\
&\stackrel{\text{a.s.}}{=} \mathbb{E}[g(f(L))|f(L)] \text{ as } e(L) \stackrel{\text{a.s.}}{=} g(f(L)) \\
&= g(f(L)) \text{ because } g(f(L)) \text{ is } \sigma(f(L))\text{-measurable} \\
&\stackrel{\text{a.s.}}{=} e(L)
\end{aligned}$$

Now, we have

$$\begin{aligned}
\mathbb{P}(A = 1|L, f(L)) &= \mathbb{E}[\mathbb{1}_{\{A=1\}}|L, f(L)] \\
&= \mathbb{E}[\mathbb{1}_{\{A=1\}}|L] \text{ because } \sigma(L, f(L)) = \sigma(L) \text{ as } \sigma(f(L)) \subseteq \sigma(L) \\
&= \mathbb{P}(A = 1|L) \\
&= e(L) \\
&\stackrel{\text{a.s.}}{=} \mathbb{P}(A = 1|f(L)) \text{ by Equation (3.1)}
\end{aligned}$$

As a result, we can conclude that A and L are conditionally independent given $f(L)$, because we showed that

$$\mathbb{E}[\mathbb{1}_{\{A=1\}}|L, f(L)] \stackrel{\text{a.s.}}{=} \mathbb{E}[\mathbb{1}_{\{A=1\}}|f(L)].$$

Conversely, let us suppose that f is a balancing score, i.e., $A \perp\!\!\!\perp L|f(L)$, and let us show that there exists a measurable function $g : \mathbb{R} \rightarrow [0, 1]$ such that

$$e(L) \stackrel{\text{a.s.}}{=} g(f(L)). \quad (3.2)$$

We have

$$\begin{aligned}
e(L) &= \mathbb{P}(A = 1|L) \\
&= \mathbb{P}(A = 1|L, f(L)) \text{ because } \sigma(L, f(L)) = \sigma(L) \\
&\stackrel{\text{a.s.}}{=} \mathbb{P}(A = 1|f(L)) \text{ because } A \perp\!\!\!\perp L|f(L) \\
&= \mathbb{E}[\mathbb{1}_{\{A=1\}}|f(L)] \\
&= g(f(L)),
\end{aligned}$$

where g is a Borel-measurable function. The existence of this function comes from Proposition A.2.10. \square

3.2 Inverse probability weighted mean

With IPW, we can introduce new estimators, such as the inverse probability weighted mean, which will be our main focus for the rest of this thesis.

Definition 3.2.1 (Inverse probability weighted mean). Let A be the treatment variable, Y be the outcome variable, L be a set of $p \in \mathbb{N}_0$ variables such that the conditional exchangeability holds ($Y^a \perp\!\!\!\perp A|L$), then the inverse probability weighted mean (IPWM) for the treatment $A = a$, μ_a^{IPW} , is given by taking the expectation of the outcome variable Y times its weight $W^{A|L}$ when $A = a$, i.e.,

$$\mu_a^{IPW} = \mathbb{E} [W^{A|L} \mathbb{1}_{\{A=a\}} Y] = \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y}{\mathbb{P}(A|L)} \right].$$

An interesting property of this estimator, and this is why it is defined this way, is that under certain assumptions it equals the potential mean $\mathbb{E}[Y^a]$ (see [31]).

Proposition 3.2.1. *Let A be the treatment variable, Y be the outcome variable, L be a set of $p \in \mathbb{N}_0$ variables such that the following assumptions hold:*

1. *Consistency: If $A = a$ then $Y^a = Y$*
2. *Conditional exchangeability: $Y^a \perp\!\!\!\perp A|L$*
3. *Positivity: $\forall a, l, \mathbb{P}(A = a|L = l) > 0$.*

Then we have that

$$\mu_a^{IPW} = \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y}{\mathbb{P}(A|L)} \right] = \mathbb{E}[Y^a].$$

Proof. We have the following

$$\begin{aligned} \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y}{\mathbb{P}(A|L)} \right] &= \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y^a}{\mathbb{P}(A|L)} \right] \text{ by consistency} \\ &= \mathbb{E} \left[\mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y^a}{\mathbb{P}(A|L)} \middle| L \right] \right] \text{ by properties of the conditional expectation} \\ &= \mathbb{E} \left[\mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}}}{\mathbb{P}(A|L)} \middle| L \right] \mathbb{E}[Y^a|L] \right] \text{ by conditional exchangeability} \\ &= \mathbb{E} [\mathbb{E}[Y^a|L]] \text{ because } \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}}}{\mathbb{P}(A|L)} \middle| L \right] = 1 \\ &= \mathbb{E}[Y^a] \text{ by properties of the conditional expectation,} \end{aligned} \tag{\#}$$

where Equation (#) is obtained by

$$\begin{aligned} \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}}}{\mathbb{P}(A|L)} \middle| L \right] &= \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}}}{\mathbb{P}(a|L)} \middle| L \right] \\ &= \frac{1}{\mathbb{P}(a|L)} \mathbb{E} [\mathbb{1}_{\{A=a\}} | L] \text{ because } \mathbb{P}(a|L) \text{ is } \sigma(L)\text{-measurable} \end{aligned}$$

$$\begin{aligned}
&= \frac{\mathbb{P}(a|L)}{\mathbb{P}(a|L)} \\
&= 1.
\end{aligned}$$

□

Remark 3.2.1. For the rest of this thesis, unless mentioned explicitly, we will suppose that the consistency, conditional exchangeability and positivity properties hold.

Before going any further, Let us first study the distribution of the variable $W^{A|L} \mathbb{1}_{\{A=a\}} Y$, as this knowledge will help us study its properties.

Proposition 3.2.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let (L, A, Y) be a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$, where A is the treatment variable, Y is the outcome variable and L is a vector of covariates. Let $a \in \{0, 1\}$, then we have, $\forall x \geq 0$,*

$$F_{W^{A|L} \mathbb{1}_{\{A=a\}} Y}(x) = \mathbb{P}(A = a, Y \geq 0, W^{a|L} Y \leq x) + \mathbb{P}(A = a, Y < 0) + \mathbb{P}(A = 1 - a),$$

and, $\forall x < 0$,

$$F_{W^{A|L} \mathbb{1}_{\{A=a\}} Y}(x) = \mathbb{P}(A = a, Y < 0, W^{a|L} Y \leq x).$$

Proof. We have the following

$$\begin{aligned}
F_{W^{A|L} \mathbb{1}_{\{A=a\}} Y}(x) &= \mathbb{P}(W^{A|L} \mathbb{1}_{\{A=a\}} Y \leq x) \\
&= \int_{\Omega} \mathbb{1}_{\{W^{A|L}(\omega) \mathbb{1}_{\{A(\omega)=a\}} Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega),
\end{aligned}$$

and let us write $A_a = \{\omega \in \Omega | A(\omega) = a\}$. Then $\Omega = A_a \cup A_{1-a}$ and

$$\begin{aligned}
&F_{W^{A|L} \mathbb{1}_{\{A=a\}} Y}(x) \\
&= \int_{A_a} \mathbb{1}_{\{W^{A|L}(\omega) \mathbb{1}_{\{A(\omega)=a\}} Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) + \int_{A_{1-a}} \mathbb{1}_{\{W^{A|L}(\omega) \mathbb{1}_{\{A(\omega)=a\}} Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) \\
&= \int_{A_a} \mathbb{1}_{\{W^{A|L}(\omega) Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) + \int_{A_{1-a}} \mathbb{1}_{\{0 \leq x\}}(\omega) d\mathbb{P}(\omega) \\
&= \begin{cases} \int_{A_a} \mathbb{1}_{\{W^{A|L}(\omega) Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) + \mathbb{P}(A = 1 - a) & \text{if } x \geq 0 \\ \int_{A_a} \mathbb{1}_{\{W^{A|L}(\omega) Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) & \text{otherwise} \end{cases}
\end{aligned}$$

Furthermore, let $Y_+ = \{\omega \in \Omega | Y(\omega) \geq 0\}$ and $Y_- = \{\omega \in \Omega | Y(\omega) < 0\}$, then $A_a = (A_a \cap Y_+) \cup (A_a \cap Y_-)$ and we obtain that

$$\begin{aligned}
&\int_{A_a} \mathbb{1}_{\{W^{A|L}(\omega) Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) \\
&= \int_{A_a} \mathbb{1}_{\{W^{a|L}(\omega) Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega)
\end{aligned}$$

$$\begin{aligned}
&= \int_{A_a \cap Y_+} \mathbb{1}_{\{W^{a|L}(\omega)Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) + \int_{A_a \cap Y_-} \mathbb{1}_{\{W^{a|L}(\omega)Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) \\
&= \begin{cases} \int_{A_a \cap Y_+} \mathbb{1}_{\{W^{a|L}(\omega)Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) + \mathbb{P}(A = a, Y < 0) & \text{if } x \geq 0 \\ \int_{A_a \cap Y_-} \mathbb{1}_{\{W^{a|L}(\omega)Y(\omega) \leq x\}}(\omega) d\mathbb{P}(\omega) & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathbb{P}(A = a, Y \geq 0, W^{a|L}Y \leq x) + \mathbb{P}(A = a, Y < 0) & \text{if } x \geq 0 \\ \mathbb{P}(A = a, Y < 0, W^{a|L}Y \leq x) & \text{otherwise} \end{cases},
\end{aligned}$$

which proves the statement. \square

3.3 Empirical estimator of the inverse probability weighted mean

An intuitive empirical estimator of the IPWM follows directly from Definition 3.2.1.

Definition 3.3.1 (Classical empirical estimator of the IPWM). Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . We construct the IPWM estimator as follows:

1. $\forall a, l$, we compute $\hat{\mathbb{P}}_n(A = a|L = l)$ an empirical estimation of $\mathbb{P}(A = a|L = l)$ (e.g., we use logistic regression to model $\mathbb{P}(A = a|L)$).
2. The IPWM estimator is given by

$$\hat{\mu}_{a,n}^{IPW} : (\mathbb{R}^p \times \{0, 1\} \times \mathbb{R})^n \rightarrow \mathbb{R} : \{(L_i, A_i, Y_i)\}_{i=1}^n \mapsto \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\hat{\mathbb{P}}_n(A_i|L_i)}.$$

Now, in order to provide some interesting results about the empirical estimator, we will consider, unless otherwise specified, that $\hat{\mathbb{P}}_n(A = a|L = l)$ is estimated using logistic regression (see [39]).

Remark 3.3.1. The logistic regression is the following process:

Let L be a vector of covariates of size $p \in \mathbb{N}_0$, $\pi = \mathbb{P}(A = 1|L)$ and consider the function

$$\text{logit} :]0, 1[\rightarrow \mathbb{R} : p \mapsto \ln \left(\frac{p}{1-p} \right).$$

Suppose that we have the following model

$$\text{logit}(\pi) = \beta^T L' \Leftrightarrow \pi = \frac{1}{1 + \exp(-\beta^T L')}$$

with $\beta \in \mathbb{R}^{p+1}$ and $L' = (1, L_1, \dots, L_p)$. If $\{(L_i, A_i)\}_{i=1}^n$ is a random sample of size n following this model, then the likelihood is obtained by

$$\begin{aligned} \mathcal{L}(\beta; \{(L_i, A_i)\}_{i=1}^n) &= \prod_{i=1}^n \mathbb{P}(A_i, L_i) \\ &= \prod_{i=1}^n \mathbb{P}(A_i | L_i) \mathbb{P}(L_i) \\ &= \prod_{i=1}^n \left(\frac{1}{1 + \exp(-\beta^T L_i')} \right)^{A_i} \left(1 - \frac{1}{1 + \exp(-\beta^T L_i')} \right)^{1-A_i} \mathbb{P}(L_i). \end{aligned}$$

Therefore, the maximum likelihood estimator, $\hat{\pi}_{ML}$, of π is obtained by taking $\hat{\beta}$ as the solution of

$$\begin{aligned} &\arg \max_{\beta \in \mathbb{R}^{p+1}} L(\beta; \{(L_i, A_i)\}_{i=1}^n) \\ &= \arg \max_{\beta \in \mathbb{R}^{p+1}} \prod_{i=1}^n \left(\frac{1}{1 + \exp(-\beta^T L_i')} \right)^{A_i} \left(1 - \frac{1}{1 + \exp(-\beta^T L_i')} \right)^{1-A_i} \\ &= \arg \min_{\beta \in \mathbb{R}^{p+1}} - \sum_{i=1}^n \left[A_i \ln \left(\frac{1}{1 + \exp(-\beta^T L_i')} \right) + (1 - A_i) \ln \left(1 - \frac{1}{1 + \exp(-\beta^T L_i')} \right) \right], \end{aligned}$$

which results in the estimator

$$\hat{\pi}_{ML} = \frac{1}{1 + \exp(-\hat{\beta}^T L')}.$$

Here, we use that

$$\mathbb{P}(A = a | L) = \left(\frac{1}{1 + \exp(-\beta^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\beta^T L')} \right)^{1-a},$$

which, we estimate by

$$\hat{\mathbb{P}}_n(A = a | L) = \left(\frac{1}{1 + \exp(-\hat{\beta}^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}^T L')} \right)^{1-a}.$$

Remark 3.3.2. There exists many other techniques to model $\mathbb{P}(A = 1 | L)$ such as a decision trees or a multilayer perceptron.

Concerning the IPWM estimator, to stress the fact that the estimator of β depends on the size of the sample, we will denote $\hat{\beta}_n$ as the resulting estimator of beta. As we suppose that $\hat{\mathbb{P}}_n(A_i = a | L_i)$ is estimated using a logistic regression, we have the following form for the IPWM estimator:

$$\hat{\mu}_{a,n}^{IPW} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L_i')} \right)^{A_i} \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L_i')} \right)^{1-A_i}},$$

where $L'_i = (1, (L_i)_1, \dots, (L_i)_d)$.

In order to show some properties of the IPWM estimator, we will need a way to approximate the expectation and variance of a function of variables (see [28, 63]).

Proposition 3.3.1. *Let $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^q$, $p, q \in \mathbb{N}_0$, be a mapping that is differentiable on a open set $O \subseteq \mathbb{R}^p$, and consider a random vector $X = (X_1, \dots, X_p)^T$ which takes values in O with $\mathbb{E}[X_i] = \mu_i \in \mathbb{R}$ for $i \in \{1, \dots, p\}$ such that $\mu = \mathbb{E}[X] = (\mu_1, \dots, \mu_p)^T \in O$. Then,*

$$\mathbb{E}[\varphi(X)] \approx \varphi(\mathbb{E}[X]),$$

and we have

$$\mathbb{V}[\varphi(X)] \approx \nabla \varphi(\mu) \mathbb{V}[X] \nabla \varphi(\mu)^T.$$

Proof. Let us start by writing φ by it Taylor expansion around μ , that is,

$$\varphi(X) = \varphi(\mu - (\mu - X)) = \varphi(\mu) + \nabla \varphi(\mu)(X - \mu) + o(\|X - \mu\|),$$

where

$$\nabla \varphi(\mu) = \begin{pmatrix} \frac{\partial}{\partial u_1} \varphi_1(u) \Big|_{u=\mu} & \cdots & \frac{\partial}{\partial u_p} \varphi_1(u) \Big|_{u=\mu} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial u_1} \varphi_q(u) \Big|_{u=\mu} & \cdots & \frac{\partial}{\partial u_p} \varphi_q(u) \Big|_{u=\mu} \end{pmatrix}.$$

As a result, we obtain that

$$\begin{aligned} \mathbb{E}[\varphi(X)] &= \mathbb{E}[\varphi(\mu) + \nabla \varphi(\mu)(X - \mu) + o(\|X - \mu\|)] \\ &= \varphi(\mu) + \nabla \varphi(\mu) \mathbb{E}[(X - \mu)] + \mathbb{E}[o(\|X - \mu\|)] \\ &= \varphi(\mathbb{E}[X]) + \mathbb{E}[o(\|X - \mu\|)] \text{ as } \mathbb{E}[(X - \mu)] = 0 \\ &\approx \varphi(\mathbb{E}[X]) \text{ discarding the term } \mathbb{E}[o(\|X - \mu\|)] \end{aligned}$$

For the variance, where we discard the term $o(\|X - \mu\|)$ in the Taylor expansion of φ around μ , we have

$$\begin{aligned} \mathbb{V}[\varphi(X)] &\approx \mathbb{V}[\varphi(\mu) + \nabla \varphi(\mu)(X - \mu)] \\ &= \nabla \varphi(\mu) \mathbb{V}[(X - \mu)] \nabla \varphi(\mu)^T \\ &= \nabla \varphi(\mu) \mathbb{V}[X] \nabla \varphi(\mu)^T. \end{aligned}$$

□

The first property of the IPWM estimator that we will examine is its bias.

Proposition 3.3.2 (Bias of $\hat{\mu}_{a,n}^{IPW}$). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . We have that*

$$\mathbb{E} [\hat{\mu}_{a,n}^{IPW}] - \mathbb{E} [Y^a] \approx \mathbb{E} [Y^a] \left(\frac{\mathbb{P}(A = a)}{\mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a} \right]} - 1 \right).$$

Moreover, if $\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a}$ is asymptotically unbiased, then so is $\hat{\mu}_{a,n}^{IPW}$ (approximately).

Proof. Let us start by observing that

$$\begin{aligned} \mathbb{E} [\hat{\mu}_{a,n}^{IPW}] &= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L'_i)} \right)^{A_i} \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L'_i)} \right)^{1-A_i}} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L'_i)} \right)^{A_i} \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L'_i)} \right)^{1-A_i}} \right] \\ &= \mathbb{E} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a}} \right], \end{aligned}$$

where $(L, A, Y) \sim F$, due to the i.i.d. property of the sample. Moreover, without Consider first the case $a = 1$. We then have

$$\begin{aligned} &\mathbb{E} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a}} \right] \\ &= \mathbb{E} \left[\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \right] = \mathbb{E} \left[\mathbb{E} \left[\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \middle| L \right] \right] \end{aligned}$$

$$\begin{aligned}
 &= \mathbb{E} \left[\frac{1}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \mathbb{E} [Y^{a=1} \mathbb{1}_{\{A=1\}} | L] \right] \\
 &= \mathbb{E} \left[\frac{\mathbb{E} [\mathbb{1}_{\{A=1\}} | L]}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \mathbb{E} [Y^{a=1} | L] \right] \text{ due to Equation (2.1)} \\
 &= \mathbb{E} \left[\frac{\mathbb{P}(A = 1 | L)}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \mathbb{E} [Y^{a=1} | L] \right] \\
 &\approx \mathbb{E} \left[\frac{\mathbb{P}(A = 1 | L)}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \right] \mathbb{E} [\mathbb{E} [Y^{a=1} | L]] \text{ by Proposition 3.3.1} \\
 &= \mathbb{E} \left[\frac{\mathbb{P}(A = 1 | L)}{\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)} \right] \mathbb{E} [Y^{a=1}] \\
 &\approx \frac{\mathbb{E} [\mathbb{P}(A = 1 | L)]}{\mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right) \right]} \mathbb{E} [Y^{a=1}] \text{ by Proposition 3.3.1} \\
 &= \frac{\mathbb{P}(A = 1)}{\mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right) \right]} \mathbb{E} [Y^{a=1}].
 \end{aligned}$$

The case $a = 0$ gives that

$$\mathbb{E} [\hat{\mu}_{0,n}^{IPW}] \approx \frac{\mathbb{P}(A = 0)}{\mathbb{E} \left[\left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right) \right]} \mathbb{E} [Y^{a=0}].$$

As a result, the bias of $\hat{\mu}_{a,n}^{IPW}$ is then (approximately)

$$\mathbb{E} [\hat{\mu}_{a,n}^{IPW}] - \mathbb{E} [Y^a] \approx \mathbb{E} [Y^a] \left(\frac{\mathbb{P}(A = a)}{\mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a} \right]} - 1 \right).$$

If we suppose that

$$\mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right)^{1-a} \right] \xrightarrow{n \rightarrow +\infty} \mathbb{P}(A = a),$$

which means that the estimator of the propensity score is asymptotically unbiased,¹ we have that $\hat{\mu}_{a,n}^{IPW}$ is also asymptotically unbiased. \square

Remark 3.3.3. We made several approximations in the proof of Proposition 3.3.2 using Proposition 3.3.1. However, due to the form of this estimator and the context (we try to estimate $\mathbb{E}[Y^a]$, where Y^a is not a variable we can obtain), there are not many options in order to provide exact information about the bias of this estimator. As such, it might not be possible to avoid such approximations.

Next, let us look at the asymptotic variance of the IPWM estimator.

Proposition 3.3.3 (Asymptotic variance of $\hat{\mu}_{a,n}^{IPW}$). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . Moreover, suppose that the following hold:*

1. *There exists $N > 0$ such that*

- $\mathbb{V} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right) \right] < N,$
- $\mathbb{V}[Y^a] < N.$

2. *There exists $\delta > 0$ such that*

- $\delta < \mathbb{E} \left[\left(\frac{1}{1 + \exp(-\hat{\beta}_n^T L')} \right) \right] < +\infty,$
- $\delta < \mathbb{P}(A = a) < +\infty,$
- $\delta < \mathbb{E}[Y^a] < +\infty.$

Then, $\hat{\mu}_{a,n}^{IPW}$ has asymptotically zero variance (approximately).

¹This supposes that the estimator obtained by logistic regression converges in mean to the mean of the conditional probability $\mathbb{P}(A = a|L)$, that is, $\mathbb{E}[\mathbb{P}(A = a|L)] = \mathbb{P}(A = a)$. This is a natural hypothesis we need to make.

Proof. Let us start by observing that

$$\begin{aligned}
\mathbb{V} [\hat{\mu}_{a,n}^{IPW}] &= \mathbb{V} \left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L'_i)} \right)^{A_i} \left(1 - \frac{1}{1+\exp(-\hat{\beta}_n^T L'_i)} \right)^{1-A_i}} \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V} \left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L'_i)} \right)^{A_i} \left(1 - \frac{1}{1+\exp(-\hat{\beta}_n^T L'_i)} \right)^{1-A_i}} \right] \\
&= \frac{1}{n} \mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)^{1-a}} \right],
\end{aligned}$$

where $(L, A, Y) \sim F$, due to the i.i.d. property of the sample. Moreover, without loss of generality, assume that $a = 1$. We then have

$$\begin{aligned}
&\frac{1}{n} \mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)^a \left(1 - \frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)^{1-a}} \right] \\
&= \frac{1}{n} \mathbb{V} \left[\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)} \right] \\
&= \frac{1}{n} \mathbb{E} \left[\left(\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)} - \mathbb{E} \left[\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)} \right] \right)^2 \right] \\
&\approx \frac{1}{n} \mathbb{E} \left[\left(\frac{Y^{a=1} \mathbb{1}_{\{A=1\}}}{\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right)} - \frac{\mathbb{P}(A=1)}{\mathbb{E} \left[\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')} \right) \right]} \mathbb{E}[Y^{a=1}] \right)^2 \right] \quad \text{by Proposition 3.3.2.}
\end{aligned}$$

Now, let us consider the function

$$\varphi : \mathbb{R}^3 \rightarrow \mathbb{R} : (x, y, z) \mapsto \frac{xy}{z},$$

which is differentiable on $\mathbb{R}^3 \setminus \{(x, y, z) \in \mathbb{R}^3 | z = 0\}$ with Jacobian given by

$$\nabla \varphi(x, y, z) = \left(\frac{y}{z} \quad \frac{x}{z} \quad -\frac{xy}{z^2} \right).$$

To simplify the notation, let us write

$$A = \frac{\mathbb{P}(A=1)}{\mathbb{E}\left[\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')}\right)\right]}, B = \frac{\mathbb{E}[Y^{a=1}]}{\mathbb{E}\left[\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')}\right)\right]}, C = -\frac{\mathbb{E}[Y^{a=1}] \mathbb{P}(A=1)}{\mathbb{E}\left[\left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')}\right)\right]^2},$$

and

$$U = \left(\frac{1}{1+\exp(-\hat{\beta}_n^T L')}\right), V = Y^{a=1}, W = \mathbb{1}_{\{A=1\}}.$$

Note that we have

$$A = \frac{\mathbb{E}[W]}{\mathbb{E}[U]}, B = \frac{\mathbb{E}[V]}{\mathbb{E}[U]} \text{ and } C = -\frac{\mathbb{E}[V] \mathbb{E}[W]}{\mathbb{E}[U]^2}.$$

We can use the Taylor expansion of φ to obtain

$$\begin{aligned} & \frac{1}{n} \mathbb{E} \left[\left(\frac{VW}{U} - \frac{\mathbb{E}[V] \mathbb{E}[W]}{\mathbb{E}[U]} \right)^2 \right] \\ & \approx \frac{1}{n} \mathbb{E} [(A(V - \mathbb{E}[V]) + B(W - \mathbb{E}[W]) - C(U - \mathbb{E}[U]))^2] \\ & = \frac{1}{n} (A^2 \mathbb{V}[V] + B^2 \mathbb{V}[W] + C^2 \mathbb{V}[U] + 2AB \mathfrak{Cov}[V, W] - 2AC \mathfrak{Cov}[V, U] - 2BC \mathfrak{Cov}[W, U]) \\ & = \frac{1}{n} \left(\frac{\mathbb{E}[V] \mathbb{E}[W]}{\mathbb{E}[U]} \right)^2 \left(\frac{\mathbb{V}[V]}{\mathbb{E}[V]^2} + \frac{\mathbb{V}[W]}{\mathbb{E}[W]^2} + \frac{\mathbb{V}[U]}{\mathbb{E}[U]^2} + 2 \frac{\mathfrak{Cov}[V, W]}{\mathbb{E}[V] \mathbb{E}[W]} - 2 \frac{\mathfrak{Cov}[V, U]}{\mathbb{E}[V] \mathbb{E}[U]} \right. \\ & \quad \left. - 2 \frac{\mathfrak{Cov}[W, U]}{\mathbb{E}[W] \mathbb{E}[U]} \right). \end{aligned}$$

The case $a = 0$ results in a similar expression.

Finally, since $\mathbb{V}[U], \mathbb{V}[V] < N$ for some $N > 0$ and $\delta < \mathbb{E}[V], \mathbb{E}[W], \mathbb{E}[U] < +\infty$ for some $\delta > 0$, we see that $\hat{\mu}_{1,n}^{IPW}$ has asymptotically zero variance (the expression is bounded by M/n for a constant $M > 0$). The same conclusion holds for $\hat{\mu}_{0,n}^{IPW}$. \square

Unfortunately, we cannot easily obtain properties concerning the consistency of $\hat{\mu}_{a,n}^{IPW}$ unless stronger assumptions are made. Furthermore, due to the form of the approximated variance derived in Proposition 3.3.3, the MSE is not usable in practice. Once again, these problems come from the nature of our estimator, which despite its clear proximity to the classical mean estimator, does not have all its nice properties.

3.4 Properties of the IPWM estimator when the propensity score is known

To conclude this chapter, we will consider the case where we know exactly the form of $\mathbb{P}(A = a|L)$ ($a \in \{0, 1\}$), i.e., we know a Borel-measurable function $f_a : \mathbb{R}^p \rightarrow [0, 1]$ such

that

$$\mathbb{P}(A = a|L) \stackrel{\text{a.s.}}{=} f_a(L).$$

The practical application of this specific case is more restrictive as it requires to know the propensity score, and might be unrealistic depending on the context. However, it allows us to obtain exact results concerning the bias and asymptotic variance of the IPWM estimator.

Proposition 3.4.1 (Bias of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A = a|L)$ is known). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . If*

$$\mathbb{P}(A = a|L) \stackrel{\text{a.s.}}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), then

$$\mathbb{E}[\hat{\mu}_{a,n}^{IPW}] - \mathbb{E}[Y^a] = 0.$$

Proof. We have

$$\begin{aligned} \mathbb{E}[\hat{\mu}_{a,n}^{IPW}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A = a|L_i)}\right] \\ &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{f_a(L_i)}\right] \text{ as } \mathbb{P}(A = a|L_i) \stackrel{\text{a.s.}}{=} f_a(L_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{f_a(L_i)}\right] \\ &= \mathbb{E}\left[\frac{Y \mathbb{1}_{\{A=a\}}}{f_a(L)}\right] \text{ as } (L, A, Y) \sim F, \text{ since } (L_i, A_i, Y_i) \stackrel{i.i.d.}{\sim} F \\ &= \mathbb{E}\left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A = a|L)}\right] \text{ as } \mathbb{P}(A = a|L) \stackrel{\text{a.s.}}{=} f_a(L) \\ &= \mathbb{E}[Y^a] \text{ by Proposition 3.2.1} \end{aligned}$$

□

Proposition 3.4.2 (Asymptotic variance of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A = a|L)$ is known). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . If*

$$\mathbb{P}(A = a|L) \stackrel{\text{a.s.}}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), then $\hat{\mu}_{a,n}^{IPW}$ has asymptotically zero variance.

Proof. We have

$$\begin{aligned}
\mathbb{V} [\hat{\mu}_{a,n}^{IPW}] &= \mathbb{V} \left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A=a|L_i)} \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V} \left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A=a|L_i)} \right] \text{ as } (L_i, A_i, Y_i) \text{ are independent} \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A=a|L)} \right] \text{ as } (L, A, Y) \sim F, \text{ since } (L_i, A_i, Y_i) \stackrel{i.i.d.}{\sim} F \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{f_a(L)} \right] \text{ as } \mathbb{P}(A=a|L) \stackrel{a.s.}{=} f_a(L) \\
&= \frac{1}{n^2} \sum_{i=1}^n V \text{ with } V \text{ the variance of } \frac{Y \mathbb{1}_{\{A=a\}}}{f_a(L)} \\
&= \frac{V}{n} \xrightarrow{n \rightarrow +\infty} 0.
\end{aligned}$$

□

From that we directly obtain the mean squared error (see Definition 2.3.3).

Proposition 3.4.3 (Mean squared error of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A=a|L)$ is known). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . If*

$$\mathbb{P}(A=a|L) \stackrel{a.s.}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), then

$$\text{MSE}(\hat{\mu}_{a,n}^{IPW}) = \frac{\mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A=a|L)} \right]}{n}.$$

Proof. This is a direct consequence of Proposition 3.4.1 and the fact that

$$\mathbb{V} [\hat{\mu}_{a,n}^{IPW}] = \frac{\mathbb{V} \left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A=a|L)} \right]}{n},$$

as developed in the proof of Proposition 3.4.2. □

In addition to the exact result of the variance and bias, we also obtain that the estimator of the inverse probability weighted mean converges almost surely to $\mathbb{E}[Y^a]$ if $\mathbb{E}[|Y^a|] < +\infty$.

Proposition 3.4.4 (Convergence almost surely of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A = a|L)$ is known).
 Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i .
 If

$$\mathbb{P}(A = a|L) \stackrel{a.s.}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), and if

$$\mathbb{E}[|Y^a|] < +\infty,$$

then

$$\hat{\mu}_{a,n}^{IPW} \xrightarrow{a.s.} \mathbb{E}[Y^a].$$

Proof. We have that $\forall i \in \{1, \dots, n\}$,

$$\mathbb{E}\left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right] = \mathbb{E}[Y^a],$$

as $(L_i, A_i, Y_i) \stackrel{i.i.d.}{\sim} F$, and

$$\begin{aligned} \mathbb{E}\left[\left|\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right|\right] &= \mathbb{E}\left[\left|\frac{Y_i^a \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right|\right] \text{ by consistency} \\ &= \mathbb{E}\left[\mathbb{E}\left[\left|\frac{Y_i^a \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right| \middle| L_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[|Y_i^a| \frac{\mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)} \middle| L_i\right]\right] \text{ as } \frac{\mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)} \geq 0 \\ &= \mathbb{E}\left[\mathbb{E}[|Y_i^a| | L_i] \mathbb{E}\left[\frac{\mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)} \middle| L_i\right]\right] \text{ by conditional exchangeability} \\ &= \mathbb{E}[\mathbb{E}[|Y_i^a| | L_i]] \text{ as } \mathbb{E}\left[\frac{\mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)} \middle| L_i\right] = 1 \\ &= \mathbb{E}[|Y_i^a|] \\ &= \mathbb{E}[|Y^a|] \text{ as } (L_i, A_i, Y_i) \stackrel{i.i.d.}{\sim} F \\ &< +\infty. \end{aligned}$$

Therefore, the claim follows from the SLLN (see Theorem 2.3.1). \square

Corollary 3.4.1 (Convergence in probability of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A = a|L)$ is known).
 Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i .
 If

$$\mathbb{P}(A = a|L) \stackrel{a.s.}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), and if

$$\mathbb{E}[|Y^a|] < +\infty,$$

then

$$\hat{\mu}_{a,n}^{IPW} \xrightarrow{\mathbb{P}} \mathbb{E}[Y^a].$$

Proof. This directly follows from Proposition 3.4.4, as the convergence almost surely implies the convergence in probability. \square

Finally, we obtain the asymptotic distribution of the estimator of the inverse probability weighted mean as a result of the central limit theorem.

Proposition 3.4.5 (Asymptotic distribution of $\hat{\mu}_{a,n}^{IPW}$ when $\mathbb{P}(A = a|L)$ is known). *Let $\{(L_i, A_i, Y_i)\}_{i=1}^n$ be a sample of size n i.i.d. according to a distribution F , where L_i is a vector of covariates, A_i is the treatment and Y_i is the outcome for an individual i . If*

$$\mathbb{P}(A = a|L) \stackrel{a.s.}{=} f_a(L)$$

for some known Borel-measurable function $f_a : \mathbb{R}^p \rightarrow \mathbb{R}$ ($a \in \{0, 1\}$), then

$$\sqrt{n}(\hat{\mu}_{a,n}^{IPW} - \mathbb{E}[Y^a]) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}\left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A = a|L)}\right]\right).$$

Proof. Since $(L_i, A_i, Y_i) \stackrel{i.i.d.}{\sim} F$, we have that $\forall i \in \{1, \dots, n\}$,

$$\mathbb{E}\left[\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right] = \mathbb{E}[Y^a],$$

and $\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)} \stackrel{i.i.d.}{\sim} F'$ for a distribution F' . Therefore, by Theorem A.2.1, we have

$$\sqrt{n}(\hat{\mu}_{a,n}^{IPW} - \mathbb{E}[Y^a]) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}\left[\frac{Y \mathbb{1}_{\{A=a\}}}{\mathbb{P}(A = a|L)}\right]\right).$$

\square

As we can see, when the propensity score is known, we find the classical results of the mean estimator. This is a direct consequence of the fact that $\hat{\mu}_{a,n}^{IPW}$ is a mean estimator of $\mathbb{E}[Y^a]$ when considering the sample $\left\{\frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\mathbb{P}(A_i = a|L_i)}\right\}_{i=1}^n$.

Chapter 4

Statistical Robustness

4.1 Concept of robustness

Before we can discuss the robustness of the inverse probability weighting estimator, we first need to introduce the concept of robustness itself. Note that this chapter is mainly based on [27, 29, 30]

The robustness of an estimator is the capability of the estimator to still provide a good approximation for a parameter while having a dataset with extreme values such as outliers, corrupted data, and so on. The main idea is to quantify how well an estimator performs when we give it a dataset with outliers that may not have been detected before.

This is a major field of statistics that is being studied heavily due to the fact that an estimator that is robust can perform very well in situations where an estimator that is not robust will perform poorly (at least we hope it does). We will not go into too many details, as we are mainly interested in the robustness of the inverse probability weighting estimator. There exist many techniques to study the robustness of a classical estimator in the context of inference. In this thesis we will define some of them together with a number of their properties, as they will also be used in the context of causal inference and applied to the estimator we are interested in, that is, the inverse probability weighted mean.

Before we begin to talk about the different tools, we need to introduce a concept that will be useful for what follows, namely, the concept of a statistical functional.

Definition 4.1.1 (Statistical functional). Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space, where \mathcal{X} is some space, $\mathcal{B}(\mathcal{X})$ is a σ -algebra on \mathcal{X} , and let \mathcal{F} be the space (or a subspace) of all probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A statistical functional is a function

$$T : \mathcal{F} \rightarrow \mathcal{Y},$$

where \mathcal{Y} is an Hilbert space

Remark 4.1.1. We want \mathcal{Y} to be a Hilbert space in order to have a real vector space equipped with a norm, as this will be useful for the definition of the influence function of the statistical functional.

Proposition 4.1.1. *Let \mathcal{F} be the space of all probability measure on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Then \mathcal{F} is a convex space, that is, $\forall \mathcal{P}, \mathcal{Q} \in \mathcal{F}$ and $\forall \lambda \in [0, 1]$,*

$$(1 - \lambda)\mathcal{P} + \lambda\mathcal{Q} \in \mathcal{F}.$$

Proof. We thus need to prove that $\forall \mathcal{P}, \mathcal{Q} \in \mathcal{F}$ and $\forall \lambda \in [0, 1]$, $\mathcal{W} = (1 - \lambda)\mathcal{P} + \lambda\mathcal{Q}$ is a probability measure:

1. $\forall B \in \mathcal{B}(\mathcal{X}), \mathcal{W}(B) \geq 0 \Leftrightarrow \forall B \in \mathcal{B}(\mathcal{X}), (1 - \lambda)\mathcal{P}(B) + \lambda\mathcal{Q}(B) \geq 0$, which is true since $(1 - \lambda), \lambda, \mathcal{P}(B)$ and $\mathcal{Q}(B)$ are all positive.
2. $\mathcal{W}(\emptyset) = 0 \Leftrightarrow (1 - \lambda)\mathcal{P}(\emptyset) + \lambda\mathcal{Q}(\emptyset) = 0$, which is true because $\mathcal{P}(\emptyset) = \mathcal{Q}(\emptyset) = 0$.
3. $\forall (B_i)_{i \in \mathbb{N}}$ where $B_i \in \mathcal{B}(\mathcal{X})$ and $B_i \cap B_j = \emptyset \forall i \neq j$, we have

$$\begin{aligned} \mathcal{W}\left(\bigcup_{i=1}^{+\infty} B_i\right) &= \sum_{i=1}^{+\infty} \mathcal{W}(B_i) \\ &\Leftrightarrow (1 - \lambda)\mathcal{P}\left(\bigcup_{i=1}^{+\infty} B_i\right) + \lambda\mathcal{Q}\left(\bigcup_{i=1}^{+\infty} B_i\right) = \sum_{i=1}^{+\infty} ((1 - \lambda)\mathcal{P}(B_i) + \lambda\mathcal{Q}(B_i)) \end{aligned}$$

which is true due to the fact that

$$\mathcal{P}\left(\bigcup_{i=1}^{+\infty} B_i\right) = \sum_{i=1}^{+\infty} \mathcal{P}(B_i) \text{ and } \mathcal{Q}\left(\bigcup_{i=1}^{+\infty} B_i\right) = \sum_{i=1}^{+\infty} \mathcal{Q}(B_i).$$

4. $\mathcal{W}(\mathcal{X}) = 1 \Leftrightarrow (1 - \lambda)\mathcal{P}(\mathcal{X}) + \lambda\mathcal{Q}(\mathcal{X}) = 1$, which is true as $\mathcal{P}(\mathcal{X}) = \mathcal{Q}(\mathcal{X}) = 1$ and $(1 - \lambda) + \lambda = 1$.

□

Remark 4.1.2. For the sake of clarity and conciseness, we will occasionally use Bochner integrals implicitly throughout this chapter (see Appendix A.2.8). This choice allows for more streamlined notation and more readable proofs. In particular, this approach will be used when dealing with integrals of the form

$$\int_{\mathcal{X}} x dF(x),$$

where $\mathcal{X} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_0$, and F is a probability distribution.

Let us give two examples of classical statistical functionals.

Example 4.1.1. The mean functional is defined by

$$T_{mean} : \mathcal{F} \rightarrow \mathbb{R}^p : F \mapsto \mathbb{E}_F[X].$$

Note that for this functional \mathcal{F} is a subspace of all probability measures, as this functional requires that $\mathbb{E}_F[X]$ exists.

Example 4.1.2. The statistical functional for the covariance matrix is defined by

$$T_{var} : \mathcal{F} \rightarrow \mathcal{S}^p : F \mapsto \mathbb{E}_F [(X - \mathbb{E}_F[X])(X - \mathbb{E}_F[X])^T]$$

where \mathcal{S}^p is the space of positive semi-definite matrices of dimension p .

Remark 4.1.3. For the rest of this chapter, we will consider the case where $\mathcal{Y} = \mathbb{R}^p$, $p \in \mathbb{N}_0$.

Let us mention quickly that a point of interest is to determine whether an estimator is Fisher-consistent.

Definition 4.1.2 (Fisher-consistency). Let $\mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a family of probability measures associated with a parametric distribution depending on a parameter $\theta \in \mathbb{R}^p$ (e.g., the set of probability measure associated with a normal distribution $\mathcal{N}(\mu, \sigma^2)$). A statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^p$ is said to be Fisher-consistent with respect to \mathcal{F}_Θ (or simply Fisher-consistent if no confusion is possible) if $\forall \theta \in \Theta$,

$$T(F_\theta) = \theta.$$

In other words, a statistical functional is Fisher-consistent if it returns the real parameter of all parametric distributions in a specific family (this supposes that this functional's main goal is to compute this parameter).

Example 4.1.3. Considering the mean functional and the family

$$\mathcal{F}_\mathbb{R} = \{F_{(\mu)} | F_{(\mu)} \text{ is a probability measure associated with } \mathcal{N}(\mu, 1), \mu \in \mathbb{R}\},$$

we have, $\forall \mu \in \mathbb{R}$,

$$T_{mean}(F_\mu) = \mathbb{E}_{F_\mu}[X] = \mu.$$

Therefore, the mean functional is Fisher-consistent with respect to this family of probability measures.

4.2 Breakdown point

The first measure that might appear in the context of robustness is what we call the breakdown point.

4.2.1 Empirical breakdown point

We will first define the empirical version of the breakdown point, which will lead to the theoretical version.

Definition 4.2.1 (Empirical breakdown point). Suppose that $T_n : \mathbb{R}^{q \times n} \rightarrow \mathbb{R}^p$ ($q, p \in \mathbb{N}_0$) is some estimator (e.g., a mean or median estimator) and $\mathbf{x} = (x_1, \dots, x_n)$, with $x_i \in \mathbb{R}^q$, is a sample of size n , then the empirical breakdown point of T_n is defined by

$$\varepsilon(T_n, \mathbf{x}) = \frac{1}{n} \min \left\{ m \left| \sup_{\mathbf{x}^{(m)}} \|T_n(\mathbf{x}) - T_n(\mathbf{x}^{(m)})\| = +\infty \right. \right\}$$

where the supremum is taken on the set of samples $\mathbf{x}^{(m)}$ obtained by replacing m observations of \mathbf{x} by arbitrary values.

Basically, this empirical measure tells us how many “badly placed” observations are needed to result in divergence to infinity of the estimator.

Remark 4.2.1. If T_n is such that $\forall m \in \{0, \dots, n\}, \sup_{\mathbf{x}^{(m)}} \|T_n(\mathbf{x}) - T_n(\mathbf{x}^{(m)})\| < +\infty$, then the minimum does not exist. In such cases, we will consider $\varepsilon(T_n, \mathbf{x})$ to be 1.

Let us compute the empirical breakdown point of three classical estimators in order to exemplify this definition.

Example 4.2.1. For $\mathbf{X} = (X_1, \dots, X_n)$, the empirical breakdown point of the mean estimator

$$T_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i$$

is $\varepsilon(T_n, \mathbf{x}) = \frac{1}{n}$ as only one observation, that tends towards infinity, causes the estimator to also tend towards infinity.

Example 4.2.2. For $\mathbf{X} = (X_1, \dots, X_n)$, the empirical breakdown point of the variance estimator

$$T_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2$$

is $\varepsilon(T_n, \mathbf{x}) = \frac{1}{n}$ for the same reason as the mean estimator.

Example 4.2.3. For $\mathbf{X} = (X_1, \dots, X_n)$, the empirical breakdown point of the median estimator

$$T_n(\mathbf{X}) = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2} \left(X_{(\frac{n}{2})} + X_{(\frac{n+1}{2})} \right) & \text{if } n \text{ is even} \end{cases},$$

where $X_{(k)}$ is the k -th order statistic, is $\varepsilon(T_n, \mathbf{x}) = \lceil \frac{n}{2} \rceil \frac{1}{n}$ as we need to alter the half of the observations (or half plus one depending on the parity of n) to cause the estimator to diverge to infinity.

4.2.2 Theoretical breakdown point

In order to introduce the theoretical version, we will need a metric on \mathcal{F} (the set of probability measure on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$). A commonly used metric is Kolmogorov's metric.

Definition 4.2.2 (Kolmogorov's metric). Kolmogorov's metric is defined by

$$d_K : \mathcal{F}^2 \rightarrow [0, +\infty[: (\mathcal{P}, \mathcal{Q}) \mapsto \sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}}(x)|$$

where $F_{\mathcal{P}}$ (resp. $F_{\mathcal{Q}}$) is the cumulative distribution function associated with the probability measure \mathcal{P} (resp. \mathcal{Q}).

Proposition 4.2.1. *The Kolmogorov's metric is a metric.*

Proof. In order to prove the claim, we need to prove the following:

1. $\forall \mathcal{P}, \mathcal{Q} \in \mathcal{F}, d_K(\mathcal{P}, \mathcal{Q}) \geq 0$, which is true since $\sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}}(x)| \geq 0$.
2. $\forall \mathcal{P}, \mathcal{Q} \in \mathcal{F}, d_K(\mathcal{P}, \mathcal{Q}) = d_K(\mathcal{Q}, \mathcal{P})$, which is true since $\sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}}(x)| = \sup_{x \in \mathbb{R}^p} |F_{\mathcal{Q}}(x) - F_{\mathcal{P}}(x)|$.
3. $\forall \mathcal{P}, \mathcal{Q} \in \mathcal{F}, d_K(\mathcal{P}, \mathcal{Q}) = 0 \Leftrightarrow \mathcal{P} = \mathcal{Q}$. This is obtained by the fact that $\sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}}(x)| = 0 \Leftrightarrow \forall x \in \mathbb{R}^p, F_{\mathcal{P}}(x) = F_{\mathcal{Q}}(x)$ and that the cumulative distribution function characterizes the distribution.
4. $\forall \mathcal{P}, \mathcal{Q}, \mathcal{R} \in \mathcal{F}, d_K(\mathcal{P}, \mathcal{Q}) \leq d_K(\mathcal{P}, \mathcal{R}) + d_K(\mathcal{R}, \mathcal{Q})$. This is always true as

$$\begin{aligned} \sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}}(x)| &= \sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{R}}(x) + F_{\mathcal{R}}(x) - F_{\mathcal{Q}}(x)| \\ &\leq \sup_{x \in \mathbb{R}^p} (|F_{\mathcal{P}}(x) - F_{\mathcal{R}}(x)| + |F_{\mathcal{R}}(x) - F_{\mathcal{Q}}(x)|) \\ &\leq \sup_{x \in \mathbb{R}^p} |F_{\mathcal{P}}(x) - F_{\mathcal{R}}(x)| + \sup_{x \in \mathbb{R}^p} |F_{\mathcal{R}}(x) - F_{\mathcal{Q}}(x)|. \end{aligned}$$

□

Remark 4.2.2. Note that $d_K(\mathcal{P}, \mathcal{Q}) \in [0, 1]$ for any pair of distributions $(\mathcal{P}, \mathcal{Q}) \in \mathcal{F}^2$.

We can now define the theoretical breakdown point.

Definition 4.2.3 (Theoretical breakdown point). Let $T : \mathcal{F} \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a statistical functional and $\mathcal{P} \in \mathcal{F}$. Then the breakdown point of T at \mathcal{P} with respect to Kolmogorov's metric is defined as

$$\varepsilon^*(T, \mathcal{P}, d_K) = \inf \left\{ \varepsilon > 0 \left| \sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \varepsilon} \|T(\mathcal{P}) - T(\mathcal{Q})\| = +\infty \right. \right\}.$$

The idea behind this theoretical measure is to tell us what is the smallest “deviation” from our original distribution (in the sense of the distance based on Kolmogorov's metric) we need to make our statistical functional tend towards infinity.

Remark 4.2.3. If T is such that

$$\inf \left\{ \varepsilon > 0 \left| \sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \varepsilon} \|T(\mathcal{P}) - T(\mathcal{Q})\| = +\infty \right. \right\},$$

does not exists, then we will consider $\varepsilon^*(T, \mathcal{P}, d_K)$ to be 1.

Remark 4.2.4. In [30], the authors provide another definition of the breakdown point based on the distance of Prohorov, defined by

$$d_P : \mathcal{F}^2 \rightarrow [0, +\infty[: (\mathcal{P}, \mathcal{Q}) \mapsto \inf \{ \varepsilon > 0 | \forall A \in \mathcal{B}(\mathcal{X}), \mathcal{P}(A) \leq \mathcal{Q}(A^\varepsilon) + \varepsilon \},$$

where $A^\varepsilon = B(A, \varepsilon) = \{x \in \mathcal{X} | \exists a \in A : \|a - x\|_{\mathcal{X}} < \varepsilon\}$, and $\|\cdot\|_{\mathcal{X}}$ is some norm on \mathcal{X} . That is, A^ε is the set of all elements of \mathcal{X} that are at a distance strictly lower than ε from A (in the corresponding space). In this case, the theoretical breakdown point of an estimator T_n is defined as

$$\varepsilon^*(T_n, \mathcal{P}, d_P) = \sup \left\{ \varepsilon \leq 1 \left| \exists K \subseteq \mathbb{R}^p \text{ compact} : d_P(\mathcal{P}, \mathcal{Q}) < \varepsilon \Rightarrow \mathcal{Q}(\{T_n \in K\}) \xrightarrow{n \rightarrow +\infty} 1 \right. \right\}.$$

Nonetheless, the two definitions provide similar robustness information as they come from the same idea.

Let us now mention two results about the theoretical breakdown point concerning the statistical functionals of the mean and the median.

Proposition 4.2.2. *The theoretical breakdown point of the statistical functional of the mean*

$$T_{mean} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \mathbb{E}_F[X]$$

at \mathcal{P} is given by

$$\varepsilon^*(T_{mean}, \mathcal{P}, d_K) = 0.$$

Proof. If $1 > \varepsilon > 0$ and $N \in \mathbb{N}_0$ are fixed (we suppose that $\mathbb{E}_{\mathcal{P}}[X]$ is finite), let us consider the following distribution

$$\mathcal{P}_\varepsilon = \left(1 - \frac{\varepsilon}{2}\right) \mathcal{P} + \frac{\varepsilon}{2} \Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N},$$

where $\Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N}$ is the Dirac distribution at $\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N$.

We obtain that

$$\begin{aligned} d_K(\mathcal{P}, \mathcal{P}_\varepsilon) &= \sup_{x \in \mathbb{R}} |F_{\mathcal{P}}(x) - F_{\mathcal{P}_\varepsilon}(x)| \\ &= \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^x d\mathcal{P}(t) - \int_{-\infty}^x d\mathcal{P}_\varepsilon(t) \right| \\ &= \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^x d\mathcal{P}(t) - \left(1 - \frac{\varepsilon}{2}\right) \int_{-\infty}^x d\mathcal{P}(t) - \frac{\varepsilon}{2} \int_{-\infty}^x d\Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N}(t) \right| \\ &= \frac{\varepsilon}{2} \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^x d\mathcal{P}(t) - \int_{-\infty}^x d\Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N}(t) \right| \\ &= \frac{\varepsilon}{2} d_K(\mathcal{P}, \Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N}) \\ &\leq \frac{\varepsilon}{2} \text{ because } d_K(\mathcal{P}, \mathcal{Q}) \leq 1 \text{ for all } \mathcal{P}, \mathcal{Q} \in \mathcal{F} \\ &< \varepsilon. \end{aligned}$$

We also have that

$$\begin{aligned} \|T_{mean}(\mathcal{P}) - T_{mean}(\mathcal{P}_\varepsilon)\| &= |\mathbb{E}_{\mathcal{P}}[X] - \mathbb{E}_{\mathcal{P}_\varepsilon}[X]| \\ &= \left| \int_{\mathbb{R}} x d\mathcal{P}(t) - \left(1 - \frac{\varepsilon}{2}\right) \int_{\mathbb{R}} x d\mathcal{P}(t) - \frac{\varepsilon}{2} \int_{\mathbb{R}} x d\Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N}(t) \right| \\ &= \left| \frac{\varepsilon}{2} \mathbb{E}_{\mathcal{P}}[X] - \frac{\varepsilon}{2} \left(\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N \right) \right| \\ &= N. \end{aligned}$$

So we have showed that for any $1 > \varepsilon > 0$, we can find a distribution \mathcal{P}_ε such that $d_K(\mathcal{P}, \mathcal{P}_\varepsilon) < \varepsilon$ and $\|T_{mean}(\mathcal{P}) - T_{mean}(\mathcal{P}_\varepsilon)\| \geq N$, for arbitrary large N . Therefore, by taking N that tends towards $+\infty$, we have that

$$\sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \varepsilon} \|T_{mean}(\mathcal{P}) - T_{mean}(\mathcal{Q})\| = +\infty.$$

Since ε is arbitrarily chosen, we can conclude that $\varepsilon^*(T_{mean}, \mathcal{P}, d_K) = 0$ for all $\mathcal{P} \in \mathcal{F}$. \square

Remark 4.2.5. This proof can be adapted to the multivariate case ($T_{mean} : \mathcal{F} \rightarrow \mathbb{R}^p : F \mapsto \mathbb{E}_F[X]$ with $\mathbb{E}_F[X]$ a vector in \mathbb{R}^p) by considering the multivariate Dirac distribution $\Delta_{\mathbb{E}_{\mathcal{P}}[X] + \frac{2}{\varepsilon}N\mathbf{1}^p}$, where $\mathbf{1}^p$ is a vector of size p such that each component is $\frac{1}{\sqrt{p}}$.

Proposition 4.2.3. *The theoretical breakdown point of the statistical functional of the median*

$$T_{med} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \inf \left\{ t \in \mathbb{R} \left| \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} dF(w) \geq \frac{1}{2} \right. \right\}$$

at \mathcal{P} is given by

$$\varepsilon^*(T_{med}, \mathcal{P}, d_K) = \frac{1}{2}.$$

Proof. First, let us show that for $\frac{1}{2} > \eta > 0$, we have that

$$\sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \frac{1}{2} - \eta} \|T_{med}(\mathcal{P}) - T_{med}(\mathcal{Q})\| < +\infty.$$

Indeed, if we let $\alpha = T_{med}(\mathcal{P})$, which is finite, and $\beta = T_{med}(\mathcal{Q})$, then necessarily $\exists N \in \mathbb{N}$ that does not depend on the distribution \mathcal{Q} (which satisfies $d_K(\mathcal{P}, \mathcal{Q}) < \frac{1}{2} - \eta$) such that $|\beta| < N$.

If it was not the case, we would be able to find a sequence of distributions $(\mathcal{Q}_n)_{n \in \mathbb{N}}$ such that $\forall n \in \mathbb{N}$, $d_K(\mathcal{P}, \mathcal{Q}_n) < \frac{1}{2} - \eta$ and, for $\beta_n = T_{med}(\mathcal{Q}_n)$, $|\beta_n| > n$. Let us consider the case where $\beta_n \rightarrow +\infty$, the other case can be demonstrated analogously. Then, if we let

$$\gamma = \inf \left\{ t \in \mathbb{R} \left| \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} d\mathcal{P}(w) \geq 1 - \frac{\eta}{2} \right. \right\},$$

which is finite, we would have that

$$\begin{aligned} d_K(\mathcal{P}, \mathcal{Q}_n) &= \sup_{x \in \mathbb{R}} |F_{\mathcal{P}}(x) - F_{\mathcal{Q}_n}(x)| \\ &= \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^x d\mathcal{P}(t) - \int_{-\infty}^x d\mathcal{Q}_n(t) \right| \\ &\geq \left| \int_{-\infty}^{\gamma} d\mathcal{P}(t) - \int_{-\infty}^{\gamma} d\mathcal{Q}_n(t) \right| \\ &= \int_{-\infty}^{\gamma} d\mathcal{P}(t) - \int_{-\infty}^{\gamma} d\mathcal{Q}_n(t) \text{ because } \beta_n > \gamma \text{ for } n \text{ sufficiently large} \\ &\geq 1 - \frac{\eta}{2} - \int_{-\infty}^{\gamma} d\mathcal{Q}_n(t) \\ &\geq 1 - \frac{\eta}{2} - \frac{1}{2} \\ &= \frac{1}{2} - \frac{\eta}{2} \\ &> \frac{1}{2} - \eta, \end{aligned}$$

which contradicts the hypothesis.

Now, we can show that

$$\sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \frac{1}{2}} \|T_{med}(\mathcal{P}) - T_{med}(\mathcal{Q})\| = +\infty.$$

Let us fix $\frac{1}{2} > \varepsilon > 0$, $N \in \mathbb{N}_0$ and consider the following distribution

$$\mathcal{P}_\varepsilon = \left(\frac{1}{2} - \frac{\varepsilon}{2}\right) \mathcal{P} + \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \Delta_N.$$

Then we have that

$$\begin{aligned} d_K(\mathcal{P}, \mathcal{P}_\varepsilon) &= \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^x d\mathcal{P}(t) - \left(\frac{1}{2} - \frac{\varepsilon}{2}\right) \int_{-\infty}^x d\mathcal{P}(t) - \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \int_{-\infty}^x d\Delta_N(t) \right| \\ &= \sup_{x \in \mathbb{R}} \left| \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \int_{-\infty}^x d\mathcal{P}(t) - \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \int_{-\infty}^x d\Delta_N(t) \right| \\ &= \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) d_K(\mathcal{P}, \Delta_N) \\ &\leq \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \\ &< \left(\frac{1}{2} + \varepsilon\right). \end{aligned}$$

Moreover, we obtain that

$$\begin{aligned} T_{med}(\mathcal{P}_\varepsilon) &= \inf \left\{ t \in \mathbb{R} \left| \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} d\mathcal{P}_\varepsilon(w) \geq \frac{1}{2} \right. \right\} \\ &= \inf \left\{ t \in \mathbb{R} \left| \underbrace{\left(\frac{1}{2} - \frac{\varepsilon}{2}\right) \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} d\mathcal{P}(w)}_{\leq \frac{1}{2} - \frac{\varepsilon}{2}} + \underbrace{\left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} d\Delta_N(w)}_{=0 \text{ if } t < N \text{ and } \left(\frac{1}{2} + \frac{\varepsilon}{2}\right) \text{ otherwise}} \geq \frac{1}{2} \right. \right\} \\ &\geq N. \end{aligned}$$

So for any $\frac{1}{2} > \varepsilon > 0$, we can find a distribution \mathcal{P}_ε such that $d_K(\mathcal{P}, \mathcal{P}_\varepsilon) < \frac{1}{2} + \varepsilon$ and that $\|T_{med}(\mathcal{P}) - T_{med}(\mathcal{P}_\varepsilon)\| \geq N - \alpha$ for arbitrarily large N . So for any $\frac{1}{2} > \varepsilon > 0$, we have

$$\sup_{\mathcal{Q} \in \mathcal{F} : d_K(\mathcal{P}, \mathcal{Q}) < \frac{1}{2} + \varepsilon} \|T_{med}(\mathcal{P}) - T_{med}(\mathcal{Q})\| = +\infty$$

and, therefore, we obtain that, for any distribution $\mathcal{P} \in \mathcal{F}$, $\varepsilon^*(T_{med}, \mathcal{P}, d_K) = \frac{1}{2}$. \square

As we have seen, the median is more robust than the mean, as both the empirical and the theoretical breakdown point are greater for the median.

Remark 4.2.6. Since the definition of the empirical breakdown point is the minimal proportion of corrupted observations needed to make the estimator diverge, in some cases the empirical breakdown point will converge almost surely to the theoretical breakdown point, which is the smallest deviation (that is comprised between 0 and 1) from the original distribution needed to make the statistical functional diverge. For example, this is the case for the mean and median estimators.

4.3 Influence function

Let us now introduce another tool that will give information about the robustness of an estimator and that is complementary with the breakdown point: the influence function (or sensitivity curve).

4.3.1 Empirical influence function

Once again, we will first define the empirical version and then generalize the idea to obtain a theoretical version.

Definition 4.3.1 (Empirical influence function). Let $T_n : \mathbb{R}^{q \times n} \rightarrow \mathbb{R}^p$ be some estimator and let $\mathbf{x} = (x_1, \dots, x_n)$, with $x_i \in \mathbb{R}^q$, be a sample of size n , the empirical influence function (EIF) of T_n at $x \in \mathbb{R}^q$ is defined by

$$\text{EIF}(x; T_n, \mathbf{x}) = \frac{T_{n+1}(\mathbf{x} \cup \{x\}) - T_n(\mathbf{x})}{\frac{1}{n+1}}.$$

The empirical influence function measures/quantifies “how much” a new observation x can affect the estimator T_n .

Let us compute the empirical influence function of the mean, variance and median estimators.

Example 4.3.1. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample, and let

$$T_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

be the mean estimator. Then we have that for $\mathbf{x} = (x_1, \dots, x_n)$

$$\begin{aligned} \text{EIF}(x; T_n, \mathbf{x}) &= \frac{T_{n+1}(\mathbf{x} \cup \{x\}) - T_n(\mathbf{x})}{\frac{1}{n+1}} \\ &= (n+1) \left(\frac{1}{n+1} \left(\sum_{i=1}^n x_i + x \right) - \bar{\mathbf{x}}_n \right) \end{aligned}$$

$$\begin{aligned}
&= (n+1) \left(\frac{1}{n+1}x - \frac{1}{n+1}\bar{\mathbf{x}}_n \right) \\
&= x - \bar{\mathbf{x}}_n.
\end{aligned}$$

Example 4.3.2. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample, and let

$$T_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 = S_n^2$$

be the variance estimator. Then we have that for $\mathbf{x} = (x_1, \dots, x_n)$

$$\begin{aligned}
\text{EIF}(x; T_n, \mathbf{x}) &= \frac{T_{n+1}(\mathbf{x} \cup \{x\}) - T_n(\mathbf{x})}{\frac{1}{n+1}} \\
&= (n+1) \left(\frac{1}{n+1} \left(\sum_{i=1}^n x_i^2 + x^2 \right) - \left(\frac{1}{n+1} \sum_{i=1}^n x_i + \frac{x}{n+1} \right)^2 - \frac{1}{n} \sum_{i=1}^n x_i^2 + \bar{\mathbf{x}}_n^2 \right) \\
&= (n+1) \left(\frac{-1}{n(n+1)} \sum_{i=1}^n x_i^2 + \frac{x^2}{n+1} - \left(\frac{n}{n+1} \bar{\mathbf{x}}_n + \frac{x}{n+1} \right)^2 + \bar{\mathbf{x}}_n^2 \right) \\
&= (n+1) \left(\frac{-1}{n(n+1)} \sum_{i=1}^n x_i^2 + \frac{nx^2}{(n+1)^2} - \frac{2n}{(n+1)^2} \bar{\mathbf{x}}_n x + \frac{2n+1}{(n+1)^2} \bar{\mathbf{x}}_n^2 \right) \\
&= \frac{nx^2}{n+1} - \frac{2n}{n+1} \bar{\mathbf{x}}_n x + \frac{2n+1}{n+1} \bar{\mathbf{x}}_n^2 - \frac{1}{n} \sum_{i=1}^n x_i^2 \\
&= \frac{n}{n+1} (x^2 - 2\bar{\mathbf{x}}_n x + \bar{\mathbf{x}}_n^2) - \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{\mathbf{x}}_n^2 \right) \\
&= \frac{n}{n+1} (x - \bar{\mathbf{x}}_n)^2 - s_n^2.
\end{aligned}$$

Example 4.3.3. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample, and let

$$T_n(\mathbf{X}) = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2} \left(X_{(\frac{n}{2})} + X_{(\frac{n+1}{2})} \right) & \text{if } n \text{ is even} \end{cases}$$

be the median estimator. Then we have that for $\mathbf{x} = (x_1, \dots, x_n)$

$$\begin{aligned}
&\text{EIF}(x; T_n, \mathbf{x}) \\
&= \frac{T_{n+1}(\mathbf{x} \cup \{x\}) - T_n(\mathbf{x})}{\frac{1}{n+1}} \\
&= \begin{cases} (n+1) \left(\frac{1}{2} x_{(\frac{n+2}{2})} - \frac{1}{2} x_{(\frac{n+1}{2})} \right) & \text{if } n \text{ is odd and } x > x_{(\frac{n+2}{2})} \text{ or } x < x_{(\frac{n+1}{2})} \\ (n+1) \left(\frac{1}{2} x - \frac{1}{2} x_{(\frac{n+1}{2})} \right) & \text{if } n \text{ is odd and } x_{(\frac{n+1}{2})} \leq x \leq x_{(\frac{n+2}{2})} \\ (n+1) \left(x_{(\frac{n+2}{2})} - \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})} \right) \right) & \text{if } n \text{ is even and } x > x_{(\frac{n+2}{2})} \text{ or } x < x_{(\frac{n+1}{2})} \\ (n+1) \left(x - \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})} \right) \right) & \text{if } n \text{ is even and } x_{(\frac{n+1}{2})} \leq x \leq x_{(\frac{n+2}{2})} \end{cases}
\end{aligned}$$

From these examples, we can conclude that the mean estimator and the variance estimator are not robust as their EIF are not bounded. In contrast, the EIF of the median estimator is bounded, which proves again that the median estimator is robust.

4.3.2 The notion of derivation

Before introducing the theoretical influence function, let us discuss the notion of derivation, which will be linked to the influence function. Note that this section is based on [2, 3, 8, 16, 66].

The basic idea behind all the different notions of a derivative is to approximate a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ at some point $x \in \mathcal{X}$ by a mapping $A : \mathcal{X} \rightarrow \mathcal{Y}$ that has specific properties. In [2, 3], \mathcal{X} and \mathcal{Y} are considered to be linear topological spaces. Let us define the concept of topological space as in [8, 66].

Definition 4.3.2 (Power set). Let \mathcal{X} be a set, the power set of \mathcal{X} , noted $\wp(\mathcal{X})$ (or $2^{\mathcal{X}}$), is defined as

$$\wp(\mathcal{X}) = \{\mathcal{S} | \mathcal{S} \subseteq \mathcal{X}\}.$$

Definition 4.3.3 (Topological space). Let \mathcal{X} be some set and consider $\mathcal{N} : \mathcal{X} \rightarrow \wp(\wp(\mathcal{X})) : x \mapsto \mathcal{N}(x)$, where $\mathcal{N}(x)$ is a non-empty set of subsets of \mathcal{X} . The couple $(\mathcal{X}, \mathcal{N})$ is called a topological space if, $\forall x \in \mathcal{X}$, the following conditions hold:

1. If $N \in \mathcal{N}(x)$ then $x \in N$.
2. If $\mathcal{S} \subseteq \mathcal{X}$ such that $\exists N \in \mathcal{N}(x)$ with $N \subseteq \mathcal{S}$ then $\mathcal{S} \in \mathcal{N}(x)$.
3. If $N, N' \in \mathcal{N}(x)$ then $N \cap N' \in \mathcal{N}(x)$.
4. If $N \in \mathcal{N}(x)$ then $\exists M \in \mathcal{N}(x)$ such that $\forall m \in M, N \in \mathcal{N}(m)$.

We also need to define the concepts of continuity of a mapping between two topological spaces and the product of topological spaces.

Definition 4.3.4 (Continuity). Let $(\mathcal{X}, \mathcal{N})$ and $(\mathcal{Y}, \mathcal{M})$ be two topological spaces and let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a mapping, then f is said to be continuous if $\forall x \in \mathcal{X}, \forall M \in \mathcal{M}(f(x))$, we have $f^{-1}(M) \in \mathcal{N}(x)$.

Definition 4.3.5 (Product of topological spaces). Let $(\mathcal{X}, \mathcal{N})$ and $(\mathcal{Y}, \mathcal{M})$ be two topological spaces. The product topological space of $(\mathcal{X}, \mathcal{N})$ and $(\mathcal{Y}, \mathcal{M})$ is the topological

space $(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ where \mathcal{T} is defined by

$$\mathcal{T} : \mathcal{X} \times \mathcal{Y} \rightarrow \wp(\wp(\mathcal{X} \times \mathcal{Y})) : (x, y) \mapsto \{\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y} \mid \exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq \mathcal{S}\}.$$

Definition 4.3.6 (Interior of a set). Let $(\mathcal{X}, \mathcal{N})$ be a topological space and let $\mathcal{S} \subseteq \mathcal{X}$. The interior of \mathcal{S} , written \mathcal{S}° , is the set of all $x \in \mathcal{X}$ such that $\mathcal{S} \in \mathcal{N}(x)$.

Definition 4.3.7 (Open set). Let $(\mathcal{X}, \mathcal{N})$ be a topological space and let $\mathcal{S} \subseteq \mathcal{X}$. \mathcal{S} is said to be an open set if $\forall x \in \mathcal{S}, \mathcal{S} \in \mathcal{N}(x)$.

Proposition 4.3.1. *Let $(\mathcal{X}, \mathcal{N})$ be a topological space and let $\mathcal{S} \subseteq \mathcal{X}$. Then $\mathcal{S}^\circ \subseteq \mathcal{S}$ and $\forall s \in \mathcal{S}^\circ, \mathcal{S}^\circ \in \mathcal{N}(s)$.*

Proof. The case where $\mathcal{S}^\circ = \emptyset$ is trivial. Now, consider $x \in \mathcal{S}^\circ \neq \emptyset$ and $N \in \mathcal{N}(x)$ such that $\forall n \in N, \mathcal{S} \in \mathcal{N}(n)$. We have, by definition, that $x \in \mathcal{S}$. Moreover, $N \subseteq \mathcal{S}^\circ$, which implies that $\mathcal{S}^\circ \in \mathcal{N}(x)$. \square

Proposition 4.3.2. *Let $(\mathcal{X}, \mathcal{N})$ be a topological space and let $\mathcal{S} \subseteq \mathcal{X}$. Then \mathcal{S} is an open set if and only if $\mathcal{S} = \mathcal{S}^\circ$.*

Proof. If \mathcal{S} is an open set, then $\forall x \in \mathcal{S}, \mathcal{S} \in \mathcal{N}(x)$, which is the definition of \mathcal{S}° , so $\mathcal{S} = \mathcal{S}^\circ$. Inversely, if $\mathcal{S} = \mathcal{S}^\circ$, then $\forall x \in \mathcal{S}, \mathcal{S}^\circ = \mathcal{S} \in \mathcal{N}(x)$, therefore, \mathcal{S} is an open set. \square

Let us prove that the space defined in Definition 4.3.5 is indeed a topological space.

Proposition 4.3.3. *Let $(\mathcal{X}, \mathcal{N})$ and $(\mathcal{Y}, \mathcal{M})$ be two topological spaces and let \mathcal{T} be defined as*

$$\mathcal{T} : \mathcal{X} \times \mathcal{Y} \rightarrow \wp(\wp(\mathcal{X} \times \mathcal{Y})) : (x, y) \mapsto \{\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y} \mid \exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq \mathcal{S}\}.$$

Then $(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ is a topological space.

Proof. First, it is clear that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \mathcal{T}(x, y) \neq \emptyset$ since $\mathcal{T}(x, y)$ contains all the elements of the form $N \times M$ where $N \in \mathcal{N}(x)$ and $M \in \mathcal{M}(y)$.

Furthermore, $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, if $T \in \mathcal{T}(x, y)$, then $(x, y) \in T$ because by definition $\exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq T$ with $x \in N, y \in M$.

Now, if $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y}$ such that $\exists T \in \mathcal{T}(x, y)$ with $T \subseteq \mathcal{S}$, then we have that $\mathcal{S} \in \mathcal{T}(x, y)$. This is due to the fact that if $T \in \mathcal{T}(x, y)$ then, by definition, $\exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq T$, and that $N \times M \subseteq \mathcal{S}$, therefore $\mathcal{S} \in \mathcal{T}(x, y)$.

Let us now show that if $T, T' \in \mathcal{T}(x, y)$, then we have $T \cap T' \in \mathcal{T}(x, y)$. Again, by definition, $\exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq T$ and $\exists N' \in \mathcal{N}(x), M' \in \mathcal{M}(y) : N' \times M' \subseteq T'$. Moreover, $(N \times M) \cap (N' \times M') \subseteq T \cap T'$ and $(N \times M) \cap (N' \times M') = (N \cap N') \times (M \cap M')$ with $N \cap N' \in \mathcal{N}(x)$ and $M \cap M' \in \mathcal{M}(y)$.

Finally, let us prove that if $T \in \mathcal{T}(x, y)$ then $\exists S \in \mathcal{T}(x, y)$ such that $\forall (s, s') \in S, T \in \mathcal{T}(s, s')$. As before, by definition, $\exists N \in \mathcal{N}(x), M \in \mathcal{M}(y) : N \times M \subseteq T$ and let us take $S = N^o \times M^o$, which is in $\mathcal{T}(x, y)$. We have that $\forall (n, m) \in N^o \times M^o, T \in \mathcal{T}(n, m)$ because $N^o \in \mathcal{N}(n), M^o \in \mathcal{M}(m)$ and $N^o \times M^o \subseteq N \times M \subseteq T$. \square

Proposition 4.3.4. *Let \mathcal{X} be a normed \mathbb{K} -vector space ($\mathbb{K} = \mathbb{R}$ or \mathbb{C}) for the norm $\|\cdot\|_{\mathcal{X}} : \mathcal{X} \rightarrow \mathbb{R} : x \mapsto \|x\|_{\mathcal{X}}$. Then $(\mathcal{X}, \mathcal{N})$, where \mathcal{N} is defined as*

$$\mathcal{N} : \mathcal{X} \rightarrow \wp(\wp(\mathcal{X})) : x \mapsto \{\mathcal{S} \subseteq \mathcal{X} | \exists \varepsilon \geq 0 : B(x, \varepsilon) = \{y \in \mathcal{X} : \|x - y\|_{\mathcal{X}} < \varepsilon\} \subseteq \mathcal{S}\}$$

is a topological space.

Proof. First, $\forall x \in \mathcal{X}, \mathcal{N}(x) \neq \emptyset$. Also, if $N \in \mathcal{N}(x)$, then $B(x, \varepsilon) \subseteq N$ for some $\varepsilon \geq 0$ and therefore $x \in N$.

Now consider, $\mathcal{S} \subseteq \mathcal{X}$ such that $\exists N \in \mathcal{N}(x)$ with $B(x, \varepsilon) \subseteq N$ and $N \subseteq \mathcal{S}$, then $\mathcal{S} \in \mathcal{N}(x)$ as $B(x, \varepsilon) \subseteq \mathcal{S}$.

Let $N, N' \in \mathcal{N}(x)$, we have that $\exists \varepsilon, \eta \geq 0$ such that $B(x, \varepsilon) \subseteq N$ and $B(x, \eta) \subseteq N'$, therefore $B(x, \min\{\varepsilon, \eta\}) \subseteq N \cap N'$ and thus $N \cap N' \in \mathcal{N}(x)$.

Finally, if $N \in \mathcal{N}(x)$, then $\exists \varepsilon \geq 0$ such that $B(x, \varepsilon) \subseteq N$. We obtain that $\forall b \in \underbrace{B(x, \varepsilon)}_{\in \mathcal{N}(x)}$, $N \in \mathcal{N}(b)$ because $B(b, \varepsilon - \|x - b\|_{\mathcal{X}}) \subseteq N$, where $\underbrace{B(b, \varepsilon - \|x - b\|_{\mathcal{X}})}_{\in \mathcal{N}(x)} \in \mathcal{N}(x)$. This is due to the fact that $B(b, \varepsilon - \|x - b\|_{\mathcal{X}}) = \{y \in \mathcal{X} : \|y - b\|_{\mathcal{X}} < \varepsilon - \|x - b\|_{\mathcal{X}}\} = \{y \in \mathcal{X} : \|x - b\|_{\mathcal{X}} + \|b - y\|_{\mathcal{X}} < \varepsilon\}$ and that $\|x - y\|_{\mathcal{X}} \leq \|y - b\|_{\mathcal{X}} + \|x - b\|_{\mathcal{X}}$, therefore, $B(b, \varepsilon - \|x - b\|_{\mathcal{X}}) \subseteq B(x, \varepsilon)$. \square

Definition 4.3.8 (Linear topological space). A topological space $(\mathcal{X}, \mathcal{N})$ is called a linear topological space (or topological vector space) if \mathcal{X} is a \mathbb{K} -vector space ($\mathbb{K} = \mathbb{R}$ or \mathbb{C}) such that the addition and scalar multiplication operators

$$\bullet + : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X} : (x, y) \mapsto x + y$$

$$\bullet \cdot : \mathbb{K} \times \mathcal{X} \rightarrow \mathcal{X} : (k, x) \mapsto k \cdot x$$

are continuous (with respect to the corresponding topologies).

Proposition 4.3.5. *Let \mathcal{X} be a normed \mathbb{K} -vector space ($\mathbb{K} = \mathbb{R}$ or \mathbb{C}) for the norm $\|\cdot\|_{\mathcal{X}} : \mathcal{X} \rightarrow \mathbb{R} : x \mapsto \|x\|_{\mathcal{X}}$, then $(\mathcal{X}, \mathcal{N})$, where \mathcal{N} is defined as*

$$\mathcal{N} : \mathcal{X} \rightarrow \wp(\wp(\mathcal{X})) : x \mapsto \{\mathcal{S} \subseteq \mathcal{X} \mid \exists \varepsilon \geq 0 : B(x, \varepsilon) = \{y \in \mathcal{X} : \|x - y\|_{\mathcal{X}} < \varepsilon\} \subseteq \mathcal{S}\},$$

is a linear topological space.

Proof. In order to prove the claim, we must show that the addition and scalar multiplication operators

$$\bullet + : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X} : (x, y) \mapsto x + y$$

$$\bullet \cdot : \mathbb{K} \times \mathcal{X} \rightarrow \mathcal{X} : (k, x) \mapsto k \cdot x$$

are continuous.

First, let us consider the addition operator. Let $(x, y) \in \mathcal{X} \times \mathcal{X}$ and let $M \in \mathcal{N}(x + y)$, then $\exists \varepsilon \geq 0$ such that $B(x + y, \varepsilon) \subseteq M$. We obtain that $(+)^{-1}(B(x + y, \varepsilon)) \subseteq (+)^{-1}(M)$, and

$$\begin{aligned} (+)^{-1}(B(x + y, \varepsilon)) &= \{(w, z) \in \mathcal{X} \times \mathcal{X} \mid w + z \in B(x + y, \varepsilon)\} \\ &= \{(w, z) \in \mathcal{X} \times \mathcal{X} \mid w + z \in \{t \in \mathcal{X} : \|x + y - t\|_{\mathcal{X}} < \varepsilon\}\} \\ &= \{(w, z) \in \mathcal{X} \times \mathcal{X} : \|x + y - (w + z)\|_{\mathcal{X}} < \varepsilon\} \\ &= \{(w, z) \in \mathcal{X} \times \mathcal{X} : \|(x - w) + (y - z)\|_{\mathcal{X}} < \varepsilon\} \\ &\supseteq \{(w, z) \in \mathcal{X} \times \mathcal{X} : \|(x - w)\|_{\mathcal{X}} + \|(y - z)\|_{\mathcal{X}} < \varepsilon\} \\ &\supseteq B\left(x, \frac{\varepsilon}{2}\right) \times B\left(y, \frac{\varepsilon}{2}\right) \in \mathcal{N}(x, y), \end{aligned}$$

therefore, $(+)^{-1}(B(x + y, \varepsilon)) \in \mathcal{N}(x, y)$ and $(+)^{-1}(M) \in \mathcal{N}(x, y)$. This shows that the addition is continuous.

Now, let us consider the scalar multiplication. Let $(k, x) \in \mathbb{K} \times \mathcal{X}$ and let $M \in \mathcal{N}(k \cdot x)$, then $\exists \varepsilon \geq 0$ such that $B(k \cdot x, \varepsilon) \subseteq M$. We obtain that $(\cdot)^{-1}(B(k \cdot x, \varepsilon)) \subseteq (\cdot)^{-1}(M)$, and

$$\begin{aligned} (\cdot)^{-1}(B(k \cdot x, \varepsilon)) &= \{(l, y) \in \mathbb{K} \times \mathcal{X} \mid l \cdot y \in B(k \cdot x, \varepsilon)\} \\ &= \{(l, y) \in \mathbb{K} \times \mathcal{X} \mid l \cdot y \in \{t \in \mathcal{X} : \|k \cdot x - t\|_{\mathcal{X}} < \varepsilon\}\} \\ &= \{(l, y) \in \mathbb{K} \times \mathcal{X} : \|k \cdot x - l \cdot y\|_{\mathcal{X}} < \varepsilon\} \\ &= \{(l, y) \in \mathbb{K} \times \mathcal{X} : \|k \cdot x - l \cdot x + l \cdot x - l \cdot y\|_{\mathcal{X}} < \varepsilon\} \\ &= \{(l, y) \in \mathbb{K} \times \mathcal{X} : \|(k - l) \cdot x + l \cdot (x - y)\|_{\mathcal{X}} < \varepsilon\} \end{aligned}$$

$$\begin{aligned}
 &\supseteq \{(l, y) \in \mathbb{K} \times \mathcal{X} : \|(k - l) \cdot x\|_{\mathcal{X}} + \|l \cdot (x - y)\|_{\mathcal{X}} < \varepsilon\} \\
 &= \{(l, y) \in \mathbb{K} \times \mathcal{X} : |k - l| \|x\|_{\mathcal{X}} + |l| \|x - y\|_{\mathcal{X}} < \varepsilon\} \\
 &\supseteq B(k, \delta_1) \times B(x, \delta_2) \in \mathcal{N}(k, x),
 \end{aligned}$$

where

$$\delta_1 = \begin{cases} \frac{\varepsilon}{2\|x\|_{\mathcal{X}}} & \text{if } \|x\|_{\mathcal{X}} \neq 0 \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad \delta_2 = \begin{cases} \frac{\varepsilon}{2(|k| + \frac{\varepsilon}{2\|x\|_{\mathcal{X}}})} & \text{if } \|x\|_{\mathcal{X}} \neq 0 \\ \frac{\varepsilon}{1+|k|} & \text{otherwise} \end{cases}.$$

Indeed, let us take any $(l, k) \in B(k, \delta_1) \times B(x, \delta_2)$. If $\|x\|_{\mathcal{X}} \neq 0$, then

$$\begin{aligned}
 |k - l| \|x\|_{\mathcal{X}} + |l| \|x - y\|_{\mathcal{X}} &< \delta_1 \|x\|_{\mathcal{X}} + |l - k + k| \|x - y\|_{\mathcal{X}} \\
 &< \frac{\varepsilon}{2} + (|l - k| + |k|) \frac{\varepsilon}{2(|k| + \frac{\varepsilon}{2\|x\|_{\mathcal{X}}})} \\
 &< \frac{\varepsilon}{2} + \left(|k| + \frac{\varepsilon}{2\|x\|_{\mathcal{X}}}\right) \frac{\varepsilon}{2(|k| + \frac{\varepsilon}{2\|x\|_{\mathcal{X}}})} \\
 &= \varepsilon,
 \end{aligned}$$

and, if $\|x\|_{\mathcal{X}} = 0$, then

$$\begin{aligned}
 |k - l| \|x\|_{\mathcal{X}} + |l| \|x - y\|_{\mathcal{X}} &= |l - k + k| \|x - y\|_{\mathcal{X}} \\
 &< (|l - k| + |k|) \frac{\varepsilon}{1 + |k|} \\
 &< (1 + |k|) \frac{\varepsilon}{1 + |k|} \\
 &= \varepsilon,
 \end{aligned}$$

therefore, $(l, y) \in (\cdot)^{-1}(B(k \cdot x, \varepsilon))$. This shows that $(\cdot)^{-1}(B(k \cdot x, \varepsilon)) \in \mathcal{N}(k, x)$, which implies that $(\cdot)^{-1}(M) \in \mathcal{N}(k, x)$ and thus that the scalar multiplication operator is continuous. \square

Let us now come back to the idea behind derivation and consider $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{M})$ to be two linear topological spaces over the field \mathbb{R} . What is of interest is to approximate a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ at $x \in \mathcal{X}$ with a mapping $A : \mathcal{X} \rightarrow \mathcal{Y}$, such that

$$f(x + h) = f(x) + A(h) + r(h)$$

where $h \in \mathcal{X}$ and $r(h)$ represents a quantity negligible compared to h . We will not define $r(h)$ formally, as different formal definitions will result into different notions of the derivative.

Definition 4.3.9 (General derivatives). Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{M})$ be two linear topological spaces over the field \mathbb{R} . Moreover, consider $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y})$ to be two classes of mappings from \mathcal{X} to \mathcal{Y} that represent a class of approximation mappings and a class of infinitesimal mappings, respectively. We say that the mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ is differentiable at $x \in \mathcal{X}$ with respect to $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y})$, if there exists two mappings $A_x \in \mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $r_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y})$ such that $\forall h \in \mathcal{X}$,

$$f(x + h) = f(x) + A_x(h) + r_x(h).$$

In this case, the mapping A_x is called the derivative of f at x . Furthermore, the mapping

$$Df : \mathcal{X} \rightarrow \mathcal{A}(\mathcal{X}, \mathcal{Y}) : x \mapsto A_x,$$

defined for all $x \in \mathcal{X}$ such that A_x exists, is called the derivative of f .

Remark 4.3.1. For what follows, we will consider \mathcal{X} and \mathcal{Y} to normed \mathbb{R} -vector spaces.

We generally expect the following properties from $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y})$:

1. $\forall r \in \mathcal{R}(\mathcal{X}, \mathcal{Y}), r(0_{\mathcal{X}}) = 0_{\mathcal{Y}}$.
2. $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y})$ are \mathbb{R} -vector subspaces of the set $\mathcal{M}(\mathcal{X}, \mathcal{Y}) = \{m : \mathcal{X} \rightarrow \mathcal{Y}\}$ of all mappings from \mathcal{X} to \mathcal{Y} .
3. $\mathcal{A}(\mathcal{X}, \mathcal{Y}) \cap \mathcal{R}(\mathcal{X}, \mathcal{Y}) = \{0\}$ where 0 is defined as

$$0 : \mathcal{X} \rightarrow \mathcal{Y} : x \mapsto 0_{\mathcal{Y}}.$$

4. If $\mathcal{X} = \mathbb{R}$, then $\mathcal{R}(\mathcal{X}, \mathcal{Y}) = \left\{ r : \mathbb{R} \rightarrow \mathcal{Y} \mid \lim_{t \rightarrow 0} \frac{r(t)}{t} = 0_{\mathcal{Y}} \right\}$.

Proposition 4.3.6. Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{M})$ be two normed \mathbb{R} -vector spaces. If $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y})$ are two classes of mappings from \mathcal{X} to \mathcal{Y} such that properties 2 and 3 hold, then the derivative A_x , if it exists, of a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ at $x \in \mathcal{X}$ is unique.

Proof. Suppose that there exist mappings $A_x, B_x \in \mathcal{A}(\mathcal{X}, \mathcal{Y}), r_x, q_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y})$ such that $\forall h \in \mathcal{X}$,

$$f(x + h) = f(x) + A_x(h) + r_x(h) = f(x) + B_x(h) + q_x(h).$$

We then have

$$\begin{aligned} A_x(h) + r_x(h) &= B_x(h) + q_x(h) \\ \Leftrightarrow A_x(h) - B_x(h) &= q_x(h) - r_x(h) \\ \Leftrightarrow \underbrace{(A_x - B_x)(h)}_{\in \mathcal{A}(\mathcal{X}, \mathcal{Y})} &= \underbrace{(q_x - r_x)(h)}_{\in \mathcal{R}(\mathcal{X}, \mathcal{Y})}. \end{aligned}$$

Therefore, $(A_x - B_x) \in \mathcal{A}(\mathcal{X}, \mathcal{Y}) \cap \mathcal{R}(\mathcal{X}, \mathcal{Y})$ and thus $(A_x - B_x) = 0$, which implies that $A_x = B_x$. \square

Remark 4.3.2. A common choice is to take $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ to be $\mathcal{L}(\mathcal{X}, \mathcal{Y})$, the set of all continuous linear mappings from \mathcal{X} to \mathcal{Y} .

To conclude this section, we will mention three types of derivatives which are based on a similar construction of $\mathcal{R}(\mathcal{X}, \mathcal{Y})$.

Consider $\mathcal{R}(\mathcal{X}, \mathcal{Y})$ to be

$$\mathcal{R}(\mathcal{X}, \mathcal{Y}; \mathcal{C}_{\mathcal{X}}) = \left\{ r : \mathcal{X} \rightarrow \mathcal{Y} \mid \forall C \in \mathcal{C}_{\mathcal{X}} : \limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|r(th)\|_{\mathcal{Y}}}{t} = 0 \right\},$$

where $\mathcal{C}_{\mathcal{X}}$ is a covering of \mathcal{X} , that is, a set of bounded subsets of \mathcal{X} . The three types of derivative can be defined using specific coverings $\mathcal{C}_{\mathcal{X}}$.

- The Fréchet derivative is obtained by taking $\mathcal{C}_{\mathcal{X}} = \{C \subseteq \mathcal{X} \mid C \text{ is bounded}\}$.
- The Hadamard derivative is obtained by taking $\mathcal{C}_{\mathcal{X}} = \{C \subseteq \mathcal{X} \mid C \text{ is compact}\}$.
- The Gâteaux derivative is obtained by taking $\mathcal{C}_{\mathcal{X}} = \{C \subseteq \mathcal{X} \mid C \text{ is finite}\}$.

Note that

$$\{C \subseteq \mathcal{X} \mid C \text{ is finite}\} \subseteq \{C \subseteq \mathcal{X} \mid C \text{ is compact}\} \subseteq \{C \subseteq \mathcal{X} \mid C \text{ is bounded}\},$$

which means that the Fréchet differentiability implies the Hadamard differentiability, which itself implies the Gâteaux differentiability.

An equivalent definition of the Gâteaux derivative is the following.

Definition 4.3.10 (Gâteaux derivative). Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{M})$ to be two normed \mathbb{R} -vector spaces and let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a mapping, then f is Gâteaux differentiable at $x \in \mathcal{X}$ if $\forall h \in \mathcal{X}$, the limit

$$\lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists. In such case, the Gâteaux derivative of f at $x \in \mathcal{X}$ in the direction of $h \in \mathcal{X}$ is the value of this limit, written $f'_x(h)$.

Basically, a mapping f is Gâteaux differentiable at some point if it is differentiable (in the classical sense) along any line that passes through this point.

Proposition 4.3.7. Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{N})$ to be two normed \mathbb{R} -vector spaces and let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a mapping that is Gâteaux differentiable at $x \in \mathcal{X}$, then its derivative is homogeneous, that is, $\forall \alpha \in \mathbb{R}$

$$f'_x(\alpha h) = \alpha f'_x(h).$$

Proof. Let $\alpha \in \mathbb{R}$ and $x \in \mathcal{X}$ such that f is Gâteaux differentiable at x .

First, consider that $\alpha \neq 0$, we have

$$\begin{aligned}
 f'_x(\alpha h) &= \lim_{t \rightarrow 0} \frac{f(x + t\alpha h) - f(x)}{t} \\
 &= \lim_{t \rightarrow 0} \frac{\alpha}{\alpha} \frac{f(x + t\alpha h) - f(x)}{t} \\
 &= \alpha \lim_{t \rightarrow 0} \frac{f(x + t\alpha h) - f(x)}{t\alpha} \\
 &= \alpha \lim_{u \rightarrow 0} \frac{f(x + uh) - f(x)}{u} \\
 &= \alpha f'_x(h),
 \end{aligned} \tag{£}$$

where (£) is due to the continuity of scalar multiplication in \mathcal{X} . If $\alpha = 0$, then

$$\begin{aligned}
 f'_x(\alpha h) &= f'_x(0) \\
 &= \lim_{t \rightarrow 0} \frac{f(x + t0) - f(x)}{t} \\
 &= \lim_{t \rightarrow 0} \frac{f(x) - f(x)}{t} \\
 &= 0 \\
 &= \alpha f'_x(h).
 \end{aligned}$$

□

To show that the two definitions are equivalent, we will consider $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ to be $\mathcal{H}(\mathcal{X}, \mathcal{Y})$, the set of homogeneous mappings from \mathcal{X} to \mathcal{Y} .

Proposition 4.3.8. *Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{N})$ be two normed \mathbb{R} -vector spaces, $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a mapping and $x \in \mathcal{X}$. Then both definitions of the Gâteaux derivative are equivalent, that is, $\exists A_x \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$ and $r_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$ such that $\forall h \in \mathcal{X}$,*

$$f(x + h) = f(x) + A_x(h) + r_x(h),$$

if and only if, $\forall h \in \mathcal{X}$, the limit

$$\lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists.

Proof. Consider that $\exists A_x \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$ and $r_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$ such that $\forall h \in \mathcal{X}$,

$$f(x + h) = f(x) + A_x(h) + r_x(h).$$

Then, $\forall h \in \mathcal{X}$ we have

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t} &= \lim_{t \rightarrow 0} \frac{f(x) + A_x(th) + r_x(th) - f(x)}{t} \\ &= \lim_{t \rightarrow 0} \frac{tA_x(h)}{t} + \lim_{t \rightarrow 0} \frac{r_x(th)}{t} \\ &= A_x(h), \end{aligned}$$

as r_x is such that $\limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|r_x(th)\|_{\mathcal{Y}}}{t} = 0$ for all finite subset C of \mathcal{X} , in particular for $C = \{h\}$. Therefore the limit exists $\forall h \in \mathcal{X}$.

Reciprocally, consider that $\forall h \in \mathcal{X}$, the limit

$$\lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists, and is $f'_x(h)$. Then, $\forall h \in \mathcal{X}$, we have

$$f(x + h) = f(x) + f'_x(h) + r_x(h),$$

where $r_x(h) = f(x + h) - f(x) - f'_x(h)$. As showed in Proposition 4.3.7, $f'_x(h) \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$. We just need to show that $r_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$, that is, $\forall C \subseteq \mathcal{X}$ with C finite,

$$\limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|r_x(th)\|_{\mathcal{Y}}}{t} = 0.$$

Let $C \subseteq \mathcal{X}$ with C finite, we have

$$\begin{aligned} &\limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|r_x(th)\|_{\mathcal{Y}}}{t} \\ &= \limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|f(x + th) - f(x) - f'_x(th)\|_{\mathcal{Y}}}{t} \\ &= \limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|f(x + th) - f(x) - tf'_x(h)\|_{\mathcal{Y}}}{t} \\ &= \limsup_{t \rightarrow 0} \sup_{h \in C} \left\| \frac{f(x + th) - f(x) - tf'_x(h)}{t} \right\|_{\mathcal{Y}} \\ &= \limsup_{t \rightarrow 0} \sup_{h \in C} \left\| \frac{f(x + th) - f(x)}{t} - f'_x(h) \right\|_{\mathcal{Y}}. \end{aligned}$$

Now since, C is finite and that we have $\forall h \in \mathcal{X}$

$$\lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t} = f'_x(h),$$

we can conclude that $\limsup_{t \rightarrow 0} \sup_{h \in C} \frac{\|r_x(th)\|_{\mathcal{Y}}}{t} = 0$. □

Proposition 4.3.9. Let $(\mathcal{X}, \mathcal{N}), (\mathcal{Y}, \mathcal{N})$ be two normed \mathbb{R} -vector spaces, $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a mapping and $x \in \mathcal{X}$. If f is Gâteaux differentiable at $x \in \mathcal{X}$, that is, $\exists f'_x \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$ and $r_x \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$ such that $\forall h \in \mathcal{X}$,

$$f(x + h) = f(x) + f'_x(h) + r_x(h),$$

then f'_x is unique.

Proof. Using the result of Proposition 4.3.6, to prove the statement we just need to show that $\mathcal{H}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$ are such that they are both \mathbb{R} -vector subspaces of $\mathcal{M}(\mathcal{X}, \mathcal{Y}) = \{m : \mathcal{X} \rightarrow \mathcal{Y}\}$ and that

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) \cap \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\}) = \{0 : \mathcal{X} \rightarrow \mathcal{Y} : x \mapsto 0_{\mathcal{Y}}\}.$$

First, we have that $\forall \alpha, \lambda, \xi \in \mathbb{R}, \forall f, g \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$ and $\forall x \in \mathcal{X}$,

$$\lambda f(\alpha x) + \xi g(\alpha x) = \lambda \alpha f(x) + \xi \alpha g(x) = \alpha(\lambda f(x) + \xi g(x)),$$

therefore, $\lambda f + \xi g \in \mathcal{H}(\mathcal{X}, \mathcal{Y})$. Also, $\forall \lambda, \xi \in \mathbb{R}, \forall r, q \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$ and $\forall C \subseteq \mathcal{X}$ with C finite,

$$\begin{aligned} & \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|\lambda r(th) + \xi q(th)\|_{\mathcal{Y}}}{t} \\ & \leq \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|\lambda r(th)\|_{\mathcal{Y}} + \|\xi q(th)\|_{\mathcal{Y}}}{t} \\ & \leq \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|\lambda r(th)\|_{\mathcal{Y}}}{t} + \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|\xi q(th)\|_{\mathcal{Y}}}{t} \\ & = |\lambda| \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|r(th)\|_{\mathcal{Y}}}{t} + |\xi| \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|q(th)\|_{\mathcal{Y}}}{t} \\ & = 0. \end{aligned}$$

As such, $\limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|\lambda r(th) + \xi q(th)\|_{\mathcal{Y}}}{t} = 0$ and $\lambda r + \xi q \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$.

Now, consider $f \in \mathcal{H}(\mathcal{X}, \mathcal{Y}) \cap \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\})$, then $\forall C \subseteq \mathcal{X}$ with C finite, we have

$$\begin{aligned} 0 &= \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|f(th)\|_{\mathcal{Y}}}{t} \text{ as } f \in \mathcal{R}(\mathcal{X}, \mathcal{Y}; \{C \subseteq \mathcal{X} | C \text{ is finite}\}) \\ &= \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{\|tf(h)\|_{\mathcal{Y}}}{t} \text{ as } f \in \mathcal{H}(\mathcal{X}, \mathcal{Y}) \\ &= \limsup_{t \rightarrow 0} \limsup_{h \in C} \frac{|t| \|f(h)\|_{\mathcal{Y}}}{t} \text{ as } t > 0 \end{aligned}$$

$$= \text{sign}(t) \sup_{h \in C} \|f(h)\|_{\mathcal{Y}},$$

as a result, $\sup_{h \in C} \|f(h)\|_{\mathcal{Y}} = 0$, and thus since C is finite, $\forall h \in \mathcal{X}$, we have that $\|f(h)\|_{\mathcal{Y}} = 0$.

This means that $f(h) = 0 \forall h \in \mathcal{X}$ and therefore that

$$f : \mathcal{X} \rightarrow \mathcal{Y} : x \mapsto 0_{\mathcal{Y}}.$$

□

Remark 4.3.3. In what follows, we will consider $(\mathcal{Y}, \mathcal{M})$ to be the \mathbb{R} -vector space \mathbb{R}^p equipped with the Euclidean norm $\|\cdot\|$, as this particular context of the derivative is of interest for the subject of robustness.

Remark 4.3.4. The Gâteaux derivative is not always a linear continuous mapping in h . As a result, another definition of a derivative that addresses this issue is the Gâteaux-Levy derivative, which is the Gâteaux derivative if it is linear and continuous.

4.3.3 Theoretical influence function

The previous section serves as a brief summary of the theory behind the Gâteaux derivative, whose concept is used to define the theoretical influence function of a statistical functional.

Definition 4.3.11 (Gâteaux derivative for statistical functionals). Let $T : \mathcal{F} \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a statistical functional and $F, G \in \mathcal{F}$. Then T is differentiable in Gâteaux's sense at F in the direction of G if the limit

$$\lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon G) - T(F)}{\varepsilon}$$

exists. In this case, the limit is written $T'_F(G)$.

Note that as \mathcal{F} is not a linear topological space (because it is not a \mathbb{R} -vector space), the definition of this Gâteaux derivative does not follow directly from Definition 4.3.10. Rather, it is an adaptation of the idea in order to be applied in the context of statistical functionals.

In fact, we could properly define the Gâteaux derivative of a mapping T taking values in \mathbb{R}^p and defined on $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, the set of all finite signed measures¹ on the underlying measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. An element of $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ is a function of the form

$$\mu : \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R} : B \mapsto \mu(B),$$

where

- $\mu(\emptyset) = 0$.

¹All the developments on finite signed measures are based on [17].

- For each sequence of disjoint sets, $(B_n)_{n \in \mathbb{N}_0}$,

$$\mu \left(\bigcup_{i=1}^{+\infty} B_i \right) = \sum_{i=1}^{+\infty} \mu(B_i).$$

Proposition 4.3.10. *The set $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, equipped with the two operations*

- $+$: $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})^2 \rightarrow M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$: $(\mu, \nu) \mapsto \mu + \nu$: $\mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$: $B \mapsto \mu(B) + \nu(B)$,
- \cdot : $\mathbb{R} \times M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$: $(\lambda, \mu) \mapsto \lambda \cdot \mu$: $\mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$: $B \mapsto \lambda \mu(B)$,

is an \mathbb{R} -vector space.

Proof. In order to be an \mathbb{R} -vector space, $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ must have the following properties:

- $\forall \mu, \nu, \iota \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, $\mu + (\nu + \iota) = (\mu + \nu) + \iota$ and $\mu + \nu = \nu + \mu$. This is a direct consequence of the definition of the addition.
- There must exist a unique element $0 \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, such that $0 + \mu = \mu + 0 = \mu$. Taking $0 : \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R} : B \mapsto 0$, we have that $0 \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ and that it respects the conditions.
- $\forall \mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, there must exist a unique $\nu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ such that $\mu + \nu = \nu + \mu = 0$. We just need to take $\nu : \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R} : B \mapsto -\mu(B)$.
- $\forall \lambda, \xi \in \mathbb{R}$, and $\forall \mu, \nu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, we must have $\lambda(\mu + \nu) = \lambda\mu + \lambda\nu$, $(\lambda + \xi)\mu = \lambda\mu + \xi\mu$, $(\lambda\xi)\mu = \lambda(\xi\mu)$ and $1\mu = \mu$. Once again, it follows directly from the definition of the scalar multiplication.

□

We can obtain the following theorem proved in [17].

Proposition 4.3.11 (Jordan's decomposition). *Let $\mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$. Then there exist two finite positive measures μ^+ and μ^- such that $\forall B \in \mathcal{B}(\mathcal{X})$,*

$$\mu(B) = \mu^+(B) - \mu^-(B).$$

Furthermore, we have $\mu^+(B) = \sup\{\mu(A) | A \in \mathcal{B}(\mathcal{X}), A \subseteq B\}$ and $\mu^-(B) = \sup\{-\mu(A) | A \in \mathcal{B}(\mathcal{X}), A \subseteq B\}$.

From this result, we can define a norm on $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, that is the norm of total variation.

Definition 4.3.12 (Norm of total variation). Let $\mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$. The norm of total variation of μ is defined as

$$\|\mu\|_{\text{tv}} = |\mu|(X),$$

where $|\mu| = \mu^+ + \mu^-$ if Jordan's decomposition of μ is $\mu = \mu^+ - \mu^-$.

Proposition 4.3.12. *The norm of total variation is a norm.*

Proof. In order to prove the claim, we must show the following:

- $\forall \mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}), \|\mu\|_{\text{tv}} \geq 0$. This is trivial in this case.
- $\|\mu\|_{\text{tv}} = 0$ if and only if μ is the null measure. We have

$$\begin{aligned} \|\mu\|_{\text{tv}} = 0 &\Leftrightarrow |\mu|(X) = 0 \\ &\Leftrightarrow \mu^+(X) + \mu^-(X) = 0 \\ &\Leftrightarrow \mu^+(X) = 0 \text{ and } \mu^-(X) = 0 \\ &\Leftrightarrow \mu^+ \text{ and } \mu^- \text{ are the null measure} \\ &\Leftrightarrow \mu \text{ is the null measure.} \end{aligned}$$

- $\forall \lambda \in \mathbb{R}, \|\lambda\mu\|_{\text{tv}} = |\lambda| \|\mu\|_{\text{tv}}$. We have

$$\begin{aligned} \|\lambda\mu\|_{\text{tv}} &= |\lambda\mu|(X) \\ &= (\lambda\mu)^+(X) + (\lambda\mu)^-(X) \\ &= (\text{sign}(\lambda)|\lambda|\mu)^+(X) + (\text{sign}(\lambda)|\lambda|\mu)^-(X) \\ &= |\lambda|((\text{sign}(\lambda)\mu)^+(X) + (\text{sign}(\lambda)\mu)^-(X)) \\ &= |\lambda|(\mu^+(X) + \mu^-(X)) \\ &= |\lambda||\mu|(X) \\ &= |\lambda| \|\mu\|_{\text{tv}}. \end{aligned}$$

- $\forall \mu, \nu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}), \|\mu + \nu\|_{\text{tv}} \leq \|\mu\|_{\text{tv}} + \|\nu\|_{\text{tv}}$. We have

$$\begin{aligned} \|\mu + \nu\|_{\text{tv}} &= |\mu + \nu|(X) \\ &= (\mu + \nu)^+(X) + (\mu + \nu)^-(X) \\ &\leq \mu^+(X) + \mu^-(X) + \nu^+(X) + \nu^-(X) \\ &= |\mu|(X) + |\nu|(X) \\ &= \|\mu\|_{\text{tv}} + \|\nu\|_{\text{tv}}. \end{aligned}$$

□

From what precedes, we directly obtain the following.

Theorem 4.3.1. *The set $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ equipped with the norm of total variation is a normed \mathbb{R} -vector space and therefore also a linear topological space.*

In this context, we can redefine the Gâteaux derivative properly.

Definition 4.3.13 (Gâteaux derivative for functionals of finite signed measure). Let $T : M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a functional of finite signed measure, then T is Gâteaux differentiable at $\mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ if $\forall \nu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, the limit

$$\lim_{t \rightarrow 0} \frac{T(\mu + t\nu) - T(\mu)}{t}$$

exists. In this case, the Gâteaux derivative of T at $\mu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ in the direction of $\nu \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ is the value of this limit, written $T'_\mu(\nu)$.

As a result, the Gâteaux derivative for statistical functionals can be seen as the restriction of the Gâteaux derivative for finite signed measures on the set of probability measures (which is a convex subset of all the finite signed measures, see Proposition 4.1.1). Note that there is a small variation between the two definitions concerning the direction of the derivative. If we take a statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^p$, the Gâteaux derivative, as defined at the beginning of this section (see Definition 4.3.11), of T at F in the direction of G is

$$\lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon G) - T(F)}{\varepsilon} = \lim_{\varepsilon \rightarrow 0} \frac{T(F + \varepsilon(G - F)) - T(F)}{\varepsilon},$$

which correspond to the Gâteaux derivative, as defined just above (see Definition 4.3.13), of T at F in the direction of $G - F$. This variation is due to the fact that in the first definition we only work with probability measures and that $G - F$ is not a probability measure.

Remark 4.3.5 (Notation). In what follows, we will use the notation of the Gâteaux derivative for statistical functionals.

Remark 4.3.6. From now on, we will implicitly suppose to work with random vectors that take values in \mathbb{R}^q , $q \in \mathbb{N}_0$. Therefore, \mathcal{F} is a subset of all probability distribution on $(\mathcal{X}, \mathcal{B}(\mathcal{X})) = (\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$.

We can finally define the theoretical influence function of a statistical functional.

Definition 4.3.14 (Theoretical influence function). For $T : \mathcal{F} \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) a statistical functional, the influence function (IF) of T at $x \in \mathbb{R}^q$, $q \in \mathbb{N}_0$, for a distribution $F \in \mathcal{F}$ is defined by

$$\text{IF}(x; T, F) = \lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon \Delta_x) - T(F)}{\varepsilon} = T'_F(\Delta_x)$$

if this limit exists, where Δ_x is the Dirac distribution at $x \in \mathbb{R}^q$.

Basically, the theoretical influence function tells us “how much” a singularity at x will perturb the functional T .

Remark 4.3.7. In practice, an easier way of computing the influence function is to use the fact that

$$\text{IF}(x; T, F) = \left. \frac{dT(F_\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0}$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_x$.

Remark 4.3.8. Note that the theoretical influence function defined as before requires that the functional is differentiable in Gâteaux’s sense for every distribution in the direction of the Dirac distribution, which is not guaranteed and imposes some restrictions on the distribution.

We show examples for the mean, median, and variance.

Example 4.3.4. Consider the statistical functional of the mean

$$T_{\text{mean}} : \mathcal{F} \rightarrow \mathbb{R}^p : F \mapsto \mathbb{E}_F[X],$$

and $x \in \mathbb{R}^p$. Then we have

$$T_{\text{mean}}(F_\varepsilon) = \mathbb{E}_{(1-\varepsilon)F + \varepsilon\Delta_x}[X] = (1 - \varepsilon)\mathbb{E}_F[X] + \varepsilon\mathbb{E}_{\Delta_x}[X] = (1 - \varepsilon)\mathbb{E}_F[X] + \varepsilon x.$$

Therefore, we obtain that

$$\text{IF}(x; T_{\text{mean}}, F) = \left. \frac{dT_{\text{mean}}(F_\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} = x - \mathbb{E}_F[X] = x - T_{\text{mean}}(F).$$

If we take the example of a standard normal distribution we obtain the influence function as displayed in Figure 4.1.

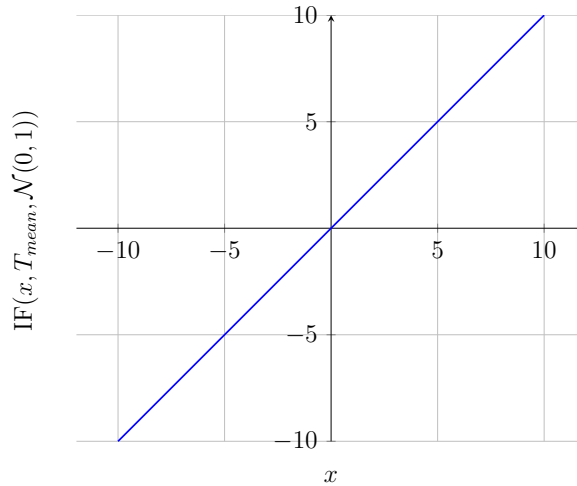


Figure 4.1: Influence function of the statistical functional of the mean for the standard normal distribution.

We can directly see that this function is not bounded, which signifies that the mean can be influenced and take arbitrarily large values in the presence of a single outlier that tends towards infinity.

Example 4.3.5. Consider the statistical functional of the median

$$T_{med} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \inf \left\{ t \in \mathbb{R} \left| \int_{\mathbb{R}} \mathbb{1}_{\{w \leq t\}} dF(w) \geq \frac{1}{2} \right. \right\},$$

and $x \in \mathbb{R}$. We will only consider the case of a continuous distribution F with cumulative distribution function given by the function C and probability density function given by the function f , so that $C(T_{med}(F)) = \frac{1}{2}$.

In this context, if we write D_x the cumulative distribution function of the Dirac distribution at x , $\mu_\varepsilon = T_{med}(F_\varepsilon)$, and $\mu = T_{med}(F)$, we have

$$\begin{aligned} C_\varepsilon(\mu_\varepsilon) &= \frac{1}{2} \\ \Leftrightarrow (1 - \varepsilon)C(\mu_\varepsilon) + \varepsilon D_x(\mu_\varepsilon) &= \frac{1}{2} \\ \Rightarrow \frac{d[(1 - \varepsilon)C(\mu_\varepsilon) + \varepsilon D_x(\mu_\varepsilon)]}{d\varepsilon} \Big|_{\varepsilon=0} &= 0 \\ \Leftrightarrow \left[-C(\mu_\varepsilon) + (1 - \varepsilon)f(\mu_\varepsilon) \frac{d\mu_\varepsilon}{d\varepsilon} + D_x(\mu_\varepsilon) + \varepsilon \frac{dD_x(\mu_\varepsilon)}{d\varepsilon} \right] \Big|_{\varepsilon=0} &= 0 \\ \Leftrightarrow -\frac{1}{2} + f(\mu) \frac{d\mu_\varepsilon}{d\varepsilon} \Big|_{\varepsilon=0} + D_x(\mu) &= 0 \\ \Leftrightarrow \frac{d\mu_\varepsilon}{d\varepsilon} \Big|_{\varepsilon=0} &= \frac{1}{f(\mu)} \left(\frac{1}{2} - D_x(\mu) \right) \\ \Leftrightarrow \frac{d\mu_\varepsilon}{d\varepsilon} \Big|_{\varepsilon=0} &= \begin{cases} \frac{1}{2f(\mu)} & \text{if } x < \mu \\ -\frac{1}{2f(\mu)} & \text{if } x \geq \mu \end{cases}. \end{aligned}$$

Therefore, we obtain that

$$\begin{aligned} \text{IF}(x; T_{med}, F) &= \frac{dT_{med}(F_\varepsilon)}{d\varepsilon} \Big|_{\varepsilon=0} \\ &= \begin{cases} \frac{1}{2f(T_{med}(F))} & \text{if } x < T_{med}(F) \\ -\frac{1}{2f(T_{med}(F))} & \text{if } x \geq T_{med}(F) \end{cases}. \end{aligned}$$

Once again, let us take the example of a standard normal distribution. We then obtain the influence function as given in Figure 4.2.

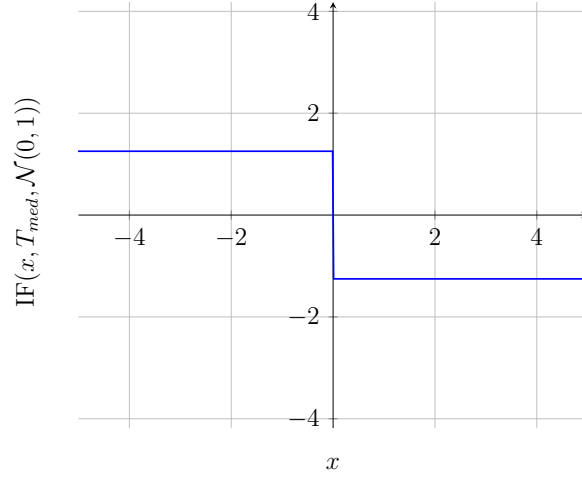


Figure 4.2: Influence function of the statistical functional of the median for the standard normal distribution.

In this case, the influence function is bounded because we have

$$\sup_{x \in \mathbb{R}} |\text{IF}(x; T_{\text{med}}, \mathcal{N}(0, 1))| = \sqrt{\frac{\pi}{2}}.$$

Example 4.3.6. Consider the statistical functional for the variance

$$T_{\text{var}} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \mathbb{E}_F[(X - \mathbb{E}_F[X])^2] = \mathbb{E}_F[X^2] - (\mathbb{E}_F[X])^2,$$

and $x \in \mathbb{R}$. We have

$$\begin{aligned} T_{\text{var}}(F_\varepsilon) &= \mathbb{E}_{F_\varepsilon}[X^2] - (\mathbb{E}_{F_\varepsilon}[X])^2 \\ &= \int_{\mathbb{R}} s^2 d((1 - \varepsilon)F + \varepsilon\Delta_x)(s) - \left(\int_{\mathbb{R}} s d((1 - \varepsilon)F + \varepsilon\Delta_x)(s) \right)^2 \\ &= (1 - \varepsilon)\mathbb{E}_F[X^2] + \varepsilon x^2 - ((1 - \varepsilon)\mathbb{E}_F[X] + \varepsilon x)^2. \end{aligned}$$

We can then obtain that

$$\begin{aligned} \text{IF}(x; T_{\text{var}}, F) &= \left. \frac{dT_{\text{var}}(F_\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} \\ &= -\mathbb{E}_F[X^2] + x^2 - 2\mathbb{E}_F[X](-\mathbb{E}_F[X] + x) \\ &= 2(\mathbb{E}_F[X])^2 - \mathbb{E}_F[X^2] - 2x\mathbb{E}_F[X] + x^2 \\ &= (\mathbb{E}_F[X] - x)^2 - T_{\text{var}}(F). \end{aligned}$$

Taking a standard normal distribution, we obtain the influence function shown in Figure 4.3.

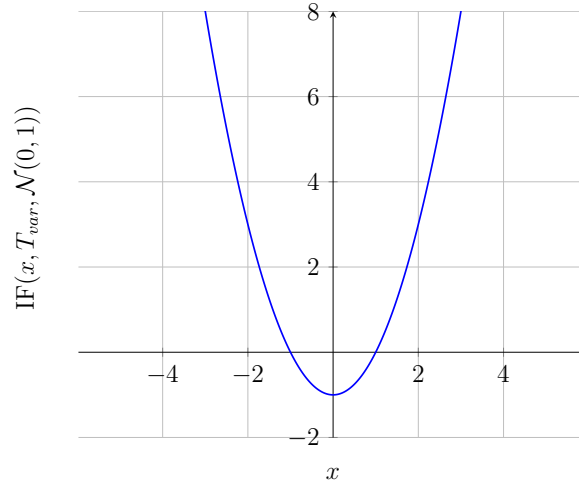


Figure 4.3: Influence function of the statistical functional of the variance for the standard normal distribution.

It is clear that this function is not bounded.

4.3.4 Link between the empirical and theoretical influence function

In this subsection, we will discuss the link that exists between the empirical and theoretical influence function.

Let $T_n : \mathbb{R}^{q \times n} \rightarrow \mathbb{R}^p$ be some estimator and consider $\mathbf{x} = (x_1, \dots, x_n)$, with $x_i \in \mathbb{R}^q$, to be a sample of observations of size n i.i.d. from a distribution F . Furthermore, suppose that there exists a statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^p$ such that for all possible samples of observations $\mathbf{x} = (x_1, \dots, x_n)$, we have $\forall n \in \mathbb{N}_0$

$$T_n(\mathbf{x}) = T(F_n),$$

where F_n is the empirical distribution of \mathbf{x} . That is,

$$F_n = \frac{1}{n} \sum_{i=1}^n \Delta_{x_i}.$$

In this case, we have for $x \in \mathbb{R}$

$$\begin{aligned} \text{EIF}(x; T_n, \mathbf{x}) &= \frac{T_{n+1}(\mathbf{x} \cup \{x\}) - T_n(\mathbf{x})}{\frac{1}{n+1}} \\ &= \frac{T\left(\frac{1}{n+1} \sum_{i=1}^n \Delta_{x_i} + \frac{1}{n+1} \Delta_x\right) - T(F_n)}{\frac{1}{n+1}} \end{aligned}$$

$$\begin{aligned}
&= \frac{T\left(\frac{n}{n+1}F_n + \frac{1}{n+1}\Delta_x\right) - T(F_n)}{\frac{1}{n+1}} \\
&= \frac{T\left(\left(1 - \frac{1}{n+1}\right)F_n + \frac{1}{n+1}\Delta_x\right) - T(F_n)}{\frac{1}{n+1}} \\
&\approx \frac{T\left(\left(1 - \frac{1}{n+1}\right)F + \frac{1}{n+1}\Delta_x\right) - T(F)}{\frac{1}{n+1}} \quad (\$) \\
&\approx \lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon\Delta_x) - T(F)}{\varepsilon} \text{ if } n \text{ is large} \\
&= \text{IF}(x; T, F).
\end{aligned}$$

We could hope that for an estimator T_n , a sample $\mathbf{X} = (X_1, \dots, X_n)$ and $\forall x \in \mathbb{R}^q$, we always have $\text{EIF}(x; T_n, \mathbf{X}) \xrightarrow{a.s.} \text{IF}(x; T, F)$. However, that is not always the case, since in the previous development in Equation (\$), we made an approximation by replacing F with F_n , which could prevent the limit from being the desired one. Nevertheless, for the mean and variance estimator, we have the convergence of the EIF to the IF since we have the following:

- Let T_n be the mean estimator as defined in Example 4.3.1, then $\text{EIF}(x; T_n, \mathbf{x}) = x - \bar{\mathbf{x}}_n \xrightarrow{a.s.} x - \mathbb{E}_F[X] = \text{IF}(x; T_{\text{mean}}, F)$ by the SLLN.
- Let T_n be the variance estimator as defined in Example 4.3.2, then $\text{EIF}(x; T_n, \mathbf{x}) = \frac{n}{n+1}(x - \bar{\mathbf{x}}_n)^2 - s_n^2 \xrightarrow{a.s.} (x - \mathbb{E}_F[X])^2 - \mathbb{E}_F[X^2] + (\mathbb{E}_F[X])^2 = \text{IF}(x; T_{\text{var}}, F)$ by the SLLN.

As an example, the empirical influence function of the median does not always converge to the theoretical influence function, as proved in [20].

4.3.5 Desired properties of the influence function of an estimator

For an estimator to be considered robust, we would like its influence function to have some specific properties. For example, being bounded, meaning that a new observation cannot influence the estimator too much.

We will just mention three other interesting properties (see [29, 30]) that an influence function of a statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^p$ might have.

- Finite rejection point: $\rho^*(T, F) = \inf\{r > 0 | \forall \|x\| > r, \text{IF}(x; T, F) = 0\}$, which tells, if it is finite, that any observation that is farther than the rejection point has no influence on the statistical functional.
- Small gross-error sensitivity: $\gamma^*(T, F) = \sup_{x \in \mathbb{R}^q} \|\text{IF}(x; T, F)\|$, which provides the maximal influence an observation can have on the statistical functional. A functional T for which $\gamma^*(T, F) < +\infty$ is said to be B -robust (B stands for bounded).

- Small local-shift sensitivity: $\lambda^*(T, F) = \sup_{x, y \in \mathbb{R}^q: x \neq y} \frac{\|\text{IF}(y; T, F) - \text{IF}(x; T, F)\|}{\|y - x\|}$, which, in case it is finite, tells that the influence function cannot vary unboundedly in a neighborhood of any points, i.e., the IF is Lipschitz continuous with Lipschitz constant $\lambda^*(T, F)$. Therefore, observations in the neighborhood of some fixed point cannot have a too large influence. In practice, this can be interpreted as a quantifier of the sensitivity to rounding errors of the functional.

These properties will be obtained for the IPWM in Chapter 5.

4.3.6 Asymptotic distribution approximation

In this section, we will discuss one of the most important properties of the influence function: under some specific conditions, the asymptotic distribution of an estimator can be approximated using the influence function of this estimator.

Consider an estimator $T_n : \mathbb{R}^{q \times n} \rightarrow \mathbb{R}^p$ for which there exists a statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^p$ such that for all sample $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. from a distribution $F \in \mathcal{F}$, we have

$$T_n(\mathbf{X}) = T(F_n),$$

where $F_n = \frac{1}{n} \sum_{i=1}^n \Delta_{X_i}$ is the empirical distribution of \mathbf{X} . Moreover, suppose that T is defined not only on \mathcal{F} but also on $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ and that it is Gâteaux differentiable (see Definition 4.3.13) at F . In this case, we have (see Proposition 4.3.8) $\forall H \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ that

$$T(F + H) = T(F) + T'_F(H) + r(H),$$

where r is such that $\forall C \subseteq \mathcal{X}$, with C finite, $\limsup_{t \rightarrow 0} \sup_{H \in C} \frac{\|r(tH)\|}{t} = 0$. In particular, for $H = G - F$, where $G \in \mathcal{F}$, we obtain

$$T(G) = T(F + (G - F)) = T(F) + T'_F(G - F) + r(G - F), \quad (4.1)$$

where

$$T'_F(G - F) = \lim_{t \rightarrow 0} \frac{T(F + t(G - F)) - T(F)}{t} = \lim_{t \rightarrow 0} \frac{T((1 - t)F + tG) - T(F)}{t}.$$

Note that if $G = \Delta_x$ for some $x \in \mathcal{X}$, then

$$T'_F(\Delta_x - F) = \text{IF}(x; T, F) = \int_{\mathcal{X}} \text{IF}(y; T, F) d\Delta_x(y).$$

In fact, we can show that under the assumption that $\forall G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, $T'_F(G - F)$ can be written as $\int_{\mathcal{X}} k(x) dG(x)$ for some function $k : \mathcal{X} \rightarrow \mathbb{R}^p$, then necessarily $k(x) = \text{IF}(x; T, F)$.

Proposition 4.3.13. *Let $T : M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow \mathbb{R}^p$ be a functional, and consider $F \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$. Furthermore, suppose that T is Gâteaux differentiable at F and that $\forall G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, the equality*

$$T'_F(G - F) = \int_{\mathcal{X}} k(y) dG(y)$$

holds for some function $k : \mathcal{X} \rightarrow \mathbb{R}^p$. Then, necessarily, $k(\cdot) = IF(\cdot; T, F)$. Also, we obtain that

$$\int_{\mathcal{X}} IF(y; T, F) dF(y) = 0.$$

As a result, we have

$$T'_F(G - F) = \int_{\mathcal{X}} IF(y; T, F) d(G - F)(y) = \int_{\mathcal{X}} IF(y; T, F) dG(y).$$

Proof. Note that k must be the same for all $G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, in particular, taking $G = \Delta_x$ for some $x \in \mathcal{X}$, we have

$$\begin{aligned} T'_F(\Delta_x - F) &= \int_{\mathcal{X}} k(y) d\Delta_x(y) \\ &= k(x). \end{aligned}$$

Since $T'_F(\Delta_x - F) = IF(x; T, F)$, we can conclude that $k(\cdot) = IF(\cdot; T, F)$. Moreover, taking $G = F$, we have

$$\begin{aligned} T'_F(F - F) &= \int_{\mathcal{X}} IF(y; T, F) dF(y) \\ \Leftrightarrow T'_F(0) &= \int_{\mathcal{X}} IF(y; T, F) dF(y) \\ \Leftrightarrow \lim_{t \rightarrow 0} \frac{T(F + t \cdot 0) - T(F)}{t} &= \int_{\mathcal{X}} IF(y; T, F) dF(y) \\ \Leftrightarrow \lim_{t \rightarrow 0} \frac{T(F) - T(F)}{t} &= \int_{\mathcal{X}} IF(y; T, F) dF(y) \\ \Leftrightarrow 0 &= \int_{\mathcal{X}} IF(y; T, F) dF(y). \end{aligned}$$

□

Remark 4.3.9. The assumption that the Gâteaux derivative at F in the direction of $H = G - F$ can be written as an integral of a function with respect to H is, in fact, the definition of the derivative used by Von Mises [67] when he introduced the concept of a Taylor-like expansion for statistical functionals. It is called the Von Mises' expansion for functional.

Heuristically, this result can be seen as the following: the derivative of a statistical functional T at some distribution F in the direction of $G - F$, for some distribution G , is simply the mean of the influences of every point $x \in \mathcal{X}$ on T with respect to the distribution G . In other words, the influence of a contamination of the distribution F by a distribution G on T is the mean (with respect to G) of the influences of a contamination of F by Δ_x on T .

Remark 4.3.10. By Proposition 4.3.13, we also obtain that $\forall H \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$,

$$T'_F(H) = \int_{\mathcal{X}} \text{IF}(y; T, F) dH(y).$$

This equality follows directly by seeing H as $(H+F)-F$ and using $\int_{\mathcal{X}} \text{IF}(y; T, F) dF(y) = 0$.

The assumption made in Proposition 4.3.13 is not too restrictive as for most statistical functionals we might consider and study, the assumption will be true. As an example, let us show that it holds for the mean functional.

Proposition 4.3.14. *Let the (extended) mean functional be defined as*

$$T_{\text{mean}} : S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow \mathbb{R}^p : F \mapsto \mathbb{E}_F[X] = \int_{\mathcal{X}} x dF(x),$$

where $S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ is an \mathbb{R} -vector subspace of $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$ and consider $F \in S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$. Then we have that T_{mean} is Gâteaux differentiable at F and $\forall G \in S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$, we have

$$(T_{\text{mean}})'_F(G - F) = \int_{\mathcal{X}} \text{IF}(x; T_{\text{mean}}, F) dG(x).$$

Proof. First, let us prove that T_{mean} is Gâteaux differentiable. We have $\forall G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$

$$\begin{aligned} (T_{\text{mean}})'_F(G) &= \lim_{t \rightarrow 0} \frac{T_{\text{mean}}(F + tG) - T_{\text{mean}}(F)}{t} \\ &= \lim_{t \rightarrow 0} \frac{\mathbb{E}_{F+tG}[X] - \mathbb{E}_F[X]}{t} \\ &= \lim_{t \rightarrow 0} \frac{\mathbb{E}_F[X] + t\mathbb{E}_G[X] - \mathbb{E}_F[X]}{t} \\ &= \mathbb{E}_G[X]. \end{aligned}$$

Furthermore, we have $\forall G \in S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$,

$$\begin{aligned} (T_{\text{mean}})'_F(G - F) &= \mathbb{E}_{G-F}[X] \\ &= \mathbb{E}_G[X] - \mathbb{E}_F[X] \end{aligned}$$

$$\begin{aligned}
&= \int_{\mathcal{X}} x dG(x) - \int_{\mathcal{X}} \mathbb{E}_F[X] dG(x) \\
&= \int_{\mathcal{X}} x - \mathbb{E}_F[X] dG(x) \\
&= \int_{\mathcal{X}} \text{IF}(x; T_{\text{mean}}, F) dG(x),
\end{aligned}$$

where the last equality is the result of Example 4.3.4. \square

Substituting $T'_F(G - F)$ by $\int_{\mathcal{X}} \text{IF}(x; T, F) dG(x)$ in Equation (4.1), we have

$$T(G) = T(F) + \int_{\mathcal{X}} \text{IF}(x; T, F) dG(x) + r(G - F). \quad (4.2)$$

In particular, we can take $G = F_n$ and obtain the main property of the influence function, which provides an approximation of the asymptotic distribution of the estimator T_n .

Theorem 4.3.2. *Let $T_n : \mathbb{R}^{q \times n} \rightarrow \mathbb{R}^p$ be an estimator for which there exists a functional $T : M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow \mathbb{R}^p$, such that for all samples $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. from a distribution $F \in \mathcal{F}$, we have*

$$T_n(\mathbf{X}) = T(F_n),$$

where $F_n = \frac{1}{n} \sum_{i=1}^n \Delta_{X_i}$ is the empirical distribution of \mathbf{X} . Moreover, suppose that the following conditions are met:

1. T is Gâteaux differentiable at F and $\forall G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$,

$$T(G) = T(F) + T'_F(G - F) + r(G - F).$$

2. $\forall G \in M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$,

$$T'_F(G - F) = \int_{\mathcal{X}} \text{IF}(x; T, F) dG(x).$$

3. $\forall i, j \in \{1, \dots, p\}$, $\text{IF}(\cdot; T, F)_i \times \text{IF}(\cdot; T, F)_j \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), F)$.

4. $\sqrt{n}r(F_n - F) \xrightarrow{\mathcal{L}} 0$.

In this case, we have

$$\sqrt{n}(T_n(\mathbf{X}) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

where $\forall i, j \in \{1, \dots, p\}$

$$\Sigma_{ij} = \int_{\mathcal{X}} \text{IF}(x; T, F)_i \text{IF}(x; T, F)_j dF(x).$$

In particular, if $p = 1$,

$$\sqrt{n}(T_n(\mathbf{X}) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \int_{\mathcal{X}} \text{IF}(x; T, F)^2 dF(x)\right).$$

Proof. Using Equation (4.2) with $G = F_n$, we obtain

$$\begin{aligned} T(F_n) &= T(F) + \int_{\mathcal{X}} \text{IF}(x; T, F) dF_n(x) + r(F_n - F) \\ \Leftrightarrow T(F_n) - T(F) &= \int_{\mathcal{X}} \text{IF}(x; T, F) d\left(\frac{1}{n} \sum_{i=1}^n \Delta_{X_i}\right)(x) + r(F_n - F) \\ \Leftrightarrow T(F_n) - T(F) &= \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{X}} \text{IF}(x; T, F) d\Delta_{X_i}(x) + r(F_n - F) \\ \Leftrightarrow T(F_n) - T(F) &= \frac{1}{n} \sum_{i=1}^n \text{IF}(X_i; T, F) + r(F_n - F) \\ \Leftrightarrow \sqrt{n}(T(F_n) - T(F)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{IF}(X_i; T, F) + \sqrt{n}r(F_n - F). \end{aligned}$$

Let us write $Z_i = \text{IF}(X_i; T, F) \forall i \in \{1 \dots, n\}$, and

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n \text{IF}(X_i; T, F) = \frac{1}{n} \sum_{i=1}^n Z_i.$$

We have that $X_i \stackrel{i.i.d.}{\sim} F$, therefore the variables Z_i are also independent and identically distributed according to a distribution G . Moreover, we have $\forall i \in \{1 \dots, n\}$,

$$\mathbb{E}[Z_i] = \int_{\mathcal{X}} \text{IF}(x; T, F) dF(x) = 0,$$

and

$$\begin{aligned} \mathbb{V}[Z_i] &= \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_i - \mathbb{E}[Z_i])^T] \\ &= \mathbb{E}[Z_i Z_i^T] \\ &= \mathbb{E}[\text{IF}(X_i; T, F) \text{IF}(X_i; T, F)^T] \\ &= \Sigma. \end{aligned}$$

As a result, by the multivariate central limit theorem (see Theorem A.2.1), we have that

$$\sqrt{n}\bar{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{IF}(X_i; T, F) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Now, using the assumption that $\sqrt{nr}(F_n - F) \xrightarrow{\mathcal{L}} 0$, replacing $T(F_n)$ by $T_n(\mathbf{X})$, and applying Slutsky's theorem (see [63]), we obtain that

$$\sqrt{n}(T_n(\mathbf{X}) - T(F)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{IF}(X_i; T, F) + \sqrt{nr}(F_n - F) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

□

Remark 4.3.11. The assumption that $\sqrt{nr}(F_n - F) \xrightarrow{\mathcal{L}} 0$ seems reasonable and should hold for most cases, as intuitively the empirical distribution should converge toward the real distribution in the space $M(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R})$. Note that there exist cases where the higher order terms in the Von Mises' expansion will induce other types of asymptotic distributions (see [67]).

Let us apply the theorem to the classical mean estimator.

Example 4.3.7. Let $T_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$ be a mean estimator, and let $F \in \mathcal{F}$ be a distribution. If we consider the (extended) mean functional

$$T_{\text{mean}} : S(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{R}) \rightarrow \mathbb{R}^p : F \mapsto \mathbb{E}_F[X] = \int_{\mathcal{X}} x dF(x),$$

we have that for all samples $\mathbf{X} = (X_1, \dots, X_n) \stackrel{i.i.d.}{\sim} F$, $T_n(\mathbf{X}) = T(F_n)$, where $F_n = \frac{1}{n} \sum_{i=1}^n \Delta_{X_i}$. Then by Proposition 4.3.14, the two first conditions of Theorem 4.3.2 hold. Moreover, $\forall i, j \in \{1, \dots, p\}$,

$$\begin{aligned} \int_{\mathcal{X}} \text{IF}(x; T, F)_i \text{IF}(x; T, F)_j dF(x) &= \int_{\mathcal{X}} (x_i - (\mathbb{E}_F[X])_i)(x_j - (\mathbb{E}_F[X])_j) dF(x) \\ &= \text{Cov}[X^{(i)}, X^{(j)}], \end{aligned}$$

where $X^{(i)}$ (resp. $X^{(j)}$) are the i -th (resp. j -th) component of the vector $X \sim F$. Now, since T is a linear operator, we have that

$$\begin{aligned} T'_F(F_n - F) &= \lim_{t \rightarrow 0} \frac{T(F + t(F_n - F)) - T(F)}{t} \\ &= \lim_{t \rightarrow 0} \frac{T(F) + tT(F_n - F) - T(F)}{t} \\ &= T(F_n - F) \\ &= T(F_n) - T(F). \end{aligned}$$

Therefore, we have the following Von Mises' expansion for T :

$$T(F) + T'_F(F_n - F) = T(F_n) = T(F) + T'_F(F_n - F) + r(F_n - F),$$

and thus r is the null mapping, which implies that $\sqrt{n}r(F_n - F) \xrightarrow{\mathcal{L}} 0$. As a result, we obtain from Theorem 4.3.2 that

$$\sqrt{n}(\bar{X}_n - \mathbb{E}_F[X]) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

where $\forall i, j \in \{1, \dots, p\}$,

$$\Sigma_{ij} = \mathfrak{Cov}[X^{(i)}, X^{(j)}].$$

As expected, this gives us back the result of the multivariate central limit theorem (see Theorem A.2.1).

4.3.7 M-estimators

To conclude this chapter on statistical robustness, we will make a small discussion about a specific type of estimator. As mentioned before, we would like an estimator to be robust, which can be assessed by looking at the properties of the influence function of the estimator under consideration. In this section, we will talk about a common type of estimators obtained by a generalized maximum likelihood (which explains their name), called M-estimators. As we shall see, their influence function has a general form. Note that we will follow the definition of M-estimators proposed in [30, 32]

Let us recall the context for an estimator obtained via the maximum likelihood (see Definition 2.3.6). We consider a sample $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. from a univariate distribution $F \in \mathcal{F}$, which can either be discrete or continuous and depends on a parameter vector $\theta \in \Theta$ (Θ is considered as an open subset of some space \mathcal{Z}). We will denote with $f(\cdot; \theta)$ the probability mass function or density of F . The goal is to provide an estimator for θ . The maximum likelihood estimator $T_n(\mathbf{X})$, which will be written $\hat{\theta}_{ML}(\mathbf{X})$ in this case, is such that

$$\hat{\theta}_{ML}(\mathbf{X}) = \arg \min_{\theta \in \Theta} (-\ell(\theta; X_1, \dots, X_n)),$$

where ℓ is the log-likelihood function defined as

$$\ell : \Theta \rightarrow \mathbb{R} : \theta \mapsto \ell(\theta; X_1, \dots, X_n) = \sum_{i=1}^n \ln(f(X_i; \theta)).$$

If we write it in another way, we obtain

$$\hat{\theta}_{ML}(\mathbf{X}) = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \rho(X_i, \theta),$$

where

$$\rho : \mathcal{X} \times \Theta \rightarrow \mathbb{R} : (x, \theta) \mapsto -\ln(f(x; \theta)).$$

The idea behind generalized maximum likelihood is to consider a specific function ρ that will produce an estimator that has some robust properties. Concerning the function ρ , we

assume that it is strictly convex on Θ and differentiable with respect to its second argument on Θ , with the function

$$\psi : \mathcal{X} \times \Theta \rightarrow \mathbb{R} : (x, \theta) \mapsto \frac{\partial}{\partial \theta} \rho(x, \theta)$$

as derivative. Moreover, we suppose that there exists a minimum (which will be global), that is,

$$\arg \min_{\theta \in \Theta} \sum_{i=1}^n \rho(X_i, \theta)$$

exists. From the preceding definitions, we obtain that the generalized maximum likelihood estimator is a solution, in t , of the equation

$$\sum_{i=1}^n \psi(X_i, t) = 0. \quad (4.3)$$

Note that we will identify the function ψ with the obtained generalized maximum likelihood estimator it defines, hence we will use the notation $\hat{\theta}_\psi(\mathbf{X})$.

Remark 4.3.12. If we do not make the assumption that ρ is strictly convex, then there may exist other solutions to Equation (4.3) than $\hat{\theta}_\psi(\mathbf{X})$, which will represent the local extrema of $\sum_{i=1}^n \rho(X_i, \theta)$. This is why, in order to obtain a unique solution to Equation (4.3) (if it exists), we will only consider a strictly convex function ρ .

Let us consider $F_n = \frac{1}{n} \sum_{i=1}^n \Delta_{X_i}$ to be the empirical distribution of \mathbf{X} , and define the functional T_ψ such that for a distribution G , $T_\psi(G)$ is the solution to the following equation in t ,

$$\int_{\mathcal{X}} \psi(x, t) dG(x) = 0. \quad (4.4)$$

Possibly after restriction to a subset of \mathcal{F} , we assume that Equation (4.4) possesses at least one solution for all distributions G . Furthermore, we assume that for a distribution F , depending on a parameter θ , we have $T_\psi(F) = \theta$, that is, T_ψ is Fisher-consistent. Now, consider the functional T_ψ at F_n , we get that it satisfies Equation (4.3), therefore, we obtain $T_\psi(F_n) = \hat{\theta}_\psi(\mathbf{X})$.

What we would like is to obtain the influence function of the functional T_ψ . For that, we consider $F_t = (1 - t)F + t\Delta_x$ with $t \in [0, 1]$ and $x \in \mathcal{X}$. Using Equation (4.4), we have

$$\int_{\mathcal{X}} \psi(y, T_\psi(F_t)) dF_t(y) = 0 \Rightarrow \frac{\partial}{\partial t} \int_{\mathcal{X}} \psi(y, T_\psi(F_t)) dF_t(y) = 0,$$

where

$$\frac{\partial}{\partial t} \int_{\mathcal{X}} \psi(y, T_\psi(F_t)) dF_t(y)$$

$$\begin{aligned}
 &= \frac{\partial}{\partial t}(1-t) \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y) + \frac{\partial}{\partial t} t \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) d\Delta_x(y) \\
 &= \underbrace{\frac{\partial}{\partial t}(1-t) \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y)}_{A(t)} + \underbrace{\frac{\partial}{\partial t} t \psi(x, T_{\psi}(F_t))}_{B(t)}.
 \end{aligned}$$

For A , we obtain

$$\begin{aligned}
 A(t) &= - \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y) + (1-t) \frac{\partial}{\partial t} \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y) \\
 &= - \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y) + (1-t) \int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(y, T_{\psi}(F_t)) dF(y) \quad (\spadesuit) \\
 &= - \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF(y) + (1-t) \int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_{\psi}(F_t)} \frac{\partial}{\partial t} T_{\psi}(F_t) dF(y),
 \end{aligned}$$

where for Equation (\spadesuit) , we assume that the order of integration and derivation can be swapped. Also, we make the assumption that ψ is differentiable with respect to its second argument. Likewise, for B , we have

$$\begin{aligned}
 B(t) &= \psi(x, T_{\psi}(F_t)) + t \frac{\partial}{\partial t} \psi(x, T_{\psi}(F_t)) \\
 &= \psi(x, T_{\psi}(F_t)) + t \frac{\partial}{\partial u} \psi(x, u) \Big|_{u=T_{\psi}(F_t)} \frac{\partial}{\partial t} T_{\psi}(F_t).
 \end{aligned}$$

Evaluating at $t = 0$, we get

$$\begin{aligned}
 &\frac{\partial}{\partial t} \int_{\mathcal{X}} \psi(y, T_{\psi}(F_t)) dF_t(y) \Big|_{t=0} = 0 \\
 &\Leftrightarrow A(0) + B(0) = 0 \\
 &\Leftrightarrow - \underbrace{\int_{\mathcal{X}} \psi(y, T_{\psi}(F)) dF(y)}_{=0 \text{ by Equation (4.4)}} + \frac{\partial}{\partial t} T_{\psi}(F_t) \Big|_{t=0} \int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_{\psi}(F)} dF(y) + \psi(x, T_{\psi}(F)) = 0 \\
 &\Leftrightarrow \frac{\partial}{\partial t} T_{\psi}(F_t) \Big|_{t=0} = - \frac{\psi(x, T_{\psi}(F))}{\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_{\psi}(F)} dF(y)}.
 \end{aligned}$$

As mentioned in Remark 4.3.7,

$$\frac{\partial}{\partial t} T_{\psi}(F_t) \Big|_{t=0} = \text{IF}(x; T_{\psi}, F),$$

therefore, we have the following general form for the influence function of an M-estimator based on the function ψ :

$$\text{IF}(x; T_{\psi}, F) = - \frac{\psi(x, T_{\psi}(F))}{\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_{\psi}(F)} dF(y)},$$

where we assume that $\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_\psi(F)} dF(y)$ exists and is not null.

We see that the influence function is proportional to the function ψ . As a result, we can choose ψ such that the influence function has specific properties such as B-robustness (i.e., $\gamma^*(T_\psi, F) = \sup_{x \in \mathbb{R}} |\text{IF}(x; T_\psi, F)| < +\infty$). Also, supposing that the conditions of Theorem 4.3.2 are met, we obtain the asymptotic variance for $\sqrt{n}(\hat{\theta}_\psi(\mathbf{X}) - T_\psi(F))$ as

$$\int_{\mathcal{X}} \text{IF}(x; T_\psi, F)^2 dF(x) = \frac{\int_{\mathcal{X}} \psi(x, T_\psi(F))^2 dF(x)}{\left(\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(y, u) \Big|_{u=T_\psi(F)} dF(y) \right)^2}.$$

In the particular case where θ is a location parameter, that is, the density or probability mass function $f(\cdot; \theta)$ satisfies

$$f(x; \theta) = g(x - \theta),$$

for some function g . Then, we have $\psi(x, t) = \psi(x - t)$, and we obtain the following:

$$\text{IF}(x; T_\psi, F) = \frac{\psi(x - T_\psi(F))}{\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(u) \Big|_{u=y-T_\psi(F)} dF(y)}.$$

Let us finish this section with an example of an M-estimator.

Example 4.3.8. Suppose that we have a sample $\mathbf{X} = (X_1, \dots, X_n)$ such that $X_i \stackrel{i.i.d.}{\sim} F$, where F is a univariate continuous distribution depending on a location parameter $\mu \in \mathbb{R}$. We consider the function

$$\rho : \mathbb{R} \rightarrow \mathbb{R} : (x - t) \mapsto (x - t)^2 \text{ and } \psi : \mathbb{R} \rightarrow \mathbb{R} : (x - t) \mapsto -2(x - t).$$

It is clear that ρ is strictly convex, differentiable with respect to t and that a global minimum exists. The functional T_ψ at a distribution G is the solution, in t , of the equation

$$\int_{\mathcal{X}} -2(x - t) dG(x) = 0.$$

Since we have $\int_{\mathcal{X}} -2(x - t) dG(x) = -2\mathbb{E}_G[X] + 2t$, we obtain that $T_\psi : \mathcal{F} \rightarrow \mathbb{R} : G \mapsto \mathbb{E}_G[X]$, that is, T_ψ is the classical mean functional. Also, T_ψ is Fisher-consistent and we have

$$T_\psi(F_n) = \hat{\theta}_\psi(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i.$$

Let us now compute the influence function based on the formula for M-estimators. We have

$$\begin{aligned}
 \text{IF}(x; T_\psi, F) &= \frac{\psi(x - T_\psi(F))}{\int_{\mathcal{X}} \frac{\partial}{\partial u} \psi(u) \Big|_{u=y-T_\psi(F)} dF(y)} \\
 &= \frac{-2(x - \mathbb{E}_F[X])}{\int_{\mathcal{X}} -2dF(y)} \\
 &= x - \mathbb{E}_F[X],
 \end{aligned}$$


which is, indeed, the influence function of the mean functional. We obtain the same result for the asymptotic variance of $\sqrt{n}(\hat{\theta}_\psi(\mathbf{X}) - T_\psi(F))$ as in Example 4.3.7, but in one dimension.

Chapter 5

Robustness of the inverse probability weighted mean estimator

5.1 State of the art

In this section, we will discuss the state of the art in robustness against outliers of the inverse probability weighted mean estimator.

While inverse probability weighting is a way of creating estimators of different types (e.g., mean and variance), the inverse probability weighted mean (IPWM) or its improved version, the augmented inverse probability weighted mean (AIPWM), are the ones that have been of main interest for their practical use, as in causal inference we are mainly interested in the average treatment effect (ATE) [10, 12, 31, 33, 35, 54, 59, 70], which is computed as $\mu_1 - \mu_0 = \mathbb{E}[Y^{a=1} - Y^{a=0}]$. The IPWM is known to lack robustness against model specifications, especially in the case where some probabilities $\mathbb{P}(A_i = a|L_i)$ are close to zero (i.e., have a low propensity score), meaning that the associated weights tend towards infinity. Therefore, many research was performed on a way of creating a more robust estimator (in the sense of model specification) that has the so-called doubly robust property [10, 35, 36, 59]. In addition, the IPWM also has some flaws concerning its variance, as its value can vary greatly depending on the dataset used to compute it. Therefore, some alterations and modifications to the estimator were proposed to reduce its variance [33, 40]. Finally, some implementations of IPW are proposed in libraries for the  software such as in [64]. However, only a few publications are concerned with robustness against outliers, in which we are interested in statistical robustness, and even then they are mainly focused on an empirical point of view [9, 71]. There seem to be no rigorous mathematical foundations of robustness in this context, which results in the fact that the few articles speaking about the robustness against outliers of the inverse probability weighted mean do not provide very detailed proofs, nor well-defined theorems. Moreover, to the best of our knowledge, no rigorous work seems to have been done on the breakdown point, nor the influence function (either empirically or theoretically) of the

IPWM. Therefore, this chapter aims to apply the concepts developed in Chapter 4 to the IPWM.

5.2 Breakdown point of the IPWM

Let us dive in directly into the results concerning the breakdown point of the IPWM.

Proposition 5.2.1 (Empirical breakdown point of IPWM). *Let $\mathbf{x} = \{(l_i, a_i, y_i)\}_{i=1}^n$ be a sample of observations, where l_i is a vector of covariates, a_i is the treatment and y_i is the outcome. Consider the estimator of the inverse probability weighted mean*

$$\hat{\mu}_{a,n}^{IPW} : (\mathbb{R}^p \times \{0, 1\} \times \mathbb{R})^n \rightarrow \mathbb{R} : \{(L_i, A_i, Y_i)\}_{i=1}^n \mapsto \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\hat{\mathbb{P}}_n(A_i|L_i)}$$

where $\hat{\mathbb{P}}_n(A_i|L_i)$ is calculated using any method (as long as we do not have 0 as an estimate which would result in an infinite weight), then

$$\varepsilon(\hat{\mu}_{a,n}^{IPW}, \mathbf{x}) = \frac{1}{n}.$$

Proof. This results directly from the fact that if only a single outcome y_j , such that $a_j = a$, is corrupted and given a value that tends towards infinity, then the inverse probability weighted mean will diverge to infinity regardless of the weight $1/\hat{\mathbb{P}}_n(a_i|l_i)$, which has a value in $[1, +\infty[$, associated with this observation.

Therefore, only corrupting a single observation can cause the estimator to diverge, which implies that

$$\varepsilon(\hat{\mu}_{a,n}^{IPW}, \mathbf{x}) = \frac{1}{n}.$$

□

So we showed that the empirical breakdown point of the IPWM is the same as that of the regular mean estimator, which was expected since both estimators are constructed similarly (the normal mean is just the sum of all outcomes divided by n , while the IPWM is the sum of all outcomes times a weight divided by n).

Proposition 5.2.2 (Theoretical breakdown point of IPWM). *Let $(L, A, Y) \sim F \in \mathcal{F}$, be a random vector where L is a vector of covariates, A the treatment variable and Y the outcome variable. Consider the statistical functional of the IPWM defined as*

$$T_a^{IPWM} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \mathbb{E}_F [W^{A|L} \mathbb{1}_{\{A=a\}} Y],$$

for $a \in \{0, 1\}$, and a probability distribution $\mathcal{P} \in \mathcal{F}$, then

$$\varepsilon^*(T_a^{IPWM}, \mathcal{P}, d_K) = 0.$$

Proof. Using Proposition 4.2.2, we know that $\forall \mathcal{P} \in \mathcal{F}$,

$$\varepsilon^*(T_{mean}, \mathcal{P}, d_K) = 0.$$

So, if we let $Z = W^{A|L} \mathbb{1}_{\{A=a\}} Y$, then Z is a random variable that takes values in \mathbb{R} such that $Z \sim F_{\mathcal{P}}$ for some distribution $F_{\mathcal{P}} \in \mathcal{F}$. Therefore, we get that

$$\varepsilon^*(T_a^{IPWM}, \mathcal{P}, d_K) = \varepsilon^*(T_{mean}, F_{\mathcal{P}}, d_K) = 0.$$

□

The proof of Proposition 5.2.2 relies on the fact that the statistical functional of the IPWM at some distribution \mathcal{P} is just the statistical functional of the mean at a distribution $F_{\mathcal{P}}$ depending on \mathcal{P} . Therefore, we can use the results that are known about the mean.

5.3 Influence function of IPWM

Let us now provide the results concerning the influence function of the IPWM.

5.3.1 Computation of the influence function

First, we will obtain both the empirical and theoretical influence functions.

Proposition 5.3.1 (Empirical influence function of IPWM). *Let $\mathbf{x} = \{(l_i, a_i, y_i)\}_{i=1}^n$ be a sample of observations, where l_i is a vector of covariates, a_i is the treatment and y_i is the outcome. Consider the estimator of the inverse probability weighted mean*

$$\hat{\mu}_{a,n}^{IPW} : (\mathbb{R}^p \times \{0, 1\} \times \mathbb{R})^n \rightarrow \mathbb{R} : \{(L_i, A_i, Y_i)\}_{i=1}^n \mapsto \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\hat{\mathbb{P}}_n(A_i|L_i)}$$

where $\hat{\mathbb{P}}_n(A_i|L_i)$ is calculated using any method (as long as we do not have 0 as an estimate which would result as an infinite weight), then the empirical influence function of the IPWM at $x = (\ell, \alpha, v) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ is given by

$$\text{EIF}((\ell, \alpha, v); \hat{\mu}_{a,n}^{IPW}, \mathbf{x}) = \sum_{i=1}^n y_i \mathbb{1}_{\{a_i=a\}} \left(\frac{1}{\hat{\mathbb{P}}_{n+1}(a_i|l_i)} - \frac{(n+1)}{n \hat{\mathbb{P}}_n(a_i|l_i)} \right) + \frac{v \mathbb{1}_{\{\alpha=a\}}}{\hat{\mathbb{P}}_{n+1}(\alpha|\ell)}.$$

Proof. Consider $x = (\ell, \alpha, v) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ and a sample $\mathbf{x} = \{(l_i, a_i, y_i)\}_{i=1}^n$ of size $n \in \mathbb{N}_0$, then we have

$$\begin{aligned} \text{EIF}((\ell, \alpha, v); \hat{\mu}_{a,n}^{IPW}, \mathbf{x}) &= \frac{\mu_{a,n+1}^{IPWM}(\mathbf{x} \cup \{(\ell, \alpha, v)\}) - \hat{\mu}_{a,n}^{IPW}(\mathbf{x})}{\frac{1}{n+1}} \\ &= (n+1) \left(\sum_{i=1}^n \frac{y_i \mathbb{1}_{\{a_i=a\}}}{(n+1) \hat{\mathbb{P}}_{n+1}(a_i|l_i)} + \frac{v \mathbb{1}_{\{\alpha=a\}}}{(n+1) \hat{\mathbb{P}}_{n+1}(\alpha|\ell)} - \sum_{i=1}^n \frac{y_i \mathbb{1}_{\{a_i=a\}}}{n \hat{\mathbb{P}}_n(a_i|l_i)} \right) \\ &= \sum_{i=1}^n \frac{y_i \mathbb{1}_{\{a_i=a\}}}{\hat{\mathbb{P}}_{n+1}(a_i|l_i)} + \frac{v \mathbb{1}_{\{\alpha=a\}}}{\hat{\mathbb{P}}_{n+1}(\alpha|\ell)} - \sum_{i=1}^n \frac{(n+1) y_i \mathbb{1}_{\{a_i=a\}}}{n \hat{\mathbb{P}}_n(a_i|l_i)} \\ &= \sum_{i=1}^n y_i \mathbb{1}_{\{a_i=a\}} \left(\frac{1}{\hat{\mathbb{P}}_{n+1}(a_i|l_i)} - \frac{(n+1)}{n \hat{\mathbb{P}}_n(a_i|l_i)} \right) + \frac{v \mathbb{1}_{\{\alpha=a\}}}{\hat{\mathbb{P}}_{n+1}(\alpha|\ell)} \end{aligned}$$

□

Proposition 5.3.2 (Theoretical influence function of IPWM). *Let $(L, A, Y) \sim F \in \mathcal{F}$, be a random vector where L is a vector of covariates, A the treatment variable and Y the outcome variable. Consider the statistical functional of the IPWM defined as*

$$T_a^{IPWM} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \mathbb{E}_F [W^{A|L} \mathbb{1}_{\{A=a\}} Y],$$

for $a \in \{0, 1\}$, and a probability distribution $\mathcal{P} \in \mathcal{F}$, then the theoretical influence function of the IPWM at $x = (\ell, \alpha, v) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ is given by

$$\text{IF}((\ell, \alpha, v); T_a^{IPWM}, \mathcal{P}) = \begin{cases} \frac{v}{\mathbb{P}(A=\alpha|L=\ell)} - T_a^{IPWM}(\mathcal{P}) & \text{if } \alpha = a \\ -T_a^{IPWM}(\mathcal{P}) & \text{otherwise} \end{cases}.$$

Proof. Consider $x = (\ell, \alpha, v) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$, and write $Z = W^{A|L} \mathbb{1}_{\{A=a\}} Y$. Then

$$\begin{aligned} \text{IF}((\ell, \alpha, v); T_a^{IPWM}, \mathcal{P}) &= \lim_{\varepsilon \rightarrow 0} \frac{T_a^{IPWM}((1-\varepsilon)\mathcal{P} + \varepsilon\Delta_x) - T_a^{IPWM}(\mathcal{P})}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\mathbb{E}_{(1-\varepsilon)\mathcal{P} + \varepsilon\Delta_x}[Z] - \mathbb{E}_{\mathcal{P}}[Z]}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{(1-\varepsilon)\mathbb{E}_{\mathcal{P}}[Z] + \varepsilon\mathbb{E}_{\Delta_x}[Z] - \mathbb{E}_{\mathcal{P}}[Z]}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{-\varepsilon\mathbb{E}_{\mathcal{P}}[Z] + \varepsilon\mathbb{E}_{\Delta_x}[Z]}{\varepsilon} \\ &= -\mathbb{E}_{\mathcal{P}}[Z] + \mathbb{E}_{\Delta_x}[Z] \\ &= W^{\alpha|\ell} \mathbb{1}_{\{\alpha=a\}} v - T_a^{IPWM}(\mathcal{P}) \\ &= \frac{\mathbb{1}_{\{\alpha=a\}} v}{\mathbb{P}(A=\alpha|L=\ell)} - T_a^{IPWM}(\mathcal{P}) \end{aligned}$$

$$= \begin{cases} \frac{v}{\mathbb{P}(A=\alpha|L=\ell)} - T_a^{IPWM}(\mathcal{P}) & \text{if } \alpha = a \\ -T_a^{IPWM}(\mathcal{P}) & \text{otherwise} \end{cases}.$$

□

The form of the theoretical influence function of the IPWM, in the case $\alpha = a$, is what we could expect: it is the influence function of the statistical functional of the mean for a specific distribution.

5.3.2 Properties of the influence function of the IPWM

Now that we have the influence function of the IPWM, we can study its properties. We will look at the rejection point, gross-error sensitivity, and local-shift sensitivity as defined in Section 4.3.5.

Remark 5.3.1. For the following, we will need to define a norm on $\mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$. We can see $\mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ as a subset of \mathbb{R}^{p+2} and, therefore, use the classical norm of \mathbb{R}^{p+2} :

$$\|\cdot\| : \mathbb{R}^{p+2} \rightarrow [0, +\infty[: (x_1, \dots, x_{p+2}) \mapsto \sqrt{\sum_{i=1}^{p+2} x_i^2}.$$

Also, the classical addition and scalar multiplication of \mathbb{R}^{p+2} will be used.

Proposition 5.3.3. *Let $(L, A, Y) \sim F \in \mathcal{F}$, be a random vector where L is a set of p variables, A the treatment variable and Y the outcome variable. Consider the statistical functional of the IPWM defined as*

$$T_a^{IPWM} : \mathcal{F} \rightarrow \mathbb{R} : F \mapsto \mathbb{E}_F [W^{A|L} \mathbb{1}_{\{A=a\}} Y],$$

for $a \in \{0, 1\}$, and a probability distribution $\mathcal{P} \in \mathcal{F}$, then

1. $\rho^*(T_a^{IPWM}, \mathcal{P}) = +\infty$.
2. $\gamma^*(T_a^{IPWM}, \mathcal{P}) = +\infty$.
3. $\lambda^*(T_a^{IPWM}, \mathcal{P}) = +\infty$.

Proof. Let $a \in \{0, 1\}$ and $\mathcal{P} \in \mathcal{F}$. Using Proposition 5.3.2, we get that the influence function of T_a^{IPW} at $(\ell, \alpha, v) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ is

$$\text{IF}((\ell, \alpha, v); T_a^{IPWM}, \mathcal{P}) = \begin{cases} \frac{v}{\mathbb{P}(A=\alpha|L=\ell)} - T_a^{IPWM}(\mathcal{P}) & \text{if } \alpha = a \\ -T_a^{IPWM}(\mathcal{P}) & \text{otherwise} \end{cases}.$$

From this formula it is clear that

$$\gamma^*(T_a^{IPWM}, \mathcal{P}) = \sup_{(x, y, z) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}} |\text{IF}((x, y, z); T_a^{IPW}, \mathcal{P})| = +\infty$$

by considering the sequence $((x, a, z_n))_{n \in \mathbb{N}_0}$ where $x \in \mathbb{R}^p$ and $z_n \in \mathbb{R} \xrightarrow{n \rightarrow +\infty} +\infty$. Then

$$|\text{IF}((x, a, z_n); T_a^{IPW}, \mathcal{P})| = \left| \frac{z_n}{\mathbb{P}(A = a|L = x)} - T_a^{IPWM}(\mathcal{P}) \right| \xrightarrow{n \rightarrow +\infty} +\infty,$$

because T_a^{IPWM} and $\mathbb{P}(A = a|L = x)$ are constants.

Concerning the rejection point, let us prove by contradiction that it is not finite. Suppose that $\rho^*(T_a^{IPWM}, \mathcal{P}) = N$ for some $N \in \mathbb{N}_0$, that is, $\forall (x, y, z) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ such that $\|(x, y, z)\| > N$,

$$\text{IF}((x, y, z); T_a^{IPW}, \mathcal{P}) = 0.$$

Let us take $(0, a, z)$ where $z \in \mathbb{R}$ is such that $|z| > N + 1$, and

$$\frac{z}{\mathbb{P}(A = a|L = 0)} \neq T_a^{IPWM}(\mathcal{P}).$$

In this case, we obtain that $\|(0, a, z)\| = \sqrt{a^2 + (N + 1)^2} > N$ and

$$\begin{aligned} \text{IF}((0, a, z); T_a^{IPW}, \mathcal{P}) &= \frac{z}{\mathbb{P}(A = a|L = 0)} - T_a^{IPWM}(\mathcal{P}) \\ &\neq 0. \end{aligned}$$

This contradicts the fact that

$$\rho^*(T_a^{IPWM}, \mathcal{P}) = \inf\{r > 0 | \forall \|x\| > r, \text{IF}(x; T_a^{IPWM}, \mathcal{P}) = 0\} = N.$$

As a result, we can conclude that $\rho^*(T_a^{IPWM}, \mathcal{P}) = +\infty$.

Finally, let us show that $\lambda^*(T_a^{IPWM}, \mathcal{P}) = +\infty$, that is,

$$\sup_{(x, y, z) \neq (u, v, w) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}} \frac{|\text{IF}((x, y, z); T_a^{IPW}, \mathcal{P}) - \text{IF}((u, v, w); T_a^{IPW}, \mathcal{P})|}{\|(x, y, z) - (u, v, w)\|} = +\infty.$$

Consider $((x, a, z_n))_{n \in \mathbb{N}_0}, ((x, 1 - a, z_n))_{n \in \mathbb{N}_0}$ to be two sequences such that $x \in \mathbb{R}^p, z_n \in \mathbb{R}$, and $z_n \xrightarrow{n \rightarrow +\infty} +\infty$. We have

$$\begin{aligned} &\frac{|\text{IF}((x, a, z_n); T_a^{IPW}, \mathcal{P}) - \text{IF}((x, 1 - a, z_n); T_a^{IPW}, \mathcal{P})|}{\|(x, a, z_n) - (x, 1 - a, z_n)\|} \\ &= \frac{\left| \frac{z_n}{\mathbb{P}(A=a|L=x)} - T_a^{IPWM}(\mathcal{P}) - (-T_a^{IPWM}(\mathcal{P})) \right|}{\|(x, a, z_n) - (x, 1 - a, z_n)\|} \\ &= \frac{\left| \frac{z_n}{\mathbb{P}(A=a|L=x)} \right|}{|2a - 1|} \\ &= \left| \frac{z_n}{\mathbb{P}(A = a|L = x)} \right| \xrightarrow{n \rightarrow +\infty} +\infty \text{ since } a \in \{0, 1\}. \end{aligned}$$

From this, we can conclude that $\lambda^*(T_a^{IPWM}, \mathcal{P}) = +\infty$. □

CHAPTER 5. ROBUSTNESS OF THE INVERSE PROBABILITY WEIGHTED MEAN ESTIMATOR

All the properties in terms of robustness of the inverse probability weighted mean show that it is not a robust estimator.

Chapter 6

Simulations and empirical analysis

In order to illustrate and apply the theory we have developed in the preceding chapters, we will devote this chapter to simulations. To begin with, we will dive into the generation of causal data. Then, we will use those data to perform simulations to study the empirical robustness of the estimator of the inverse probability weighted mean. Note that this chapter is mainly based on [44, 45].

6.1 Causal data

Let us start with the main problem we face when we want to generate causal data, that is, what do we mean by causal data? In Chapter 2, we have formally defined the concept of causality for random variables. However, there is no clear definition of a dataset that contains causal relations. Here, we will consider that a causal dataset is a dataset \mathcal{D} of the form

$$\mathcal{D} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad n, p \in \mathbb{N}_0,$$

where each row is an observation of the variables X_1, \dots, X_p that satisfies a structural causal model. A structural causal model is a 5-tuple $(\mathcal{V}, \mathcal{C}, \mathcal{E}, \mathcal{F}, \mathcal{S})$ where

- $\mathcal{V} = \{X_1, \dots, X_n\}$ is a set of random variables/vectors.
- $\mathcal{E} = \{\varepsilon_i \mid i \in \{1, \dots, p\}\}$ is a set of random variables independent from each other and from \mathcal{V} .
- $(\mathcal{V}, \mathcal{C})$ is a directed acyclic graph, where $\mathcal{C} \subseteq \mathcal{V}^2$ is a set of edges representing causal relationships.
- $\mathcal{F} = \{f_i \mid i \in \{1, \dots, p\}\}$ is a set of functions such that each f_i takes as argument the parents of the variable X_i in $(\mathcal{V}, \mathcal{C})$ and the random variable $\varepsilon_i \in \mathcal{E}$.

- $\forall i \in \{1, \dots, p\}$, let $\text{Parents}(X_i) = \{X_{i_1}, \dots, X_{i_q}\}$, $1 \leq i_1 < \dots < i_q \leq n$, $q \in \{1, \dots, p\}$. $\mathcal{S} = \{X_i = f_i(X_{i_1}, \dots, X_{i_q}, \varepsilon_i) \mid i \in \{1, \dots, p\}\}$ is a system of equations representing the relationships between the variables.

With a structural causal model, we model causal relationships through a DAG that tells us which variables have an influence on which other variables. We are using a DAG to avoid any cyclic relation as this is not of interest in this thesis. To quantify the influence that a variable X_i has, we use a system of equations and a set functions that describe how each of the parents of X_i in the DAG, i.e., the variables that have a causal relationship with it, modify its value. To still be able to represent the randomness of each X_i , we use the random variable ε_i that takes part in the equation that defines X_i .

Note that this is just a way to represent specific causal relationships, and there may be many others. As an example, we could consider interferences between the random variables ε_i , or even consider cyclic relationships between the random variables X_i .

Let us illustrate a structural causal model with an example.

Example 6.1.1. At First, we just consider three variables X, Y and Z , so $\mathcal{V} = \{X, Y, Z\}$. We also consider the causal graph in Figure 6.1.

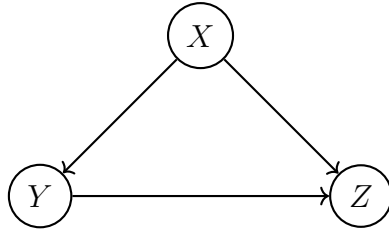


Figure 6.1: Causal graph of a simple structural causal model.

Let $\mathcal{E} = \{\varepsilon_X, \varepsilon_Y, \varepsilon_Z\}$ where each $\varepsilon \in \mathcal{E}$ is distributed according to $\mathcal{N}(0, 1)$ and is independent from each other element of \mathcal{E} and \mathcal{V} .

Note that, we can add the variables in \mathcal{E} on our causal graph as in Figure 6.2.

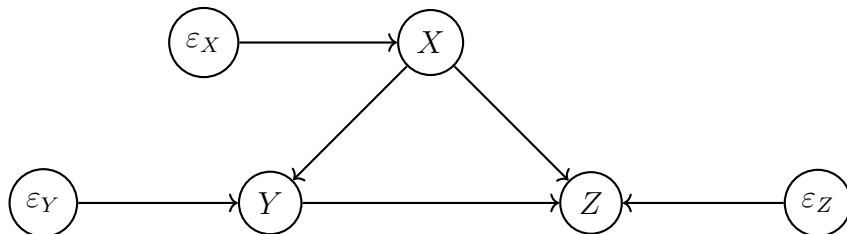


Figure 6.2: Extended causal graph of a simple structural causal model.

Moreover, consider the following system of equations:¹

$$\mathcal{S} = \begin{cases} X = \varepsilon_X^2 \\ Y = \sin(X) + \sqrt[3]{\varepsilon_Y} \\ Z = XY + \varepsilon_Z \end{cases}.$$

What precedes defines a structural causal model, and if we are able to generate data from it and obtain a dataset \mathcal{D} , we can study all the related causal questions we might have.

6.2 Generation of causal data

Now we know what we want to generate, that is, a dataset \mathcal{D} obtained from a structural causal model $(\mathcal{V}, \mathcal{C}, \mathcal{E}, \mathcal{F}, \mathcal{S})$. Therefore, the next logical question is how to generate this dataset?

In this section, we provide a possible implementation (pseudo-code) of such an algorithm. Note that the implementation in the  language is provided in Appendix B.1.

The idea is pretty straightforward and can be resumed as in Algorithm 1.

Algorithm 1 Generation of a causal dataset

Require: $N \in \mathbb{N}_0$ a number of observations to generate, and $(\mathcal{V}, \mathcal{C}, \mathcal{E}, \mathcal{F}, \mathcal{S})$ a structural causal model

Ensure: \mathcal{D} is a causal dataset of N observations obtained from $(\mathcal{V}, \mathcal{C}, \mathcal{E}, \mathcal{F}, \mathcal{S})$

```

1: procedure GENERATE_CAUSAL_DATASET( $N, \mathcal{V}, \mathcal{C}, \mathcal{E}, \mathcal{F}$ )
2:    $\mathcal{D} \leftarrow \text{NULL}$ 
3:    $\text{parents} \leftarrow \text{NULL}$ 
4:   for  $X \in \mathcal{V}$  do
5:      $\text{parents}[X] \leftarrow \text{get\_parents}(X, \mathcal{C})$  ▷ Parent variables of  $X$  in  $(\mathcal{V}, \mathcal{C})$ 
6:   end for
7:    $\text{ordered\_vars} \leftarrow \text{topological\_sort}(\mathcal{V}, \mathcal{C})$  ▷ Topological sort of  $\mathcal{V}$ 
8:    $i \leftarrow 0$ 
9:   while  $i < N$  do
10:     $\mathcal{D}[i] \leftarrow \text{NULL}$ 
11:    for  $X \in \text{ordered\_vars}$  do
12:       $\text{current\_function} \leftarrow \mathcal{F}[X]$  ▷ Function  $f \in \mathcal{F}$  associated with  $X$ 
13:       $\text{current\_epsilon} \leftarrow \mathcal{E}[X]$  ▷ Variable  $\varepsilon \in \mathcal{E}$  associated with  $X$ 
14:       $\text{epsilon\_value} \leftarrow \text{generate}(\text{current\_epsilon})$  ▷ Generate a value from  $\varepsilon$ 
15:       $\text{current\_parents} \leftarrow \text{NULL}$ 
16:      for  $Y \in \text{parents}[X]$  do
17:         $\text{current\_parents}[Y] \leftarrow \mathcal{D}[i][Y]$  ▷ Generated value of  $Y$  for  $\mathcal{D}[i]$ 

```

¹The set \mathcal{F} can be obtained from \mathcal{S} .

```

18:         end for
19:          $\mathcal{D}[i][X] \leftarrow \text{current\_function}(\text{current\_parents}, \text{epsilon\_value})$ 
20:     end for
21: end while
22: return  $\mathcal{D}$ 
23: end procedure

```

For each observation, the algorithm generates the values of the different variables based on a topological sort of them. This sort ensures that the value of any variable X can be computed, as all the other variables on which X depends have already been generated. Note that this algorithm is general as it only requires one to have a causal graph and the equations defining the variables. The user is free to give any kind of generating algorithm for the random variables in \mathcal{E} and any kind of functions for \mathcal{F} . As a result, we obtain that, despite the simplicity of the algorithm, we can generate a large variety of causal datasets using the algorithm.

A possible algorithm of the function `get_parents` is described in Algorithm 2. Concerning the function `topological_sort`, we refer to the implementation provided in [18]. The `generate` function is meant to be given by the user.

Algorithm 2 Computing parent nodes of a node

Require: V a vectrex, and \mathcal{C} a set of edges


Ensure: \mathcal{P} is the set of parent nodes of V in \mathcal{C} .

```

1: procedure GET_PARENTS( $V, \mathcal{C}$ )
2:    $\mathcal{P} \leftarrow \text{NULL}$ 
3:   for  $(V_1, V_2) \in \mathcal{C}$  do
4:     if  $V = V_2$  then
5:        $\mathcal{P} \leftarrow \text{append}(\mathcal{P}, V_1)$  ▷ Add  $V_1$  in  $\mathcal{P}$ 
6:     end if
7:   end for
8:    $\mathcal{P} \leftarrow \text{unique}(\mathcal{P})$  ▷ Delete duplicates
9:   return  $\mathcal{P}$ 
10: end procedure

```

6.3 Simulations

We now have everything we need to generate a causal dataset. Using that, we can visualize the empirical robustness properties of the estimator of the inverse probability weighted mean in different setups. In addition, we can check whether the results we obtain are consistent with what the theory tells us. Note that the  script for all simulations is given in Appendix B.2.

Recall that the inverse probability weighted mean is used in a context where we want to analyze the causal effect of a treatment on the outcome of a patient. For each patient, we are in possession of a vector of covariates $L = (X_1, \dots, X_p)$, $p \in \mathbb{N}_0$, a binary variable A indicating whether the patient has received treatment or not, and an outcome Y . In this context, we have the causal graph given in Figure 6.3.

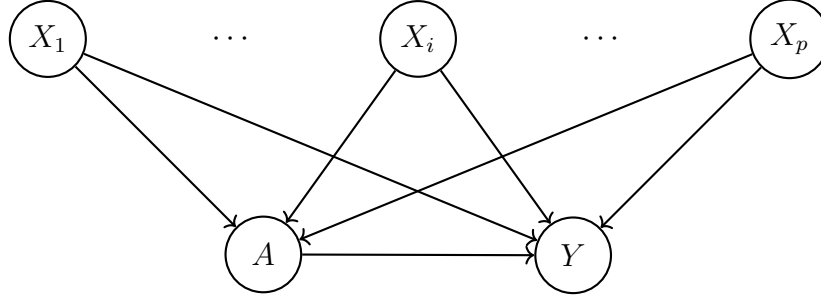


Figure 6.3: Causal graph of the treatment effect and the outcome.

Note that we could have used the simplified graph given in Figure 6.4. However, we want to emphasise the distinct effects of the different covariates in L . This is why we consider the first causal graph.

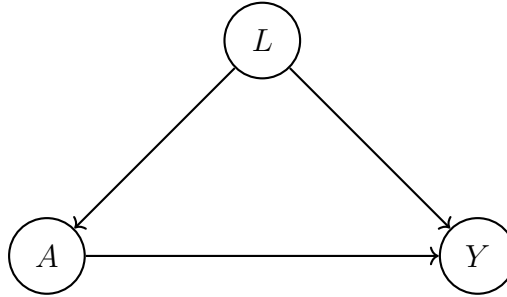


Figure 6.4: Simplified causal graph of the treatment effect and the outcome.

Here, we will only consider a basic case to illustrate the empirical robustness properties of $\hat{\mu}_{a,n}^{IPW}$. There is no need to consider complex contexts to see that this estimator is not robust. So, suppose we have the following setup:

- L is only composed of one continuous covariate X that takes values in \mathbb{R} .
- Y is a continuous outcome that takes values in \mathbb{R} .
- $\mathcal{V} = \{X, A, Y\}$.
- $\mathcal{C} = \{(X, A), (X, Y), (A, Y)\}$.
- $\mathcal{E} = \{\varepsilon_X, \varepsilon_A, \varepsilon_Y\}$ with

$$\varepsilon_X, \varepsilon_A, \varepsilon_Y \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

- \mathcal{S} is the following system of equations:

$$\mathcal{S} = \begin{cases} X = \varepsilon_X \\ A = \mathbb{1}_{\{X + \varepsilon_A > 0\}} \\ Y = A \exp(X) + \varepsilon_Y \end{cases}.$$

In this case, we get the (extended) causal graph given in Figure 6.5.

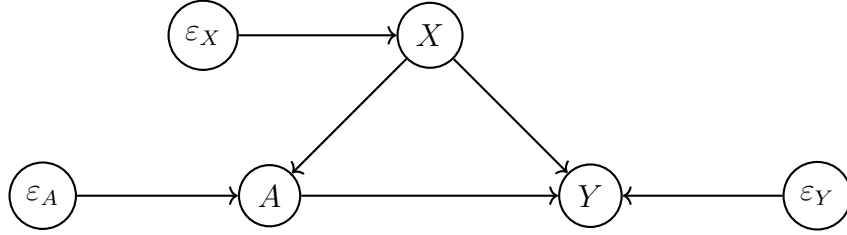


Figure 6.5: Extended causal graph for the case.

This structural causal model satisfies the following properties:

- We have the consistency property, that is, if $A = a$, $a \in \{0, 1\}$, then $Y^a = Y$. Indeed, the potential outcome Y^a is the outcome Y of the patient that would have been observed if A were equal to a , which is the case.
- We have conditional exchangeability, that is, $Y^a \perp\!\!\!\perp A|X$. When X is known, A depends only on ε_A which is independent of any other random variable, in particular of Y^a .
- We have the positivity property, that is, $\forall a \in \{0, 1\}, \forall x \in \mathbb{R}, \mathbb{P}(A = a|X = x) > 0$. Indeed,

$$\mathbb{P}(A = a|X = x) = \mathbb{P}(\mathbb{1}_{\{x + \varepsilon_A > 0\}} = a|X = x),$$

which is the probability that ε_A takes values greater (resp. less or equal) than $-x$ if $a = 1$ (resp. $a = 0$) given that $X = x$. Since $\varepsilon_A \sim \mathcal{N}(0, 1)$, this probability is always positive.

Therefore, by Proposition 3.2.1,

$$\mu_a^{IPW} = \mathbb{E} \left[\frac{\mathbb{1}_{\{A=a\}} Y}{\mathbb{P}(A = a|L)} \right] = \mathbb{E}[Y^a].$$

Recall that our estimator of μ_a^{IPW} , namely $\hat{\mu}_{a,n}^{IPW}$, is computed as follows:

$$\hat{\mu}_{a,n}^{IPW} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}_{\{A_i=a\}}}{\hat{\mathbb{P}}_n(A_i|L_i)},$$

where $\hat{\mathbb{P}}_n(A_i|L_i)$ is obtained using logistic regression.

Using the generating process described in Algorithm 1, we obtain the dataset² represented in Figure 6.6 for the structural causal model considered.

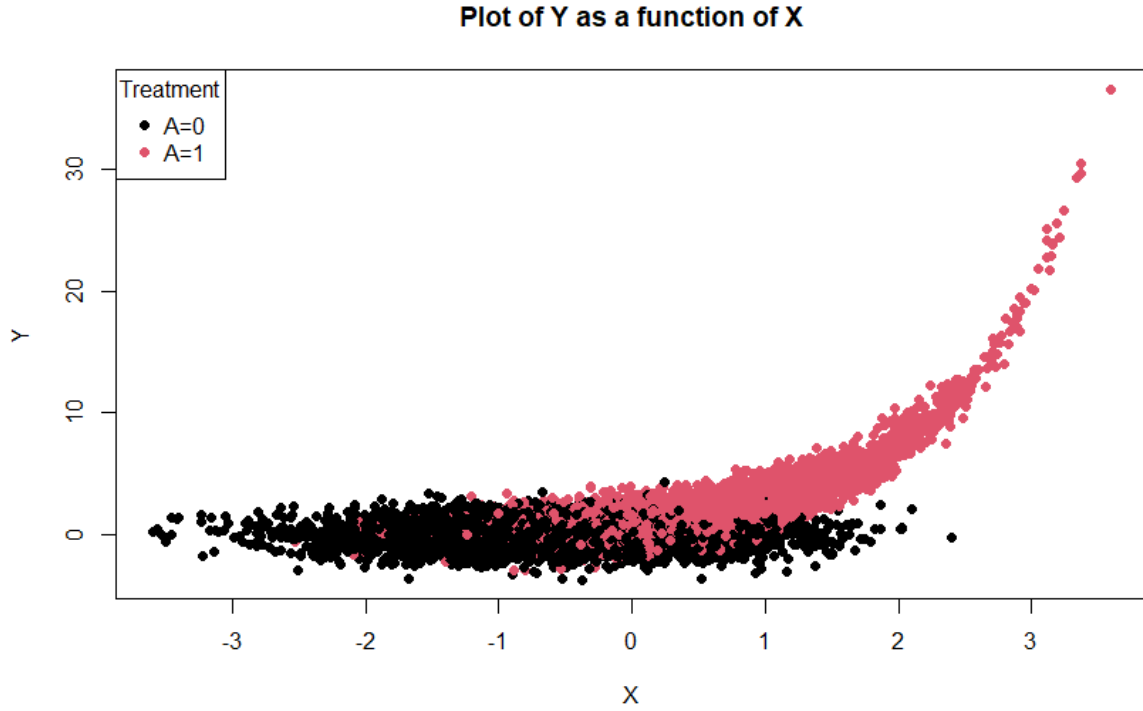


Figure 6.6: Plot of the causal dataset.

The estimates of μ_a^{IPW} , $a \in \{0, 1\}$, are given in Table 6.1, as well as the estimates of the conditional mean $\mathbb{E}[Y|A = a]$. An estimator $\hat{\mu}_{a,n}^{cond}$ of $\mathbb{E}[Y|A = a]$ can be obtained as

$$\hat{\mu}_a^{cond} = \frac{1}{\#\{i|A_i = a\}} \sum_{i:A_i=a} Y_i.$$

	$a = 1$	$a = 0$
$\hat{\mu}_{a,n}^{IPW}$	1.670	-0.019
$\hat{\mu}_{a,n}^{cond}$	2.514	0.003

Table 6.1: Table of estimates by treatment.

²We generate a dataset $\mathcal{D} = \{(l_i, a_i, y_i)\}_{i=1}^n$ containing $n = 10^4$ observations.

From Table 6.1, we can see that there is a non-negligible difference between both estimates when $a = 1$. This comes from the fact that the conditional mean estimator is just a mean over all observations such that $A = 1$. Therefore, it does not take into account that the value of A is influenced by X , resulting in a biased estimate of $\mathbb{E}[Y^a]$.

Let us now look at the empirical influence function of $\hat{\mu}_{a,n}^{IPW}$ for this dataset. From Proposition 5.3.1, the empirical influence function of $\hat{\mu}_{a,n}^{IPW}$, $a \in \{0, 1\}$, at $x = (\ell, \alpha, v) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}$ is given by³

$$\text{EIF}((\ell, \alpha, v); \hat{\mu}_{a,n}^{IPW}, \mathcal{D}) = \sum_{i=1}^n y_i \mathbb{1}_{\{a_i=a\}} \left(\frac{1}{\hat{\mathbb{P}}_{n+1}(a_i|l_i)} - \frac{(n+1)}{n\hat{\mathbb{P}}_n(a_i|l_i)} \right) + \frac{v \mathbb{1}_{\{\alpha=a\}}}{\hat{\mathbb{P}}_{n+1}(\alpha|\ell)}.$$

The empirical influence functions of $\hat{\mu}_{a,n}^{IPW}$, depending on the values of a and α , are displayed in Figure 6.7.

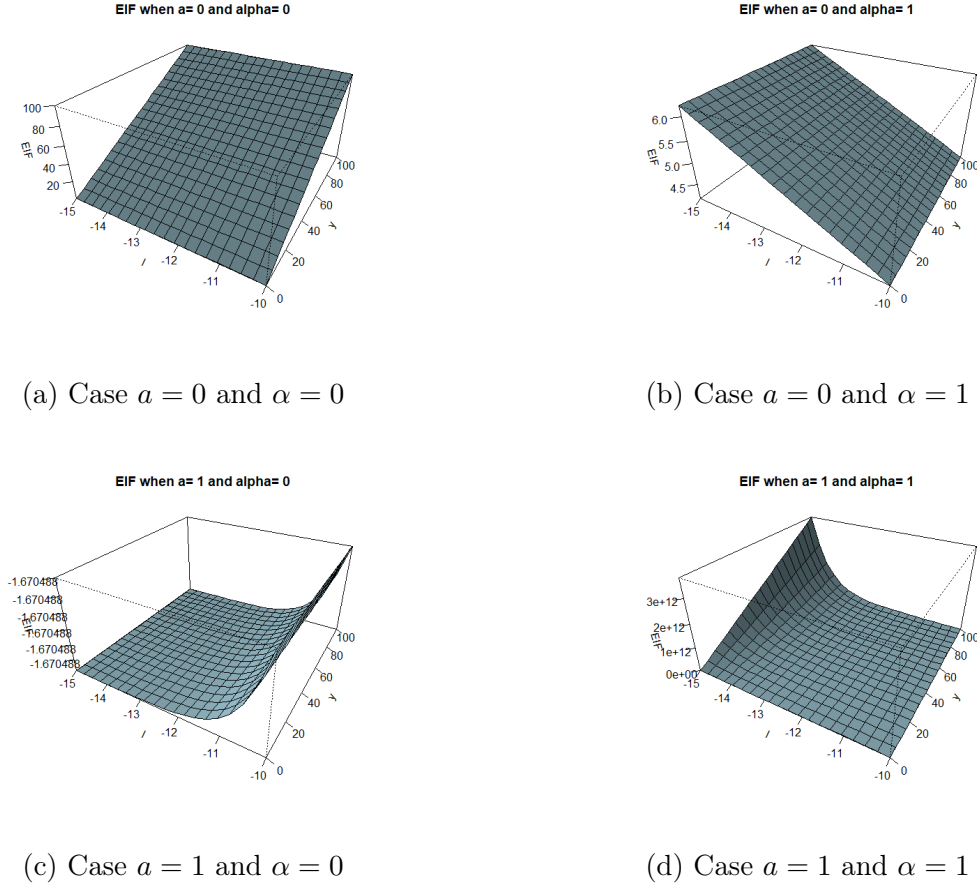


Figure 6.7: Empirical influence functions for the different values of a and α .

³In Appendix B.2, we use the definition of the empirical influence function rather than the formula obtained in Proposition 5.3.1 in order to facilitate its implementation.

We clearly see that in all cases the empirical influence function is not bounded. This was expected as, by Proposition 5.3.2, we have that the theoretical influence function is not bounded when $a = \alpha$, and in the empirical context, when $a \neq \alpha$, we still find that the influence function is not bounded because the logistic regression is influenced by α and ℓ even when $a \neq \alpha$. As a result, we can say without any doubt that $\hat{\mu}_{a,n}^{IPW}$ is not a robust estimator, as is also the case for the classical mean estimator.

We finish this section by looking at the approximate asymptotic distribution of $\hat{\mu}_{a,n}^{IPW}$. For that, we generate $N = 10^3$ datasets using the same structural causal model as before. The histograms of the resulting estimates, as well as their QQ-plots, are given in Figure 6.8 and Figure 6.9, respectively. The results of the Shapiro-Wilk tests for normality are reported in Table 6.2.

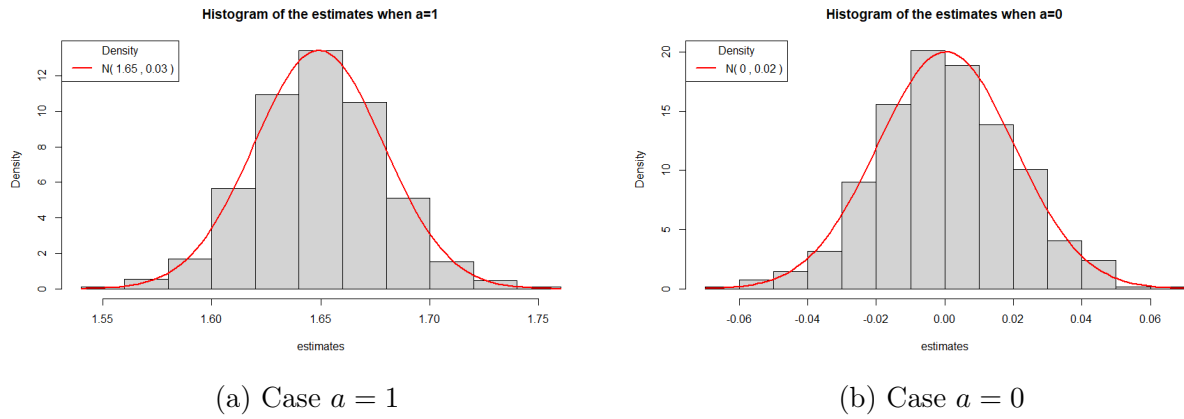


Figure 6.8: Histograms of the estimates of $\hat{\mu}_{a,n}^{IPW}$.

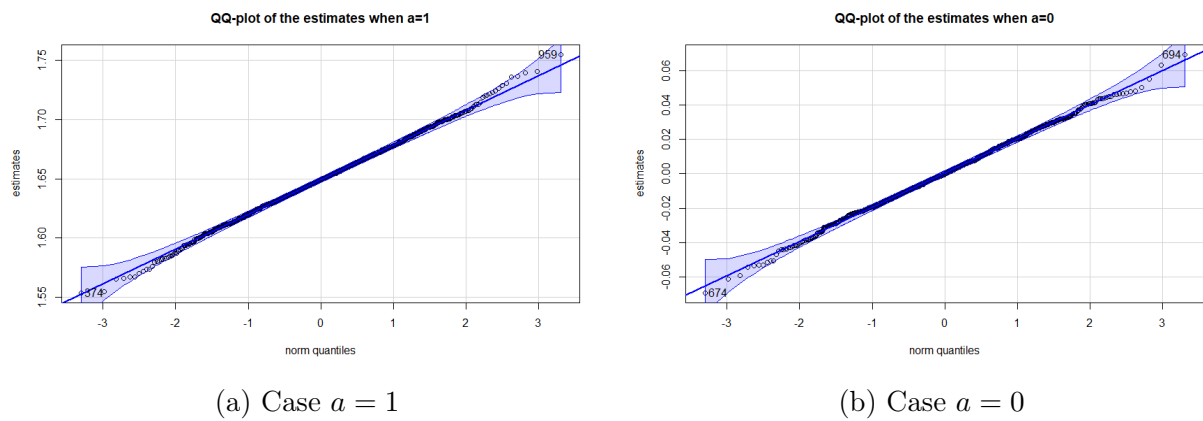


Figure 6.9: QQ-plots of the estimates of $\hat{\mu}_{a,n}^{IPW}$.

	W -Statistic	p -value	Decision
$a = 1$	0.99877	0.709	Do not reject the normality
$a = 0$	0.99884	0.758	Do not reject the normality

Table 6.2: Shapiro-Wilk tests for normality of the estimates of $\hat{\mu}_{a,n}^{IPW}$.

As we can see, it seems plausible that the estimates are normally distributed in both cases. This is what we could expect as we know, by Proposition 3.4.5, that when the propensity scores are known, the asymptotic distribution is a normal distribution. Although our estimator used a logistic regression, it seems to satisfy this convergence in distribution.

Chapter 7

Conclusion


Throughout this thesis, our objective was to provide an overview of the fields of causal inference and statistical robustness. Our aim was to formalize the application of classical robustness analysis within a causal framework, thereby establishing a solid mathematical foundation for future work.

In Chapter 2, we have defined the concept of causality, explained why it is an important topic, and gave the basic definitions in this field. We have seen that we could model situations where there exist causal links between the different variables with causal graphs. Furthermore, we have discussed the tools used in classical inference and mentioned that they could be used in the causal context.

In Chapter 3, we have formalized the process of inverse probability weighting. Using that, we have defined the inverse probability weighted mean (IPWM) and provided some of its properties in cases where the propensity score is known and when it is not.

The fundamentals of statistical robustness theory were developed in Chapter 4. We focused on a rigorous approach to define the breakdown point and the influence function both empirically and theoretically. For the theoretical influence function, we explored the theory behind the Gâteaux derivative that characterizes it. We also explored its possible applications such as obtaining an approximate asymptotic distribution of an estimator, or building estimators that have specific robustness properties.

In Chapter 5, we used the concepts previously developed in Chapter 4 to study the robustness properties of the IPWM estimator. We computed both the empirical and theoretical versions of its breakdown point and influence function. Using these results, it was clear that this estimator is not robust and is highly influenced by data perturbations. This corroborates the results known for the classical mean estimator, which was to be expected due to the similarity of the two estimators.

Finally, we have discussed how to generate causal datasets and provided an  implementation. We then used this algorithm to perform simulations and study the empirical robustness of the IPWM estimator. The results confirm the theoretical findings we obtained.

In conclusion, we can say without doubt that the IPWM estimator lacks robustness against outliers and we should not rely on such estimator to provide accurate estimates in the

presence of corrupted data. Future research could aim to enhance the robustness against outliers of the IPWM estimator or develop new estimators with stronger robustness properties.

Appendix A

Mathematical prerequisites

A.1 Functional analysis

In this section, the basics of functional analysis are given. This is mainly based on [17, 55].

A.1.1 Vector and metric spaces

Definition A.1.1 (Vector space). Let V be a non-empty set, let \mathbb{K} be a field ($\mathbb{K} = \mathbb{R}$ or \mathbb{C}), and consider the following two functions:

- $+: V^2 \rightarrow V : (x, y) \mapsto x + y$.
- $\cdot : \mathbb{K} \times V \rightarrow V : (k, y) \mapsto k \cdot y$.

$(V, +, \cdot)$ (or simply V if no confusion is possible) is said to be a \mathbb{K} -vector space if the following properties hold:

1. $\forall x, y, z \in V, x + y = y + x$ and $x + (y + z) = (x + y) + z$.
2. There exists a unique element $0_V \in V$ such that $\forall x \in V, 0_V + x = x + 0_V = x$.
3. $\forall x \in V$, there exists a unique element $y \in V$ such that $x + y = y + x = 0_V$.
4. $\forall k, l \in \mathbb{K}, \forall x, y \in V, k \cdot (x + y) = k \cdot x + k \cdot y, (k + l) \cdot x = k \cdot x + l \cdot x, (kl) \cdot x = k \cdot (l \cdot x)$ and $1 \cdot x = x$.

Definition A.1.2 (Linear mapping). Let V, W be two \mathbb{K} -vector spaces and let $T : V \rightarrow W$ be a mapping. T is said to be a linear mapping if $\forall x, y \in V, \forall k, l \in \mathbb{K}$,

$$T(k \cdot x + l \cdot y) = k \cdot T(x) + l \cdot T(y).$$

We use the notation $\mathcal{L}(V, W)$ to denote the set of all linear mappings from V to W .

Definition A.1.3 (Metric space). Let V be a set, and let $d : V^2 \rightarrow \mathbb{R}$ be a function. d is said to be a metric (or distance) if $\forall x, y, z \in V$, we have

- $d(x, y) \geq 0$.
- $d(x, y) = 0 \Leftrightarrow x = y$.
- $d(x, y) = d(y, x)$.
- $d(x, z) \leq d(x, y) + d(y, z)$.

In such a case, (V, d) (or simply V if no confusion is possible) is said to be a metric space.

Definition A.1.4 (Dense subset). Let (V, d) be a metric space, and let $S \subseteq V$. S is said to be a dense subset of V if $\forall v \in V$ and $\forall \varepsilon > 0$, $\exists s \in S$ such that $d(v, s) < \varepsilon$.

Definition A.1.5 (Separable space). Let (V, d) be a metric space. (V, d) (or simply V if no confusion is possible) is said to be separable if it possesses a countable dense subset.

Definition A.1.6 (Convergent sequence). Let (V, d) be a metric space, and let $(v_n)_{n \in \mathbb{N}_0}$ be a sequence of elements of V . $(v_n)_{n \in \mathbb{N}_0}$ is said to converge to an element $v \in V$ if $\forall \varepsilon > 0$, $\exists N \in \mathbb{N}_0$ such that

$$n \geq N \Rightarrow d(v_n, v) < \varepsilon.$$

In such case, we use the notation $v_n \longrightarrow v$ to denote that the sequence $(v_n)_{n \in \mathbb{N}_0}$ converges to v .

Definition A.1.7 (Cauchy sequence). Let (V, d) be a metric space, and let $(v_n)_{n \in \mathbb{N}_0}$ be a sequence of elements of V . $(v_n)_{n \in \mathbb{N}_0}$ is said to be a Cauchy sequence if $\forall \varepsilon > 0$, $\exists N \in \mathbb{N}_0$ such that

$$p, q \geq N \Rightarrow d(v_p, v_q) < \varepsilon.$$

Proposition A.1.1. *Let (V, d) be a metric space, and let $(v_n)_{n \in \mathbb{N}_0}$ be a sequence of elements of V that converges to $v \in V$. The following properties hold:*

- $\lim_{n \rightarrow +\infty} d(v_n, v) = 0$.
- The limit v is unique, that is, if $v' \in V$ is such that $v_n \longrightarrow v'$, then $v' = v$.
- $(v_n)_{n \in \mathbb{N}_0}$ is a Cauchy sequence.

Definition A.1.8 (Complete metric space). Let (V, d) be a metric space. (V, d) is said to be complete if any Cauchy sequence is convergent.

A.1.2 Hilbert spaces

Definition A.1.9 (Norm). Let V be a \mathbb{K} -vector space. A norm on V is a function $\|\cdot\| : V \rightarrow [0, +\infty[$ such that $\forall x, y \in V, \forall k \in \mathbb{K}$, we have

- $\|x\| = 0 \Leftrightarrow x = 0$.
- $\|k \cdot x\| = |k| \|x\|$.
- $\|x + y\| \leq \|x\| + \|y\|$.

$(V, \|\cdot\|)$ (or simply V if no confusion is possible) is said to be a normed vector space.

Remark A.1.1 (Notation). If $\|\cdot\|$ is a norm on V , then

$$d : V^2 \rightarrow [0, +\infty[: (x, y) \mapsto \|x - y\|$$

is a metric on V , called the associated metric.

Definition A.1.10 (Banach space). Let $(V, \|\cdot\|)$ be a normed vector space, and let d be the associated metric. $(V, \|\cdot\|)$ (or simply V if no confusion is possible) is said to be a Banach space if (V, d) is complete.

Definition A.1.11 (Inner/scalar product). Let V be an \mathbb{R} -vector space. An inner product (or scalar product) on V is a function $\langle \cdot, \cdot \rangle_V : V^2 \rightarrow \mathbb{R}$ such that $\forall x, y, z \in V, \forall k, l \in \mathbb{K}$, we have

- $\langle x, x \rangle_V \geq 0$.
- $\langle x, x \rangle_V = 0 \Leftrightarrow x = 0$.
- $\langle x, y \rangle_V = \langle y, x \rangle_V$.
- $\langle k \cdot x + l \cdot y, z \rangle_V = k \langle x, z \rangle_V + l \langle y, z \rangle_V$.

$(V, \langle \cdot, \cdot \rangle_V)$ (or simply V if no confusion is possible) is said to be an inner product space.

Remark A.1.2 (Notation). If $\langle \cdot, \cdot \rangle_V$ is a scalar product on V , then

$$\|\cdot\| : V \rightarrow [0, +\infty[: x \mapsto \sqrt{\langle x, x \rangle_V}$$

is a norm on V , called the induced norm.

Definition A.1.12 (Hilbert space). Let $(V, \langle \cdot, \cdot \rangle_V)$ be an inner product space, and let $\|\cdot\|_V$ be the induced norm. $(V, \langle \cdot, \cdot \rangle_V)$ is said to be a Hilbert space if $(V, \|\cdot\|_V)$ is a Banach space.

A.2 Measure theory and probability

In this section, the basics of measure theory are given. This is mainly based on [17, 25, 26, 34, 37, 42, 65].

A.2.1 σ -algebra

Definition A.2.1 (σ -algebra). Let Ω be a set and \mathcal{A} be a collection of subsets of Ω . \mathcal{A} is said to be a σ -algebra on Ω if the following properties hold:

- \mathcal{A} is non-empty.
- If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.
- If $(A_n)_{n \in \mathbb{N}_0}$ is a sequence of elements of \mathcal{A} , then $\bigcup_{n \in \mathbb{N}_0} A_n \in \mathcal{A}$.

Definition A.2.2 (Measurable space). A measurable space is a pair (Ω, \mathcal{A}) , where Ω is a set and \mathcal{A} is σ -algebra on Ω .

Proposition A.2.1 (Smallest σ -algebra). Let Ω be a set and \mathcal{C} be a collection of subsets of Ω . Then, there exists a smallest σ -algebra on Ω containing \mathcal{C} , noted $\sigma(\mathcal{C})$. Moreover, if we write $\mathcal{Z}_{\mathcal{C}} = \{\mathcal{A} \mid \mathcal{A} : \sigma\text{-algebra on } \Omega \text{ and } \mathcal{C} \subseteq \mathcal{A}\}$, then we have

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{A} \in \mathcal{Z}_{\mathcal{C}}} \mathcal{A}.$$

Definition A.2.3 (Borel σ -algebra). Let (Ω, \mathcal{N}) be a topological space. The Borel σ -algebra on Ω , noted $\mathcal{B}(\Omega, \mathcal{N})$ (or simply $\mathcal{B}(\Omega)$ if no confusion is possible), is defined as

$$\mathcal{B}(\Omega, \mathcal{N}) = \sigma(\{O \subseteq \Omega \mid O \text{ is an open set of } \Omega\}).$$

Proposition A.2.2. We have that

$$\mathcal{B}(\mathbb{R}) = \sigma(\{]-\infty, x] \mid x \in \mathbb{R}\}),$$

and, for $p \in \mathbb{N}_0$,

$$\mathcal{B}(\mathbb{R}^p) = \sigma \left(\left\{ \prod_{i=1}^p] - \infty, x_i] \mid x_1, \dots, x_p \in \mathbb{R} \right\} \right).$$

A.2.2 Measurability

Definition A.2.4 (Measurability of a function). Let (Ω, \mathcal{A}) and (Ω', \mathcal{A}') be two measurable spaces, and let $f : \Omega \rightarrow \Omega'$ be a function. f is said to be measurable with respect to \mathcal{A} and \mathcal{A}' (or simply measurable if no confusion is possible), if

$$f^{-1}(A') \in \mathcal{A}, \forall A' \in \mathcal{A}'.$$

Furthermore, the smallest σ -algebra on Ω that makes f measurable is denoted by $\sigma(f)$ and is obtained as

$$\sigma(f) = \{f^{-1}(A') \mid A' \in \mathcal{A}'\}.$$

Finally, if $\mathcal{C} = \{f_i : \Omega \rightarrow \Omega' \mid i \in I\}$, $I \subseteq \mathbb{N}_0$, is a collection of functions from Ω to Ω' , then we write $\sigma(\mathcal{C})$ (or $\sigma(f_i, i \in I)$) the smallest σ -algebra that makes f_i , $i \in I$, measurable. It is obtained as

$$\sigma(\mathcal{C}) = \sigma(\{f_i^{-1}(A') \mid A' \in \mathcal{A}', i \in I\}).$$

Definition A.2.5 (Strong measurability). Let (Ω, \mathcal{A}) and $(\mathcal{B}, \mathcal{B}(\mathcal{B}))$ be two measurable spaces, where \mathcal{B} is a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \mathcal{B}$ be a function. f is said to be strongly measurable if it is measurable and if $f(\Omega)$ is a separable space.

Remark A.2.1. If \mathcal{B} is separable, then any measurable function $f : \Omega \rightarrow \mathcal{B}$ is also strongly measurable.

Proposition A.2.3. Let (Ω, \mathcal{A}) be a measurable space, let $f : \Omega \rightarrow \mathbb{R}$, $g : \Omega \rightarrow \mathbb{R}$ be two measurable functions, and let $\alpha \in \mathbb{R}$. Then αf , $f + g$, $f - g$, and fg are measurable.

Proposition A.2.4. Let (Ω, \mathcal{A}) be a measurable space, let $(f_n : \Omega \rightarrow \mathbb{R})_{n \in \mathbb{N}_0}$ be a sequence of measurable functions. Then $\lim_{n \rightarrow +\infty} f_n$ is measurable.

Proposition A.2.5. Let (Ω, \mathcal{A}) , (Ω', \mathcal{A}') and $(\Omega'', \mathcal{A}'')$ be measurable spaces, and let $f : \Omega \rightarrow \Omega'$, $g : \Omega' \rightarrow \Omega''$ be measurable functions. We have that $g \circ f : \Omega \rightarrow \Omega''$ is measurable.

Proposition A.2.6. *Let (Ω, \mathcal{A}) and (Ω', \mathcal{A}') be measurable spaces, and let $f : \Omega \rightarrow \Omega'$ be a function. If \mathcal{C} is a collection of elements of \mathcal{A}' such that $\sigma(\mathcal{C}) = \mathcal{A}'$, then f is measurable if and only if*

$$f^{-1}(C) \in \mathcal{A}, \forall C \in \mathcal{C}.$$

Definition A.2.6 (Borel-measurability). Let (Ω, \mathcal{N}) be a topological space, and let $f : \Omega \rightarrow \mathbb{R}$ be a measurable function with respect to $\mathcal{B}(\Omega)$ and $\mathcal{B}(\mathbb{R})$, then, f is said to be Borel-measurable.

Proposition A.2.7. *Let (Ω, \mathcal{N}) be a topological space, and let $f : \Omega \rightarrow \mathbb{R}$ be a function. f is Borel-measurable if and only if $f^{-1}([-\infty, x]) \in \mathcal{B}(\Omega)$ for all $x \in \mathbb{R}$.*

Proposition A.2.8. *Let (Ω, \mathcal{N}) be a topological space, and let $f : \Omega \rightarrow \mathbb{R}$ be a continuous function, then f is Borel-measurable.*

A.2.3 Measure

Definition A.2.7 (Measure). Let (Ω, \mathcal{A}) be a measurable space. A function

$$\mu : \mathcal{A} \rightarrow [0, +\infty]$$

is said to be a measure on (Ω, \mathcal{A}) if $\mu(\emptyset) = 0$ and for any sequence $(A_n)_{n \in \mathbb{N}_0}$ of elements of \mathcal{A} such that $i \neq j \Rightarrow A_i \cap A_j = \emptyset$, we have

$$\mu \left(\bigcup_{n \in \mathbb{N}_0} A_n \right) = \sum_{n \in \mathbb{N}_0} \mu(A_n).$$

In such a case, $(\Omega, \mathcal{A}, \mu)$ is said to be a measure space.

Definition A.2.8 ((σ -)finite measure). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. μ is said to be finite if $\mu(\Omega) < +\infty$. Moreover, μ is said to be σ -finite if there exists a sequence $(A_n)_{n \in \mathbb{N}_0}$ of elements of \mathcal{A} such that $\Omega = \bigcup_{n \in \mathbb{N}_0} A_n$ and $\mu(A_n) < +\infty$ for all $n \in \mathbb{N}_0$.

Proposition A.2.9 (Basic properties of measures). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, then the following properties hold:*

- *If $A, A' \in \mathcal{A}$ and $A \subseteq A'$, then $\mu(A) \leq \mu(A')$. In the case where $\mu(A) < +\infty$, we also obtain that $\mu(A' \setminus A) = \mu(A') - \mu(A)$.*

- For any sequence $(A_n)_{n \in \mathbb{N}_0}$ of elements of \mathcal{A} , we have

$$\mu \left(\bigcup_{n \in \mathbb{N}_0} A_n \right) \leq \sum_{n \in \mathbb{N}_0} \mu(A_n).$$

- For any sequence $(A_n)_{n \in \mathbb{N}_0}$ of elements of \mathcal{A} such that $i \leq j \Rightarrow A_i \subseteq A_j$, we have

$$\mu \left(\bigcup_{n \in \mathbb{N}_0} A_n \right) = \lim_{n \rightarrow +\infty} \mu(A_n).$$

- For any sequence $(A_n)_{n \in \mathbb{N}_0}$ of elements of \mathcal{A} such that $i \leq j \Rightarrow A_i \supseteq A_j$ and such that $\mu(A_N) < +\infty$ for some $N \in \mathbb{N}_0$, we have

$$\mu \left(\bigcap_{n \in \mathbb{N}_0} A_n \right) = \lim_{n \rightarrow +\infty} \mu(A_n).$$

Definition A.2.9 (Image measure). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let (Ω', \mathcal{A}') be a measurable space, and let $f : \Omega \rightarrow \Omega'$ be a measurable function. The image measure of μ under f , noted μ_f , is defined as

$$\mu_f : \mathcal{A}' \rightarrow [0, +\infty] : A' \mapsto \mu(f^{-1}(A')).$$

Definition A.2.10 (π -system). Let \mathcal{C} be a set. \mathcal{C} is said to be a π -system (or closed under finite intersection), if $C_1, \dots, C_n \in \mathcal{C}$, $n \in \mathbb{N}_0$, implies that

$$\bigcap_{i=1}^n C_i \in \mathcal{C}.$$

Lemma A.2.1. Let (Ω, \mathcal{A}) be a measurable space and $\mathcal{C} \subseteq \mathcal{A}$ be a collection of elements of \mathcal{A} that is a π -system and such that $\sigma(\mathcal{C}) = \mathcal{A}$. If μ and ν are two σ -finite measures on \mathcal{A} such that $\forall C \in \mathcal{C}$, $\mu(C) = \nu(C)$, then $\mu = \nu$.

Definition A.2.11 (μ -almost everywhere property). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let P be a property (e.g., $P : f = g$, where f, g are two functions defined on Ω). P is said to be satisfied μ -almost everywhere, noted μ -a.e., if the set $A \subseteq \Omega$ where P does not hold is such that $\exists N \in \mathcal{A}$, with $A \subseteq N$ and $\mu(N) = 0$.

Remark A.2.2 (Notation). In the context of probability, we often use “almost surely”

instead of “almost everywhere”.

A.2.4 Probability

Definition A.2.12 (Probability measure). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. In the case where $\mu(\Omega) = 1$, μ is said to be a probability measure and $(\Omega, \mathcal{A}, \mu)$ is called a probability space.

Remark A.2.3 (Notation). Most of the time, the notation $(\Omega, \mathcal{F}, \mathbb{P})$ is used to denote a probability space, where Ω is a set, \mathcal{F} a σ -algebra on Ω and \mathbb{P} a probability measure on (Ω, \mathcal{F}) .

Definition A.2.13 (Random variable/vector). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a function. X is said to be a random vector (variable if $p = 1$), if it is measurable.

Definition A.2.14 (Distribution). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a random vector. The distribution of X is the probability measure \mathbb{P}_X on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ defined as

$$\mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \rightarrow [0, 1] : B \mapsto \mathbb{P}(X \in B).$$

In other words, \mathbb{P}_X is the image measure of \mathbb{P} under X . The notation $X \sim F$ is used to denote that F is the distribution of X .

Definition A.2.15 (Cumulative distribution function). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a random vector. The cumulative distribution function (cdf) of X is the function F_X defined as

$$F_X : \mathbb{R}^p \rightarrow [0, 1] : (x_1, \dots, x_p) \mapsto \mathbb{P}(X_1 \leq x_1, \dots, X_p \leq x_p).$$

Definition A.2.16 (Expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathbb{R}$ be a random vector. The expectation of X , noted $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega),$$

if it exists. In the case where X takes values in \mathbb{R}^p ($p \in \mathbb{N}_0$), we define the expectation of X as

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_p])^T.$$

Remark A.2.4 (Notation). If F is some distribution, then we use the notation $\mathbb{E}_F[X]$ to denote the expectation of variable X such that $X \sim F$, i.e.,

$$\mathbb{E}_F[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega),$$

where the distribution of X is F .

Definition A.2.17 (Characteristic function). Let $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mu)$, $p \in \mathbb{N}_0$, be a measure space. The characteristic function of μ , noted φ_μ , is defined as

$$\varphi_\mu : \mathbb{R}^p \rightarrow \mathbb{C} : t \mapsto \int_{\mathbb{R}^p} \exp(i \langle t, x \rangle_{\mathbb{R}^p}) d\mu(x).$$

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and $X = (X_1, \dots, X_p)$ is a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$, then its characteristic function, noted φ_X , is the characteristic function of \mathbb{P}_X . That is

$$\varphi_X : \mathbb{R}^p \rightarrow \mathbb{C} : t \mapsto \int_{\mathbb{R}^p} \exp(i \langle t, x \rangle_{\mathbb{R}^p}) d\mathbb{P}_X(x).$$

Definition A.2.18 (Convergence in distribution). Let $(X_n)_{n \in \mathbb{N}_0}$ be a sequence of random vectors (possibly all defined on different probability spaces) of size $p \in \mathbb{N}_0$, and let X be a random vector (possibly also defined on a different probability space) of size p . $(X_n)_{n \in \mathbb{N}_0}$ is said to convergence in distribution to X , noted $X_n \xrightarrow{\mathcal{L}} X$, if $\forall t \in \mathbb{R}^p$,

$$\lim_{n \rightarrow +\infty} \varphi_{X_n}(t) = \varphi_X(t).$$

Definition A.2.19 (Covariance matrix). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathbb{R}^p$ ($p \in \mathbb{N}_0$) be a random vector. The covariance matrix of X is defined as the $p \times p$ semi-definite positive matrix $\mathbb{V}[X]$ such that $\forall i, j \in \{1, \dots, p\}$,

$$(\mathbb{V}[X])_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])],$$

if this matrix exists.

Theorem A.2.1 (Multivariate central limit theorem). Let $(X_n)_{n \in \mathbb{N}_0}$ be a sequence of random vectors (possibly all defined on different probability spaces) of size $p \in \mathbb{N}_0$, independently and identically distributed according to a distribution F such that $\forall n \in \mathbb{N}_0$, $\mathbb{E}[X_n] = \mu \in \mathbb{R}^p$ and $\mathbb{V}[X_n] = \Sigma \in \mathbb{R}^{p \times p}$. If $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

where $\mathcal{N}(0, \Sigma)$ is the multivariate normal distribution with mean 0 and covariance matrix Σ .

Definition A.2.20 (Conditional probability). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ be two events such that $\mathbb{P}(B) \neq 0$. The conditional probability of A given B , written as $\mathbb{P}(A|B)$, is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Definition A.2.21 (Dirac distribution). Let (Ω, \mathcal{F}) be a measure space and let $\omega \in \Omega$. The Dirac distribution at ω is the probability measure Δ_ω defined as

$$\Delta_\omega : \mathcal{F} \rightarrow \{0, 1\} : F \mapsto \begin{cases} 1 & \text{if } \omega \in F \\ 0 & \text{if } \omega \notin F \end{cases}.$$

A.2.5 Independence

Definition A.2.22 (Independence of events). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{C} \subseteq \mathcal{F}$ be a finite collection of elements of \mathcal{F} . The elements of \mathcal{C} , called events, are independent if for all non-empty and finite subsets $\mathcal{D} = \{D_1, \dots, D_n\} \subseteq \mathcal{C}$, we have

$$\mathbb{P}\left(\bigcap_{i=1}^n D_i\right) = \prod_{i=1}^n \mathbb{P}(D_i).$$

Moreover, if $\mathcal{C} = \{C_i \mid i \in I\}$ where $I \subseteq \mathbb{N}_0$, then the elements of \mathcal{C} are independent if any finite sub-collection of elements of \mathcal{C} consists of independent events.

Remark A.2.5 (Notation). For $A, B \in \mathcal{F}$, we write $A \perp\!\!\!\perp B$ to denote that A and B are independent.

Definition A.2.23 (Independence of σ -algebras). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $I \subseteq \mathbb{N}_0$, and let $\mathfrak{C} = \{\Sigma_i \subseteq \mathcal{F} \mid i \in I\}$ be a collection of sub- σ -algebras of \mathcal{F} . The sub- σ -algebras in \mathfrak{C} are independent if any collection of events $\mathcal{C} = \{C_i \in \Sigma_i \mid i \in I\}$ consists of independent events.

Definition A.2.24 (Independence of random vectors). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathfrak{X} = \{X_i \mid i \in I\}$, $I \subseteq \mathbb{N}_0$, be a collection of random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$. The elements of \mathfrak{X} are independent if the sub- σ -algebras $\sigma(X_i)$, $i \in I$, are independent.

A.2.6 Conditional expectation and conditional independence

Definition A.2.25 (Conditional expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let Σ be a sub- σ -algebra of \mathcal{F} , and let X be random variable such that $\mathbb{E}[X]$ exists. The conditional expectation of X given Σ is the Σ -measurable almost surely unique variable Z such that $\forall S \in \Sigma$,

$$\mathbb{E}[Z\mathbb{1}_S] = \mathbb{E}[X\mathbb{1}_S].$$

The conditional expectation of X given Σ is written as $\mathbb{E}[X|\Sigma]$. Moreover, if Y_1, \dots, Y_n , $n \in \mathbb{N}_0$, are random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$, then we define the conditional expectation of X given Y_1, \dots, Y_n , noted $\mathbb{E}[X|Y_1, \dots, Y_n]$, as

$$\mathbb{E}[X|Y_1, \dots, Y_n] = \mathbb{E}[X|\sigma(Y_1, \dots, Y_n)].$$

The next proof is a modified version of the proof of [25, Lemma 7.2.4].

Proposition A.2.10. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X be random variable such that $\mathbb{E}[X]$ exists, and let $Y = (Y_1, \dots, Y_p)$, $p \in \mathbb{N}_0$, be random vector on $(\Omega, \mathcal{F}, \mathbb{P})$. The conditional expectation of X given Y is the $\sigma(Y)$ -measurable almost surely unique random variable $f(Y)$, where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is Borel-measurable, such that for all $S \in \sigma(Y)$,*

$$\mathbb{E}[f(Y)\mathbb{1}_S] = \mathbb{E}[X\mathbb{1}_S].$$

Proof. By Definition A.2.25, $\mathbb{E}[X|Y]$ is the $\sigma(Y)$ -measurable almost surely unique random variable Z such that for all $S \in \sigma(Y)$,

$$\mathbb{E}[Z\mathbb{1}_S] = \mathbb{E}[X\mathbb{1}_S].$$

In order to prove the claim, we just have to show that there exists a Borel-measurable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ such that $f(Y)$ is $\sigma(Y)$ -measurable and for all $S \in \sigma(Y)$,

$$\mathbb{E}[f(Y)\mathbb{1}_S] = \mathbb{E}[X\mathbb{1}_S].$$

Note that if $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is a Borel-measurable function, then, by Proposition A.2.5, $f(Y)$ is $\sigma(Y)$ -measurable.

First, let us suppose that $\mathbb{E}[X|Y]$ is a positive simple function (see Definition A.2.29) of the form

$$\mathbb{E}[X|Y] = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i},$$

where $\forall i \in \{1, \dots, N\}$, $\alpha_i > 0$ and $A_i \in \sigma(Y)$. If we let $B_i \in \mathcal{B}(\mathbb{R}^p)$ such that $A_i = Y^{-1}(B_i)$, and if we consider

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{B_i},$$

then f is Borel-measurable and $\mathbb{E}[X|Y] = f(Y)$.

Now, suppose that $\mathbb{E}[X|Y]$ takes positive values, by Proposition A.2.11, there exists a sequence of $\sigma(Y)$ -measurable positive simple function $(\psi_n)_{n \in \mathbb{N}_0}$ such that $\psi_n \rightarrow \mathbb{E}[X|Y]$ pointwise and $0 \leq \psi_1 \leq \dots \leq \mathbb{E}[X|Y]$. Using what precedes, $\forall n \in \mathbb{N}_0$, there exists a Borel-measurable function f_n such that $\psi_n = f_n(Y)$. If we consider the function

$$f : \mathbb{R}^p \rightarrow \mathbb{R} : (x_1, \dots, x_p) \mapsto \begin{cases} \lim_{n \rightarrow +\infty} f_n(x_1, \dots, x_p) & \text{if } (x_1, \dots, x_p) \in Y(\Omega) \\ 0 & \text{otherwise} \end{cases},$$

then f is Borel-measurable and $\mathbb{E}[X|Y] = f(Y)$.

In the general case, we let $(\mathbb{E}[X|Y])^+$ and $(\mathbb{E}[X|Y])^-$ be the positive and negative part of $\mathbb{E}[X|Y]$ respectively. By what precedes, there exists f^+, f^- two Borel-measurable functions such that $f^+(Y) = (\mathbb{E}[X|Y])^+$ and $f^-(Y) = (\mathbb{E}[X|Y])^-$. Therefore, since $\mathbb{E}[X|Y] = (\mathbb{E}[X|Y])^+ - (\mathbb{E}[X|Y])^-$, we have that $f = f^+ - f^-$ is a Borel-measurable function such that $f(Y) = \mathbb{E}[X|Y]$. \square

Definition A.2.26 (Conditional probability given a σ -algebra). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let Σ be a sub- σ -algebra of \mathcal{F} , and let $A \in \mathcal{F}$ be an event. The conditional probability of A given Σ , noted $\mathbb{P}(A|\Sigma)$, is defined as

$$\mathbb{P}(A|\Sigma) = \mathbb{E}[\mathbb{1}_A|\Sigma].$$

Furthermore, if $Y_1, \dots, Y_n, n \in \mathbb{N}_0$, are random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$, then we define the conditional probability of A given Y_1, \dots, Y_n , noted $\mathbb{P}(X|Y_1, \dots, Y_n)$, as

$$\mathbb{P}(X|Y_1, \dots, Y_n) = \mathbb{P}(X|\sigma(Y_1, \dots, Y_n)).$$

Definition A.2.27 (Conditional independence of σ -algebras). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\Sigma, \Sigma_1, \dots, \Sigma_n, n \in \mathbb{N}_0$ be sub- σ -algebras of \mathcal{F} . The sub- σ -algebras $\Sigma_1, \dots, \Sigma_n$ are conditionally independent given Σ if for all non-empty collections $\mathcal{D} = \{A_i | i \in I\}$ where $I \subseteq \{1, \dots, n\}, A_i \in \Sigma_i$, we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i \middle| \Sigma\right) \stackrel{\text{a.s.}}{=} \prod_{i \in I} \mathbb{P}(A_i | \Sigma).$$

Moreover, if $\mathcal{C} = \{\Sigma_i | i \in I\}, I \subseteq \mathbb{N}_0$, is a collection of sub- σ -algebras of \mathcal{F} , then the elements of \mathcal{C} are conditionally independent given Σ if any finite sub-collection of elements of \mathcal{C} consists of conditionally independent (given Σ) sub- σ -algebras.

Remark A.2.6 (Notation). If $\Sigma, \Sigma_1, \Sigma_2$ are sub- σ -algebras of \mathcal{F} , the notation $\Sigma_1 \perp\!\!\!\perp \Sigma_2 | \Sigma$ is used to denote that Σ_1 and Σ_2 are conditionally independent given Σ .

Definition A.2.28 (Conditional independence of random vectors). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X be random vector on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{C} = \{X_i \mid i \in I\}$, $I \subseteq \mathbb{N}_0$, be a collection of random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$. The elements of \mathcal{C} are conditionally independent given X if the sub- σ -algebras $\sigma(X_i)$, $i \in I$, are conditionally independent given $\sigma(X)$.

Remark A.2.7 (Notation). If X, X_1, X_2 are random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$, the notation $X_1 \perp\!\!\!\perp X_2 \mid X$ is used to denote that X_1 and X_2 are conditionally independent given X .

A.2.7 Integration

Definition A.2.29 (Simple function). Let f be a function from a set Ω to \mathcal{B} , a Banach space over \mathbb{R} or \mathbb{C} . f is said to be a simple function if $f(\Omega)$ is finite. In such case, if $f(\Omega) = \{\alpha_1, \dots, \alpha_N\}$, and if we let $A_i = f^{-1}(\{\alpha_i\})$, then we can write

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i}.$$

If (Ω, \mathcal{A}) is a measurable space, then we use the notation $\mathcal{S}(\Omega, \mathcal{A})$ to denote the set of all real-valued simple functions that are \mathcal{A} -measurable. Moreover, we define $\mathcal{S}^+(\Omega, \mathcal{A})$ as the set of all positive simple functions that are \mathcal{A} -measurable.

The next proof is a rewritten version of the proof of [42, Proposition 2.1.12].

Proposition A.2.11. *Let (Ω, \mathcal{A}) be a measurable space. If $f : \Omega \rightarrow [0, +\infty[$ is a measurable function, then there exists a sequence of measurable positive simple functions $(\psi_n)_{n \in \mathbb{N}_0}$ such that $\psi_n \rightarrow f$ pointwise and such that $0 \leq \psi_1 \leq \dots \leq f$.*

Proof. For each $n \in \mathbb{N}_0$ and for each $k \in \{1, \dots, n2^n\}$, we define the sets

$$A_{n,k} = f^{-1} \left(\left[\frac{k-1}{2^n}, \frac{k}{2^n} \right] \right) \text{ and } A_n = f^{-1}([n, +\infty[).$$

Since f is measurable, we have that $A_{n,k}, A_n \in \mathcal{A}$, and the functions $\mathbb{1}_{A_{n,k}}$ and $\mathbb{1}_{A_n}$ are measurable. Now, for each $n \in \mathbb{N}_0$, we define the function

$$\psi_n : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{A_{n,k}}(\omega) + n \mathbb{1}_{A_n}(\omega).$$

We have that the functions ψ_n , $n \in \mathbb{N}_0$, are measurable positive simple functions such that $0 \leq \psi_1 \leq \psi_2 \leq \dots \leq f$ and $\psi_n \rightarrow f$. \square

Definition A.2.30 (Integral with respect to a measure). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let $f \in \mathcal{S}^+(\Omega, \mathcal{A})$, such that

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i},$$

where $f(\Omega) = \{\alpha_1, \dots, \alpha_N\}$, and $A_i = f^{-1}(\{\alpha_i\})$. The integral of f with respect to μ , noted $\int f d\mu$, is defined as

$$\int f d\mu = \sum_{i=1}^N \alpha_i \mu(A_i).$$

If $f : \Omega \rightarrow \mathbb{R}$ is a positive and measurable function, then its integral with respect to μ is defined as

$$\int f d\mu = \sup \left\{ \int s d\mu \mid s \in \mathcal{S}^+(\Omega, \mathcal{A}), s \leq f \right\}.$$

Finally, if $f : \Omega \rightarrow \mathbb{R}$ is a measurable function with f^+ and f^- , its positive and negative part respectively, then its integral with respect to μ is defined as

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu,$$

if it exists. If $\int f^+ d\mu$ and $\int f^- d\mu$ are both finite, then f is said to be μ -integrable.

Remark A.2.8 (Notation). If f is μ -integrable, we often use the notation $\int_{\Omega} f(\omega) d\mu(\omega)$ or $\int_{\Omega} f(\omega) \mu(d\omega)$ to denote the integral of f with respect to μ .

Proposition A.2.12. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let $f : \Omega \rightarrow \mathbb{R}, g : \Omega \rightarrow \mathbb{R}$ be two μ -integrable functions. We have the following properties:

- If $\alpha \in \mathbb{R}$ then αf is μ -integrable and

$$\int \alpha f d\mu = \alpha \int f d\mu.$$

- $f + g$ is μ -integrable and

$$\int (f + g) d\mu = \int f d\mu + \int g d\mu.$$

- If $f \leq g$ then

$$\int f d\mu \leq \int g d\mu.$$

Proposition A.2.13. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let $f : \Omega \rightarrow \mathbb{R}$ be a function. We have that f is μ -integrable if and only if $|f|$ is μ -integrable. Moreover, if $|f|$ is μ -integrable and*

$$\int |f| d\mu = 0,$$

then $f = 0$ μ -almost-everywhere.

Theorem A.2.2 (Monotone convergence). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(f_n : \Omega \rightarrow [0, +\infty])_{n \in \mathbb{N}_0}$ a sequence of measurable functions such that $f_1 \leq f_2 \leq \dots$, and let $f : \Omega \rightarrow [0, +\infty]$ be a measurable function. If*

$$\lim_{n \rightarrow +\infty} f_n \stackrel{\mu\text{-a.e.}}{=} f,$$

then

$$\lim_{n \rightarrow +\infty} \int f_n d\mu = \int f d\mu.$$

In particular, if $(f_n : \Omega \rightarrow [0, +\infty])_{n \in \mathbb{N}_0}$ is a sequence of measurable functions, then

$$\sum_{n \in \mathbb{N}_0} \int f_n d\mu = \int \sum_{n \in \mathbb{N}_0} f_n d\mu.$$

Theorem A.2.3 (Dominated convergence). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(f_n : \Omega \rightarrow [0, +\infty])_{n \in \mathbb{N}_0}$ a sequence of measurable functions, let $g : \Omega \rightarrow [0, +\infty]$ be a μ -integrable function, and let $f : \Omega \rightarrow [0, +\infty]$ be a measurable function. If for almost every $\omega \in \Omega$*

$$\lim_{n \rightarrow +\infty} f_n(\omega) = f(\omega),$$

and, $\forall n \in \mathbb{N}_0$,

$$|f_n(\omega)| \leq g(\omega),$$

then $f, f_n, \forall n \in \mathbb{N}_0$, are μ -integrable functions and

$$\lim_{n \rightarrow +\infty} \int f_n d\mu = \int f d\mu.$$

Proposition A.2.14. *Let (Ω, \mathcal{A}) be a measurable space, let μ, ν be two measures on (Ω, \mathcal{A}) , let $\alpha \in [0, 1]$, let*

$$(1 - \alpha)\mu + \alpha\nu : \mathcal{A} \rightarrow [0, +\infty] : A \mapsto (1 - \alpha)\mu(A) + \alpha\nu(A),$$

and let $f : \Omega \rightarrow \mathbb{R}$ be a μ -integrable and ν -integrable function. We have that $(1 - \alpha)\mu + \alpha\nu$

is a measure on (Ω, \mathcal{A}) , f is $((1 - \alpha)\mu + \alpha\nu)$ -integrable, and we have

$$\int f d((1 - \alpha)\mu + \alpha\nu) = (1 - \alpha) \int f d\mu + \alpha \int f d\nu.$$

Proof. This is clear that $(1 - \alpha)\mu + \alpha\nu$ is a measure on (Ω, \mathcal{A}) . Also, since f is μ -integrable (and ν -integrable), we have that f is measurable.

Let us consider the case where $f \in \mathcal{S}^+(\Omega, \mathcal{A})$ of the form

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i},$$

where $N \in \mathbb{N}_0$, $\alpha_i > 0$ and $A_i = f^{-1}(\{\alpha_i\})$, $i \in \{1, \dots, N\}$. We have

$$\begin{aligned} \int f d((1 - \alpha)\mu + \alpha\nu) &= \sum_{i=1}^N \alpha_i ((1 - \alpha)\mu + \alpha\nu)(A_i) \\ &= (1 - \alpha) \sum_{i=1}^N \alpha_i \mu(A_i) + \alpha \sum_{i=1}^N \alpha_i \nu(A_i) \\ &= (1 - \alpha) \int f d\mu + \alpha \int f d\nu. \end{aligned}$$

As both $\int f d\mu$ and $\int f d\nu$ are finite, and $\alpha \in [0, 1]$, we have that $\int f d((1 - \alpha)\mu + \alpha\nu)$ exists and is finite. Therefore, f is $((1 - \alpha)\mu + \alpha\nu)$ -integrable.

Now, if f is positive function, then, by Proposition A.2.11, there exists a sequence of measurable positive simple functions $(\psi_n)_{n \in \mathbb{N}_0}$ such that $\psi_n \rightarrow f$ pointwise and such that $0 \leq \psi_1 \leq \dots \leq f$. By Theorem A.2.3, we have that $\forall n \in \mathbb{N}_0$, ψ_n is μ -integrable and ν -integrable (so $((1 - \alpha)\mu + \alpha\nu)$ -integrable by before), and

$$\int \psi_n d((1 - \alpha)\mu + \alpha\nu) = (1 - \alpha) \int \psi_n d\mu + \alpha \int \psi_n d\nu. \quad (\text{A.1})$$

By Theorem A.2.3, we have

$$\begin{aligned} (1 - \alpha) \int f d\mu + \alpha \int f d\nu &= (1 - \alpha) \lim_{n \rightarrow +\infty} \int \psi_n d\mu + \alpha \lim_{n \rightarrow +\infty} \int \psi_n d\nu \\ &= \lim_{n \rightarrow +\infty} \int \psi_n d((1 - \alpha)\mu + \alpha\nu) \text{ by Equation (A.1)} \\ &= \int f d((1 - \alpha)\mu + \alpha\nu) \text{ by Theorem A.2.2.} \end{aligned}$$

As a result, since $\int f d\mu$ and $\int f d\nu$ are finite, and $\alpha \in [0, 1]$, we have that $\int f d((1 - \alpha)\mu + \alpha\nu)$ exists and is finite. Therefore, f is $((1 - \alpha)\mu + \alpha\nu)$ -integrable.

In the general case, if f^+ and f^- are the positive and negative part of f respectively, then we have that $\int f^+ d\mu$, $\int f^+ d\nu$, $\int f^- d\mu$ and $\int f^- d\nu$ exists and are finite (since f is μ -integrable and ν -integrable). So, f^+ and f^- are both μ -integrable and ν -integrable. Therefore, using what precedes, we have that f^+ and f^- are both $((1 - \alpha)\mu + \alpha\nu)$ -integrable,

$$\int f^+ d((1 - \alpha)\mu + \alpha\nu) = (1 - \alpha) \int f^+ d\mu + \alpha \int f^+ d\nu,$$

and

$$\int f^- d((1 - \alpha)\mu + \alpha\nu) = (1 - \alpha) \int f^- d\mu + \alpha \int f^- d\nu.$$

By definition of the integral, we have

$$\begin{aligned} \int f d((1 - \alpha)\mu + \alpha\nu) &= \int f^+ d((1 - \alpha)\mu + \alpha\nu) - \int f^- d((1 - \alpha)\mu + \alpha\nu) \\ &= (1 - \alpha) \int f^+ d\mu + \alpha \int f^+ d\nu - \left((1 - \alpha) \int f^- d\mu + \alpha \int f^- d\nu \right) \\ &= (1 - \alpha) \left(\int f^+ d\mu - \int f^- d\mu \right) + \alpha \left(\int f^+ d\nu - \int f^- d\nu \right) \\ &= (1 - \alpha) \int f d\mu + \alpha \int f d\nu. \end{aligned}$$

□

Theorem A.2.4 (Transfer theorem). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let (Ω', \mathcal{A}') be a measurable space, and let $f : \Omega \rightarrow \Omega', g : \Omega' \rightarrow \mathbb{R}$ be two measurable functions. We have that g is μ_f -integrable if and only if $g \circ f$ is μ -integrable. In that case, we have*

$$\int_{\Omega'} g(\omega') d\mu_f(\omega') = \int_{\Omega} (g \circ f)(\omega) d\mu(\omega).$$

Theorem A.2.5 (Law of the unconscious statistician). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X = (X_1, \dots, X_p)$, $p \in \mathbb{N}_p$, be random vector on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $h : \mathbb{R}^p \rightarrow \mathbb{R}$ be a measurable function. Then $h(X_1, \dots, X_p)$ is \mathbb{P} -integrable (i.e., $\mathbb{E}[h(X_1, \dots, X_p)]$ exists) if and only if h is \mathbb{P}_X -integrable. In that case, we have*

$$\mathbb{E}[h(X_1, \dots, X_p)] = \int_{\mathbb{R}^p} h(x_1, \dots, x_p) d\mathbb{P}_X(x_1, \dots, x_p).$$

Remark A.2.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X]$ exists. By Theorem A.2.4, we have

$$\begin{aligned}\mathbb{E}[X] &= \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \\ &= \int_{\mathbb{R}} x d\mathbb{P}_X(x).\end{aligned}$$

Therefore, the expectation of X (if it exists) depends only on the distribution of X and is the integral of the identity function $\text{id} : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto x$ with respect to \mathbb{P}_X .

Definition A.2.31 (Quotient set). Let E be a set, and let $=_E$ be an equivalence relation on E . The quotient set of E by $=_E$ is the set $E/_=_E$ defined as

$$E/_=_E = \{[e]_{=_E} : e \in E\},$$

where $[e]_{=_E} = \{e' \in E : e' =_E e\}$.

Definition A.2.32 (L^p -space). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $p \in \mathbb{N}_0$. $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ denotes the space of all measurable functions $f : \Omega \rightarrow \mathbb{R}$ such that $|f|^p$ is μ -integrable. We consider the equivalence relation $f \sim_{\mu} g$, $f, g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ defined as $f \sim_{\mu} g \Leftrightarrow f \stackrel{\mu\text{-a.e.}}{=} g$, and we define $L^p(\Omega, \mathcal{A}, \mu)$ as the space obtained as the quotient of $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, by the equivalence relation $f \sim_{\mu} g$.

A.2.8 Bochner's integration

Proposition A.2.15. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \mathcal{B}$ be a strongly measurable function. There exists a sequence of strongly measurable simple functions $(\psi_n : \Omega \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f(\omega),$$

and, $\forall n \in \mathbb{N}_0$,

$$\|\psi_n(\omega)\|_{\mathcal{B}} \leq \|f(\omega)\|_{\mathcal{B}}.$$

Definition A.2.33 (Bochner's integrability). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \mathcal{B}$ be a measurable function. f is said to be Bochner μ -integrable if f is strongly measurable and if the function

$$g : \Omega \rightarrow [0, +\infty[: \omega \mapsto \|f(\omega)\|_{\mathcal{B}}$$

is μ -integrable. In such case, we define the Bochner's integral of f with respect to μ ,

noted $\int f d\mu$, as follows:

- If f is a simple function that takes the form

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i},$$

where $N \in \mathbb{N}_0$, $\alpha_i \in \mathcal{B}$ and $A_i = f^{-1}(\{\alpha_i\}) \in \mathcal{A}$, $i \in \{1, \dots, N\}$, then

$$\int f d\mu = \sum_{i=1}^N \alpha_i \mu(A_i).$$

- In the general case, let $(\psi_n : \Omega \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ be a sequence of simple Bochner μ -integrable functions such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f(\omega),$$

and for which the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} \|\psi_n(\omega)\|_{\mathcal{B}}$$

is μ -integrable, then

$$\int f d\mu = \lim_{n \rightarrow +\infty} \int \psi_n d\mu.$$

Remark A.2.10 (Notation). If f is Bochner μ -integrable, we often use the notation $\int_{\Omega} f(\omega) d\mu(\omega)$ or $\int_{\Omega} f(\omega) \mu(d\omega)$ to denote the Bochner's integral of f with respect to μ .

Remark A.2.11. In the case where $(\mathcal{B}, \|\cdot\|_{\mathcal{B}}) = (\mathbb{R}, |\cdot|)$, the Bochner's integral of f with respect to μ corresponds to the classical integral of f with respect to μ .

Proposition A.2.16. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \mathcal{B}$ be a function. We have the following properties:

- If f is Bochner μ -integrable, then we have the inequality

$$\left\| \int f d\mu \right\|_{\mathcal{B}} \leq \int \|f\|_{\mathcal{B}} d\mu.$$

- If f is Bochner μ -integrable, $g : \Omega \rightarrow \mathcal{B}$ is a Bochner μ -integrable function and

if $\alpha, \beta \in \mathbb{R}$ (or \mathbb{C}), then $\alpha f + \beta g$ is Bochner μ -integrable and

$$\int \alpha f + \beta g d\mu = \alpha \int f d\mu + \beta \int g d\mu.$$

- Suppose that $g : \Omega \rightarrow [0, +\infty[$ is a μ -integrable function and let $(\psi_n : \Omega \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ be a sequence of strongly measurable functions such that for almost all $\omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f(\omega),$$

and, $\forall n \in \mathbb{N}_0$,

$$\|\psi_n(\omega)\|_{\mathcal{B}} \leq g(\omega).$$

Then, $f, \psi_n, n \in \mathbb{N}_0$, are Bochner μ -integrable functions and we have

$$\int f d\mu = \lim_{n \rightarrow +\infty} \int \psi_n d\mu.$$

- If f is Bochner μ -integrable and $\varphi : \mathcal{B} \rightarrow \mathcal{B}'$ is a bounded linear mapping where \mathcal{B}' is a Banach space over \mathbb{R} or \mathbb{C} , then $\varphi \circ f$ is Bochner μ -integrable and we have

$$\int \varphi \circ f d\mu = \varphi \left(\int f d\mu \right).$$

Proposition A.2.17. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $(\mathbb{R}^p, \|\cdot\|_\infty)$, $p \in \mathbb{N}_0$, be a Banach space over \mathbb{R} , where

$$\|\cdot\|_\infty : \mathbb{R}^p \rightarrow [0, +\infty[: (x_1, \dots, x_p) \mapsto \sup_{i \in \{1, \dots, p\}} |x_i|,$$

and let $f : \Omega \rightarrow \mathbb{R}^p$ be a Bochner μ -integrable function. We have that

$$\int f d\mu = \left(\int f_1 d\mu, \dots, \int f_p d\mu \right)^T,$$

where $f_i, i \in \{1, \dots, p\}$, is the i -th component of f , that is,

$$f_i : \Omega \rightarrow \mathbb{R} : \omega \mapsto (f(\omega))_i.$$

Proof. We have that \mathbb{R}^p is separable space, so by Remark A.2.1, any measurable function $f : \Omega \rightarrow \mathbb{R}^p$ is strongly measurable.

If f is Bochner μ -integrable, then f is strongly measurable, and so, is measurable. Also, we have that the projection functions

$$\pi_i : \mathbb{R}^p \rightarrow \mathbb{R} : (x_1, \dots, x_p) \mapsto x_i, \quad i \in \{1, \dots, p\},$$

are continuous, and so they are measurable by Proposition A.2.8. As a result, f_i , $i \in \{1, \dots, p\}$, are measurable due to Proposition A.2.5 and to the fact that

$$f_i = \pi_i \circ f,$$

which is a composition of measurable functions. Therefore, we have that f_i , $i \in \{1, \dots, p\}$, are strongly measurable.

Moreover, since f is Bochner μ -integrable, the function

$$g : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{i \in \{1, \dots, p\}} |f_i(\omega)|$$

is μ -integrable. Note that $\forall \omega \in \Omega$, $\forall i \in \{1, \dots, p\}$, we have

$$|f_i(\omega)| \leq g(\omega).$$

Let us fix $i \in \{1, \dots, p\}$ and consider the sequence of function $(\psi_n)_{n \in \mathbb{N}_0}$ where $\forall n \in \mathbb{N}_0$, $\psi_n = f_i$. This is a sequence of strongly measurable functions such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f_i(\omega),$$

and, $\forall n \in \mathbb{N}_0$,

$$\|\psi_n(\omega)\|_\infty \leq g(\omega).$$

Therefore, by Proposition A.2.16, we have that f_i is Bochner μ -integrable, and so, is μ -integrable. We just showed that $\forall i \in \{1, \dots, p\}$, f_i is μ -integrable and that

$$\int f_i d\mu$$

exists.

Now, consider the case where f is a simple function of the form

$$f = \sum_{k=1}^K \alpha_k \mathbb{1}_{A_k},$$

where $K \in \mathbb{N}_0$, $\alpha_k \in \mathbb{R}^p$ and $A_k = f^{-1}(\{\alpha_k\})$, $k \in \{1, \dots, K\}$. We have that

$$\int f d\mu = \sum_{k=1}^K \alpha_k \mu(A_k)$$

$$\begin{aligned}
&= \left(\sum_{k=1}^K (\alpha_k)_1 \mu(A_k), \dots, \sum_{k=1}^K (\alpha_k)_p \mu(A_k) \right)^T \\
&= \left(\int f_1 d\mu, \dots, \int f_p d\mu \right)^T.
\end{aligned}$$

In the general case, consider the sequence of simple Bochner μ -integrable functions $(\zeta_n : \Omega \rightarrow \mathbb{R}^p)_{n \in \mathbb{N}_0}$ such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \zeta_n(\omega) = f(\omega),$$

and for which the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} \|\zeta_n(\omega)\|_\infty$$

is μ -integrable. We have that

$$\lim_{n \rightarrow +\infty} \int \zeta_n d\mu = \int f d\mu,$$

and $\forall n \in \mathbb{N}_0$, we have

$$\int \zeta_n d\mu = \left(\int (\zeta_n)_1 d\mu, \dots, \int (\zeta_n)_p d\mu \right)^T.$$

As a result,

$$\int f d\mu = \lim_{n \rightarrow +\infty} \left(\int (\zeta_n)_1 d\mu, \dots, \int (\zeta_n)_p d\mu \right)^T.$$

Note that $\forall i \in \{1, \dots, p\}$, $((\zeta_n)_i)_{n \in \mathbb{N}_0}$ is a sequence of simple Bochner μ -integrable functions such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} (\zeta_n)_i(\omega) = f_i(\omega),$$

and for which the function

$$h_i : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} |(\zeta_n(\omega))_i|$$

is μ -integrable. Therefore,

$$\int f_i d\mu = \lim_{n \rightarrow +\infty} \int (\zeta_n)_i d\mu.$$

By unicity of the limits, we can conclude that

$$\int f d\mu = \left(\int f_1 d\mu, \dots, \int f_p d\mu \right)^T.$$

□

Proposition A.2.18. *Let (Ω, \mathcal{A}) be a measurable space, let μ, ν be two measures on (Ω, \mathcal{A}) , let $\alpha \in [0, 1]$, let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \mathcal{B}$ be a Bochner μ -integrable and Bochner ν -integrable function. We have that f is Bochner $((1 - \alpha)\mu + \alpha\nu)$ -integrable and*

$$\int f d((1 - \alpha)\mu + \alpha\nu) = (1 - \alpha) \int f d\mu + \alpha \int f d\nu.$$

Proof. First, since f is Bochner μ -integrable and Bochner ν -integrable, f is strongly measurable. Moreover, the function

$$g : \Omega \rightarrow [0, +\infty[: \omega \mapsto \|f(\omega)\|_{\mathcal{B}}$$

is μ -integrable and ν -integrable. Therefore, by Proposition A.2.14, g is $((1 - \alpha)\mu + \alpha\nu)$ -integrable. As a result, f is Bochner $((1 - \alpha)\mu + \alpha\nu)$ -integrable.

Now, consider the case where f is a simple function of the form

$$f = \sum_{i=1}^N \alpha_i \mathbb{1}_{A_i},$$

where $N \in \mathbb{N}_0$, $\alpha_i \in \mathbb{R}^p$ and $A_i = f^{-1}(\{\alpha_i\})$, $i \in \{1, \dots, N\}$. We have that

$$\begin{aligned} \int f d((1 - \alpha)\mu + \alpha\nu) &= \sum_{k=1}^K \alpha_k ((1 - \alpha)\mu + \alpha\nu)(A_k) \\ &= (1 - \alpha) \sum_{k=1}^K \alpha_k \mu(A_k) + \alpha \sum_{k=1}^K \alpha_k \nu(A_k) \\ &= (1 - \alpha) \int f d\mu + \alpha \int f d\nu. \end{aligned}$$

In the general case, let $(\psi_n : \Omega \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ be a sequence of simple Bochner μ -integrable and Bochner ν -integrable functions such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f(\omega),$$

and for which the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} \|\psi_n(\omega)\|_{\mathcal{B}}$$

is μ -integrable and ν -integrable. Using what precedes, we have that $(\psi_n : \Omega \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ is a sequence of simple Bochner $((1 - \alpha)\mu + \alpha\nu)$ -integrable functions such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega) = f(\omega),$$

and for which the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} \|\psi_n(\omega)\|_{\mathcal{B}}$$

is $((1 - \alpha)\mu + \alpha\nu)$ -integrable. Therefore, by definition of the Bochner's integral, and using what precedes, we have

$$\begin{aligned} \int f d((1 - \alpha)\mu + \alpha\nu) &= \lim_{n \rightarrow +\infty} \int \psi_n d((1 - \alpha)\mu + \alpha\nu) \\ &= \lim_{n \rightarrow +\infty} (1 - \alpha) \int \psi_n d\mu + \alpha \int \psi_n d\nu \\ &= (1 - \alpha) \lim_{n \rightarrow +\infty} \int \psi_n d\mu + \alpha \lim_{n \rightarrow +\infty} \int \psi_n d\nu \\ &= (1 - \alpha) \int f d\mu + \alpha \int f d\nu. \end{aligned}$$

□

Theorem A.2.6 (Transfert theorem for the Bochner's integral). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let (Ω', \mathcal{A}') be a measurable space, let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a Banach space over \mathbb{R} or \mathbb{C} , and let $f : \Omega \rightarrow \Omega', g : \Omega' \rightarrow \mathcal{B}$ be two strongly measurable functions. We have that g is Bochner μ_f -integrable if and only if $g \circ f$ is Bochner μ -integrable. In that case, we have*

$$\int_{\Omega'} g(\omega') d\mu_f(\omega') = \int_{\Omega} (g \circ f)(\omega) d\mu(\omega).$$

Proof. By Proposition A.2.5, since both f and g are measurable, $g \circ f$ is measurable. Also, $g \circ f$ is strongly measurable, as $(g \circ f)(\Omega) = g(f(\Omega)) \subseteq g(\Omega')$, which is separable, and so, $(g \circ f)(\Omega)$ is also separable.

Note that g is Bochner μ_f -integrable if and only if the function

$$h_g : \Omega' \rightarrow [0, +\infty[: \omega' \mapsto \|g(\omega')\|_{\mathcal{B}}$$

is μ_f -integrable, and $g \circ f$ is Bochner μ -integrable if and only if the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \|g(f(\omega))\|_{\mathcal{B}}$$

is μ -integrable. We have $h = h_g \circ f$ and by Theorem A.2.4, we have that h_g is μ_f -integrable if and only if h is μ -integrable. In other words, g is Bochner μ_f -integrable if and only if $g \circ f$ is Bochner μ -integrable.

Now, let us suppose that g is Bochner μ_f -integrable, and so, $g \circ f$ is Bochner μ -integrable.

First, consider the case where g is a simple function of the form

$$g = \sum_{i=1}^N \beta_i \mathbb{1}_{A'_i},$$

where $N \in \mathbb{N}_0$, $\beta_i \in \mathcal{B}$ and $A'_i = g^{-1}(\{\beta_i\})$, $i \in \{1, \dots, N\}$. We have that $\forall \omega \in \Omega$,

$$\begin{aligned} (g \circ f)(\omega) &= \sum_{i=1}^N \beta_i \mathbb{1}_{A'_i}(f(\omega)) \\ &= \sum_{i=1}^N \beta_i \mathbb{1}_{f^{-1}(A'_i)}(\omega), \end{aligned}$$

so

$$g \circ f = \sum_{i=1}^N \beta_i \mathbb{1}_{f^{-1}(A'_i)}.$$

We have

$$\begin{aligned} \int_{\Omega'} g(\omega') d\mu_f(\omega') &= \sum_{i=1}^N \beta_i \mu_f(A'_i) \\ &= \sum_{i=1}^N \beta_i \mu(f^{-1}(A'_i)) \\ &= \int_{\Omega} (g \circ f)(\omega) d\mu(\omega). \end{aligned}$$

In the general case, let $(\psi_n : \Omega' \rightarrow \mathcal{B})_{n \in \mathbb{N}_0}$ be a sequence of simple Bochner μ_f -integrable functions such that $\forall \omega' \in \Omega'$,

$$\lim_{n \rightarrow +\infty} \psi_n(\omega') = g(\omega'),$$

and for which the function

$$h'_g : \Omega' \rightarrow [0, +\infty[: \omega' \mapsto \sup_{n \in \mathbb{N}_0} \|\psi_n(\omega')\|_{\mathcal{B}}$$

is μ_f -integrable. We have that, $(\psi_n \circ f)_{n \in \mathbb{N}_0}$ is a sequence of simple Bochner μ -integrable functions. Moreover, $(\psi_n \circ f)_{n \in \mathbb{N}_0}$ is such that $\forall \omega \in \Omega$,

$$\lim_{n \rightarrow +\infty} (\psi_n \circ f)(\omega) = (g \circ f)(\omega),$$

and the function

$$h' : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{n \in \mathbb{N}_0} \|(\psi_n \circ f)(\omega)\|_{\mathcal{B}}$$

is μ -integrable (by Theorem A.2.4 since $h' = h'_g \circ f$). Using what precedes, we have

$$\begin{aligned} \int_{\Omega'} g(\omega') d\mu_f(\omega') &= \lim_{n \rightarrow +\infty} \int_{\Omega'} \psi_n(\omega') d\mu_f(\omega') \\ &= \lim_{n \rightarrow +\infty} \int_{\Omega} (\psi_n \circ f)(\omega) d\mu(\omega) \\ &= \int_{\Omega} (g \circ f)(\omega) d\mu(\omega). \end{aligned}$$

□

Remark A.2.12. In the context of probability, if \mathbb{P}_X is the distribution of random vector X taking values in \mathbb{R}^p , $p \in \mathbb{N}_0$, then the expectation (if it exists) of X is the Bochner's integral of the function $\text{id} : \mathbb{R}^p \rightarrow \mathbb{R}^p : (x_1, \dots, x_p) \mapsto (x_1, \dots, x_p)$ with respect to \mathbb{P}_X . That is

$$\mathbb{E}[X] = \int_{\mathbb{R}^p} x d\mathbb{P}_X(x).$$

Indeed, if $X = (X_1, \dots, X_p)$ is random vector taking values in \mathbb{R}^p such that $\mathbb{E}[X]$ exists, then, by definition of $\mathbb{E}[X]$, $\mathbb{E}[X_i]$ exists $\forall i \in \{1, \dots, p\}$. Also, X is strongly measurable since \mathbb{R} is separable. Moreover, X is Bochner \mathbb{P} -integrable since the function

$$h : \Omega \rightarrow [0, +\infty[: \omega \mapsto \sup_{i \in \{1, \dots, p\}} |X_i(\omega)|$$

is such that

$$h = \sum_{i=1}^p \mathbb{1}_{A_i} |X_i|,$$


where $A_i = \left\{ \omega \in \Omega \left| \sup_{i \in \{1, \dots, p\}} |X_i(\omega)| = |X_i(\omega)| \right. \right\}$. So h is a sum of products of \mathbb{P} -integrable functions, and so, h is \mathbb{P} -integrable. Therefore, by Theorem A.2.6, $\text{id}_p : \mathbb{R}^p \rightarrow \mathbb{R}^p$ is Bochner \mathbb{P}_X -integrable, and we have

$$\begin{aligned} \int_{\mathbb{R}^p} x d\mathbb{P}_X(x) &= \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \\ &= \left(\int_{\Omega} X_1(\omega) d\mathbb{P}(\omega), \dots, \int_{\Omega} X_p(\omega) d\mathbb{P}(\omega) \right)^T \text{ by Proposition A.2.17} \\ &= \left(\int_{\mathbb{R}} \text{id} d\mathbb{P}_{X_1}, \dots, \int_{\mathbb{R}} \text{id} d\mathbb{P}_{X_p} \right)^T \text{ by Theorem A.2.4} \\ &= (\mathbb{E}[X_1], \dots, \mathbb{E}[X_p])^T \\ &= \mathbb{E}[X]. \end{aligned}$$

Appendix B

Scripts

B.1 Generation of causal data

Here is the complete  script of the generation of causal data discussed in Section 6.2.

```
1 ##### Generation of causal data #####
2
3 #---- Working Directory ----
4
5 (WD <- getwd())
6 if (!is.null(WD)) setwd(WD)
7
8 #---- Libraries ----
9
10 library(igraph)
11 library(ggraph)
12 library(hash)
13
14 #---- Causal Graphs ----
15
16 get_parents <- function(edges){
17   # %------%
18   # Description
19   # %------%
20   # Given a two-column edge list (parent -> child),
21   # build a hash table where each key is a child
22   # node and each value is the vector of its
23   # parent nodes.
24   # %------%
25   # Parameters
26   # %------%
27   # edges : matrix or data.frame with 2 columns
28   #   column 1 = parent node
29   #   column 2 = child node
30   # %------%
31   N_edges <- nrow(edges)
32   parents <- hash::hash()
```



```

33   for (i in 1:N_edges){
34     child <- edges[i,2]
35     parent <- edges[i,1]
36     if (!has.key(child, parents)){
37       parents[[child]] <- c(parent)
38     }else{
39       parents[[child]] <- c(parents[[child]], parent)
40     }
41   }
42   for (key in keys(parents)){
43     #We delete duplicates
44     parents[[key]] <- unique(parents[[key]])
45   }
46   return(parents)
47 }
48
49 create_DAG <- function(edges){
50   # %------%
51   # Description
52   # %------%
53   # Given a two-column edge list (parent -> child),
54   # build the corresponding DAG
55   # %------%
56   # Parameters
57   # %------%
58   # edges : matrix or data.frame with 2 columns
59   #   column 1 = parent node
60   #   column 2 = child node
61   # %------%
62   graph <- graph_from_edgelist(edges, directed = TRUE)
63 }
64
65 #---- Generation of causal dataset ----
66
67 generate_causal_dataset <- function(N, edges, epsilons, functions){
68   # %------%
69   # Description
70   # %------%
71   # Given a two-column edge list (parent -> child),
72   # a list of generating functions, and a list of
73   # functions, build a dataset satisfying the
74   # corresponding structural causal model
75   # %------%
76   # Parameters
77   # %------%
78   # N : the number of observations to be generated
79   # edges : matrix or data.frame with 2 columns
80   #   column 1 = parent node
81   #   column 2 = child node
82   # epsilons : hash table such that epsilons[[X]]
83   # is a function that takes N as an argument and


```

```

84 # return a list of N generated values
85 # functions : hash table such that functions[[X]]
86 # is a function that takes a vector of value
87 # as a first argument and a single value as a
88 # second argument, and return a single value
89 # %-----%
90 graph <- create_DAG(edges)
91 Dataset <- matrix(0, nrow = N, ncol = length(V(graph)))
92 parents <- get_parents(edges)
93 ordered_vars <- topo_sort(graph)
94 #Set the names of the columns
95 colnames(Dataset) <- names(ordered_vars)
96 for (k in 1:length(ordered_vars)){
97   #Get the name of the variable
98   X <- as_ids(ordered_vars[k])
99   epsilon <- epsilons[[X]](N)
100   #Get the index in ordered_vars of the parents variables of X
101   parents_index <- as.integer(ordered_vars[[parents[[X]]]])
102   if (length(parents_index)>0){
103     Dataset[, k] <- functions[[X]](Dataset[, parents_index], epsilon)
104   }else{
105     Dataset[, k] <- functions[[X]](epsilon)
106   }
107 }
108 return(Dataset)
109 }

```

B.2 Simulations

Here is the complete  script of the simulations discussed in Section 6.3.

```

1 ##### Simulations #####
2
3 #---- Working Directory ----
4
5 (WD <- getwd())
6 if (!is.null(WD)) setwd(WD)
7
8 #---- Libraries ----
9
10 library(scatterplot3d)
11 library(rgl)
12 library(car)
13 source("Script R - Generation.R")
14
15 #---- Inverse probability weighted mean estimator ----
16
17 ipwm <- function(a, L, A, Y){
18   logistic_reg <- glm(A~L, family = binomial(logit))
19   P <- logistic_reg$fitted

```

```

20   mean(Y*(A==a)/(P^a*(1-P)^(1-a)))
21 }
22
23 #---- Conditional mean estimator ----
24
25 cond_mean <- function(a, D){
26   mean(D[D[, "A"]==a, "Y"])
27 }
28
29 #---- Empirical influence function ----
30
31 eif <- function(ells, alpha, upsilons, a, L, A, Y){
32   n <- length(Y)
33   mu <- ipwm(a, L, A, Y)
34   A_prime <- c(A, alpha)
35   values <- NULL
36   for (i in 1:length(ells)){
37     L_prime <- c(L, ells[i])
38     Y_prime <- c(Y, upsilons[i])
39     values[i] <- (n+1)*(ipwm(a, L_prime, A_prime, Y_prime) - mu)
40   }
41   return(values)
42 }
43
44 #---- Basic Case ----
45
46 # Number of observations generated
47 N_obs <- 10^4
48
49 #Edges for the causal graph
50 edges <- rbind(
51   c("X", "A"),
52   c("A", "Y"),
53   c("X", "Y")
54 )
55
56 #Set E of random variables (the generating functions of them)
57 epsilons <- hash::hash(
58   c("X", "A", "Y"),
59   c(function(N) rnorm(N), function(N) rnorm(N), function(N) rnorm(N))
60 )
61
62 #Set F of functions that provides us the system of equation S
63 functions <- hash::hash(
64   c("X", "A", "Y"),
65   c(function(e) e, function(x,e) (x+e>0), function(d,e) d[, "A"]*exp(d[, "X"
66     ""]) + e)
67 )
68
69 #We set the seed so that we have the same generation each time
70 set.seed(42);

```

```

70 D <- generate_causal_dataset(N_obs, edges, epsilons, functions)
71
72 #Set margins
73 par(mar = c(5, 5, 4, 2))
74
75 #Plot of Y as a function of X
76 plot(D[, "X"], D[, "Y"], col=D[, "A"]+1, pch=16,
77       main = "Plot of Y as a function of X", xlab= "X", ylab = "Y")
78 legend(x="topleft", inset = c(0.0000002, 0),
79        legend=c("A=0","A=1"),col=c(1,2), pch=16, title = "Treatment")
80
81 #Computation of the different estimates
82 mu1 <- ipwm(1, D[, "X"], D[, "A"], D[, "Y"])
83
84 mu_cond1 <- cond_mean(1, D)
85
86 mu0 <- ipwm(0, D[, "X"], D[, "A"], D[, "Y"])
87
88 mu_cond0 <- cond_mean(0, D)
89
90 #Empirical influence function
91
92 N_points <- 20
93 xs <- seq(-15,-10,length.out = N_points)
94 ys <- seq(0,100,length.out = N_points)
95
96 for(a in c(0,1)){
97   for(alpha in c(0,1)){
98     zs <- outer(xs, ys, function(x,y) eif(x,alpha,y,a,D[, "X"],D[, "A"],D
99       [, "Y"]))
100     persp(xs, ys, zs, theta = 30, phi = 30, expand = 0.5, col = "
101       lightblue",
102           shade = 0.5, ticktype = "detailed", nticks = 5, xlab = "x",
103           ylab = "y",
104           zlab = "EIF", main = paste("EIF when a=", a, "and alpha=",alpha
105             ))
106   }
107 }
108
109 #Approximate asymptotic distribution
110
111 N_sim <- 10^3
112 N_obs <- 10^4
113
114 estimates_1 <- NULL
115 estimates_0 <- NULL
116
117 set.seed(42);
118 for (i in 1:N_sim){
119   D <- generate_causal_dataset(N_obs, edges, epsilons, functions)
120   estimates_1[i] <- ipwm(1, D[, "X"], D[, "A"], D[, "Y"])

```

```

117 estimates_0[i] <- ipwm(0, D[, "X"], D[, "A"], D[, "Y"])
118 }
119
120 #Case a=1
121 hist(estimates_1, freq = F, main = "Histogram of the estimates when a=1",
122       xlab = "estimates")
123 curve(dnorm(x,mean(estimates_1),sd(estimates_1)), lwd=2, col="red", add=T
124       )
125 legend(x="topleft", inset = c(0.0000002, 0),
126       legend=c(paste("N(",round(mean(estimates_1),2),",",round(sd(
127         estimates_1),2),")")),
128       col="red", lwd=2, title = "Density")
129 qqPlot(estimates_1, line = "robust", main = "QQ-plot of the estimates
130       when a=1", ylab = "estimates")
131 shapiro.test(estimates_1)
132
133 #Case a=0
134 hist(estimates_0, freq = F, main = "Histogram of the estimates when a=0",
135       xlab = "estimates")
136 curve(dnorm(x,mean(estimates_0),sd(estimates_0)), lwd=2, col="red", add=T
137       )
138 legend(x="topleft", inset = c(0.0000002, 0),
139       legend=c(paste("N(",round(mean(estimates_0),2),",",round(sd(
140         estimates_0),2),")")),
141       col="red", lwd=2, title = "Density")
142 qqPlot(estimates_0, line = "robust", main = "QQ-plot of the estimates
143       when a=0", ylab = "estimates")
144 shapiro.test(estimates_0)

```

List of Figures

2.1	Example of a simple causal graph.	10
2.2	Causal graph of the influence of the study time.	11
2.3	Example of a specific causal graph that will induced a simplified joint distribution.	14
2.4	Chain structure.	15
2.5	Causal graph with conditioning on Y	15
2.6	Fork structure.	16
2.7	Collider structure.	17
3.1	Graphical representation of the pseudo-population.	28
4.1	Influence function of the statistical functional of the mean for the standard normal distribution.	75
4.2	Influence function of the statistical functional of the median for the standard normal distribution.	77
4.3	Influence function of the statistical functional of the variance for the standard normal distribution.	78
6.1	Causal graph of a simple structural causal model.	99
6.2	Extended causal graph of a simple structural causal model.	99
6.3	Causal graph of the treatment effect and the outcome.	102
6.4	Simplified causal graph of the treatment effect and the outcome.	102
6.5	Extended causal graph for the case.	103
6.6	Plot of the causal dataset.	104
6.7	Empirical influence functions for the different values of a and α	105
6.8	Histograms of the estimates of $\hat{\mu}_{a,n}^{IPW}$	106
6.9	QQ-plots of the estimates of $\hat{\mu}_{a,n}^{IPW}$	106

List of Tables

6.1	Table of estimates by treatment.	104
6.2	Shapiro-Wilk tests for normality of the estimates of $\hat{\mu}_{a,n}^{IPW}$	107

List of Algorithms

1	Generation of a causal dataset	100
2	Computing parent nodes of a node	101

Bibliography

- [1] D. F. Andrews *et al.*, *Robust estimates of location: Survey and advances*, Princeton Legacy Library, Princeton University Press, Princeton, N.J., 1972.
- [2] Vladimir I. Averbukh and Oleg G. Smolyanov, *The theory of differentiation in linear topological spaces*, Russian Mathematical Surveys **22** (1967), no. 6, pp. 201–258, doi: 10.1070/rm1967v022n06abeh003761.
- [3] Vladimir I. Averbukh and Oleg G. Smolyanov, *The various definitions of the derivative in linear topological spaces*, Russian Mathematical Surveys **23** (1968), no. 4, pp. 67–113, doi: 10.1070/rm1968v023n04abeh003770.
- [4] Patrick Billingsley, *Probability and measure*, 3. ed., A Wiley-Interscience publication, John Wiley & Sons, Inc, New York (NY), 1995.
- [5] Stéphane Bonhomme, Koen Jochmans, and Martin Weidner, *A neyman-orthogonalization approach to the incidental parameter problem* (2024), doi: 10.48550/ARXIV.2412.10304.
- [6] George E. P. Box, *Science and statistics*, Journal of the American Statistical Association **71** (1976), no. 356, pp. 791–799, doi: 10.1080/01621459.1976.10480949.
- [7] Stephen P. Boyd and Lieven Vandenberghe, *Convex optimization*, 7th ed., Cambridge University Press, Cambridge, 2009.
- [8] Ronald Brown, *Topology and groupoids*, 3. ed., www.groupoids.org, Deganwy, 2006.
- [9] Gustavo Canavire-Bacarreza, Luis Castro Peñarrieta, and Darwin Ugarte Ontiveros, *Outliers in semi-parametric estimation of treatment effects*, Econometrics **9** (2021), no. 2, doi: 10.3390/econometrics9020019.
- [10] Weihua Cao, Anastasios A. Tsiatis, and Marie Davidian, *Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data*, Biometrika **96** (2009), no. 3, pp. 723–734, doi: 10.1093/biomet/asp033.
- [11] George Casella and Roger L. Berger, *Statistical inference*, 2. ed., Duxbury, Pacific Grove, Calif., 2002.
- [12] Chao Cheng, Liangyuan Hu, and Fan Li, *Doubly robust estimation and sensitivity analysis for marginal structural quantile models*, Biometrics **80** (2024), no. 2, ujae045, doi: 10.1093/biomtc/ujae045.

- [13] Victor Chernozhukov, Whitney K. Newey, and Rahul Singh, *Automatic debiased machine learning of causal and structural effects*, *Econometrica* **90** (2022), no. 3, pp. 967–1027, doi: 10.3982/ecta18515.
- [14] Victor Chernozhukov *et al.*, *Applied causal inference powered by ml and ai*, arXiv, March 4, 2024, doi: 10.48550/ARXIV.2403.02467.
- [15] Victor Chernozhukov *et al.*, *Double/debiased machine learning for treatment and structural parameters*, *The Econometrics Journal* **21** (2018), no. 1, pp. C1–C68, doi: 10.1111/ectj.12097.
- [16] Andreas Christmann and Arnout Van Messem, *Bouligand Derivatives and Robustness of Support Vector Machines for Regression*, *Journal of Machine Learning Research* **9** (2008), pp. 915–936.
- [17] Donald L. Cohn, *Measure theory: Second edition*, 2nd ed. 2013, Birkhäuser Advanced Texts, Springer Science and Business Media LLC, New York, NY, 2013.
- [18] Thomas H. Cormen *et al.*, *Introduction to algorithms*, Fourth edition, The MIT Press, Cambridge, Massachusetts, 2022.
- [19] Martin Cousineau *et al.*, *Estimating causal effects with optimization-based methods: a review and empirical comparison*, *European Journal of Operational Research* **304** (2023), no. 2, pp. 367–380, doi: <https://doi.org/10.1016/j.ejor.2022.01.046>.
- [20] Christophe Croux, *Limit behavior of the empirical influence function of the median*, *Statistics & Probability Letters* **37** (1998), no. 4, pp. 331–340, doi: 10.1016/S0167-7152(97)00135-1.
- [21] P.Laurie Davies and Ursula Gather, *Robust Statistics*, Papers 2004,20, Berlin, 2004.
- [22] P.Laurie Davies and Ursula Gather, *The breakdown point—examples and counterexamples*, *REVSTAT – Statistical Journal* Volume **5** (2007), no. 1, pp. 1–17, doi: 10.57805/revstat.v5i1.39.
- [23] Reinhard Diestel, *Graph theory*, Fifth edition, Graduate Texts in Mathematics, Springer-Verlag, Heidelberg, Berlin, 2017.
- [24] Zili Dong, *Wright’s path analysis: Causal inference in the early twentieth century*, *THEORIA. An International Journal for Theory, History and Foundations of Science* **39** (2024), no. 1, pp. 67–88, doi: 10.1387/theoria.24823.
- [25] Céline Esser, *Probabilités*, Course notes, University of Liège, 2022.
- [26] Gerald B. Folland, *Real analysis: Modern techniques and their applications*, 2. ed., Pure and applied mathematics, John Wiley & Sons, Inc, New York, 1999.
- [27] Gentiane Haesbroeck, *Compléments de statistique multivariée*, Course notes, University of Liège, 2023.
- [28] Gentiane Haesbroeck, *Statistique non paramétrique*, Course notes, University of Liège, 2024.

- [29] Frank R. Hampel, *The influence curve and its role in robust estimation*, Journal of the American Statistical Association **69** (1974), no. 346, pp. 383–393, doi: 10.1080/01621459.1974.10482962.
- [30] Frank R. Hampel *et al.*, *Robust statistics: the approach based on influence functions*, John Wiley & Sons, Inc, 1986, doi: 10.1002/9781118186435.
- [31] Miguel A. Hernán and James M. Robins, *Causal inference: What if*, Chapman & Hall/CRC, Boca Raton, 2024.
- [32] Peter J. Huber and Elvezio Ronchetti, *Robust statistics*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, Inc, Hoboken, N.J., 2009.
- [33] Liao Jiangang and Rohde Charles, *Variance Reduction in the Inverse Probability Weighted Estimators for the Average Treatment Effect Using the Propensity Score*, Biometrics **78** (2021), no. 2, pp. 660–667, doi: 10.1111/biom.13454.
- [34] Olav Kallenberg, *Foundations of modern probability*, Probability and Its Applications, Springer-Verlag, New York, NY, 1997, Description based on publisher supplied metadata and other sources.
- [35] Joseph D. Y. Kang and Joseph L. Schafer, *Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data*, Statistical Science **22** (2007), no. 4, pp. 523–539, doi: 10.1214/07-STS227.
- [36] Jae-Kwang Kim and David Haziza, *Doubly robust inference with missing data in survey sampling*, Statistica Sinica **24** (2014), pp. 375–394, doi: 10.5705/ss.2012.005.
- [37] Achim Klenke, *Probability theory: A comprehensive course*, 2nd ed., Universitext, Springer-Verlag, London, 2014.
- [38] Paola Lecca, *Machine Learning for Causal Inference in Biological Networks: Perspectives of This Challenge*, Frontiers in Bioinformatics **1** (2021), doi: 10.3389/fbinf.2021.746712.
- [39] Gilles Louppe, *Introduction to artificial intelligence*, Course notes of the chapter 7, Université de Liège, 2023.
- [40] Jared K. Lunceford and Marie Davidian, *Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study*, Statistics in Medicine **23** (2004), no. 19, pp. 2937–2960, doi: <https://doi.org/10.1002/sim.1903>.
- [41] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai, *Robust statistics: Theory and methods*, John Wiley & Sons Ltd, Chichester, England, 2006.
- [42] Samuel Nicolay, *Théorie de la mesure*, Course notes, University of Liège, 2024.
- [43] Henry E. Niles, *Correlation, causation and Wright’s theory of “path coefficients”*, Genetics **7** (1922), no. 3, pp. 258–273, doi: 10.1093/genetics/7.3.258.

- [44] J. Pearl, “Statistics, Causality, and Graphs”, in: *Causal Models and Intelligent Data Management* (Alex Gammerman, eds.), Springer, Berlin, 1999, pp. 3–16, ISBN: 978-3-642-58648-4, doi: 10.1007/978-3-642-58648-4_1.
- [45] Judea Pearl, *Causal inference in statistics: A primer*, John Wiley & Sons, Inc, Chichester, 2016 (Madelyn Glymour and Nicholas P. Jewell, eds.)
- [46] Charles Chapman Pugh, *Real mathematical analysis*, 2nd ed. 2015, Springer eBook Collection, Springer, Cham, 2015.
- [47] Sidney I. Resnick, *A probability path*, Springer eBook Collection, Birkhäuser, Boston, MA, 2014.
- [48] Michel Rigo, *Théorie des graphes*, Course notes, University of Liège, 2009.
- [49] James Robins, *A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect*, Mathematical Modelling **7** (1986), no. 9–12, pp. 1393–1512, doi: 10.1016/0270-0255(86)90088-6.
- [50] James M. Robins, *Marginal structural models versus structural nested models as tools for causal inference*, Statistical models in epidemiology, the environment, and clinical trials (M. Elizabeth Halloran and Donald Berry, eds.), Springer, New York, NY, 2000, pp. 95–133.
- [51] Julia M. Rohrer, *Causal inference for psychologists who think that causal inference is not for them*, Social and Personality Psychology Compass **18** (2024), no. 3, doi: 10.1111/spc3.12948.
- [52] Steven Roman, *Advanced linear algebra*, 3. ed., Graduate texts in mathematics, no. 135, Springer, New York, NY, 2008.
- [53] Paul R. Rosenbaum and Donald B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika **70** (1983), no. 1, pp. 41–55, doi: 10.1093/biomet/70.1.41.
- [54] Donald Rubin, *Estimating causal effects of treatments in experimental and observational studies*, ETS Research Bulletin Series **1972** (1972), no. 2, pp. i–31, doi: <https://doi.org/10.1002/j.2333-8504.1972.tb00631.x>.
- [55] Bryan P. Rynne and Martyn A. Youngson, *Linear functional analysis*, 2nd ed., Springer Undergraduate Mathematics Series, Springer-Verlag, London, 2008.
- [56] Olli Saarela, David A. Stephens, and Erica E. M. Moodie, *The role of exchangeability in causal inference*, Statistical Science **38** (2023), no. 3, pp. 369–385, doi: 10.1214/22-sts879.
- [57] David Schmid, John H. Selby, and Robert W. Spekkens, *Unscrambling the omelette of causation and inference: The framework of causal-inferential theories*, 2021, doi: 10.48550/ARXIV.2009.03297.
- [58] Jean-Pierre Schneiders, *Calcul Différentiel*, Course notes, University of Liège, 2021.

- [59] Shaun R. Seaman and Stijn Vansteelandt, *Introduction to double robust methods for incomplete data*, Statistical Science **33** (2018), no. 2, doi: 10.1214/18-sts647.
- [60] Robert J. Serfling, *Approximation theorems of mathematical statistics*, Paperback ed., A Wiley-Interscience publication, John Wiley & Sons, Inc, New York, N.Y., 2002.
- [61] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9*, Statistical Science **5** (1990), no. 4, pp. 465–472, doi: 10.1214/ss/1177012031.
- [62] Anastasios Tsiatis, *Semiparametric theory and missing data*, Springer Series in Statistics Ser. Springer, New York, NY, 2006.
- [63] Aad W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.
- [64] Willem M. Van Der Wal and Ronald B. Geskus, *Ipw: an r package for inverse probability weighting*, Journal of Statistical Software **43** (2011), no. 13, pp. 1–23, doi: 10.18637/jss.v043.i13.
- [65] Arnout Van Messem, *Robustness and consistency results for support vector machines*, PhD thesis, 2011, p. 175.
- [66] Jürgen Voigt, *A course on topological vector spaces*, 1st ed., Compact Textbooks in Mathematics, Birkhäuser Cham, 2020, doi: 10.1007/978-3-030-32945-7.
- [67] Richard Von Mises, *On the Asymptotic Distribution of Differentiable Statistical Functions*, The Annals of Mathematical Statistics **18** (1947), no. 3, pp. 309–348, doi: 10.1214/aoms/1177730385.
- [68] Edmund T. Whittaker and George N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge mathematical library, Cambridge University Press, Cambridge, 1996.
- [69] Sewall Wright, *Correlation and causation*, Journal of agricultural research **20** (1921), no. 7, pp. 557–585.
- [70] Xing Wu et al., *Causal inference in the medical domain: a survey*, Applied Intelligence **54** (2024), no. 6, pp. 4911–4934, doi: 10.1007/s10489-024-05338-9.
- [71] Michael Zimmert, *The Finite Sample Performance of Treatment Effects Estimators based on the Lasso*, May 14, 2018, doi: 10.48550/ARXIV.1805.05067.