Causal Populations Identification througth Hidden Distributions Estimation

Celine Beji

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Université Paris Dauphine

Causal Populations Identification through Hidden Distributions Estimation

Soutenue par
Celine Beji
Le **28/10/2021**

École doctorale n°543
ED SDOSE

Spécialité
Informatique

Composition du jury :

Raphaël PORCHER Université de Paris	*Président du jury*
Isabelle BLOCH Sorbonne Université	*Rapporteur*
Céline HUDELOT Ecole Centrale Paris	*Rapporteur*
Jamal ATIF Université Dauphine-PSL	*Directeur de thèse*
Florian YGER Université Dauphine-PSL	*Co-encadrant*

Dauphine | PSL

Acknowledgement

Parce que tout phénomène ou effet (cette thèse) n'est rarement dû qu'à une seule cause (mon travail), je me dois de remercier toutes les personnes qui ont contribué à l'élaboration de ce manuscrit.

En premier lieu, je remercie mon directeur de thèse Jamal Atif de m'avoir accordé sa confiance, d'avoir su guider et faire évoluer nos axes de recherches ainsi que pour toutes ses idées visionnaires. Je remercie également Florian Yger pour avoir co-encadré mes travaux. Merci d'avoir été à mes côtés, merci pour ta présence et ta bonne volonté sans faille. Je remercie les rapporteurs de cette thèse, Isabelle Bloch et Céline Hudelot, pour avoir lu minutieusement ce manuscrit, pour toutes leurs remarques pertinentes et leurs pistes d'améliorations. Je remercie Raphaël Porcher, président du jury, de m'encourager dans ce domaine de recherche et pour toutes les perspectives en médecine qu'il m'a fait découvrir. Je suis honorée de l'intérêt qu'il porte à mes travaux et impatiente de travailler davantage avec lui.

Je remercie mes co-auteurs, avec qui j'ai pris beaucoup de plaisir à travailler. Merci Eric pour toutes tes idées, tes encouragements, et ta bienveillance, tu as su me motiver même dans les moments les plus difficiles. Merci Michael pour ton travail et ta détermination, merci de m'avoir soutenu et d'avoir été au fond du code avec moi. Je tiens également à remercier toute l'équipe MILES pour tous les groupes de lecture, leurs conseils et leur aide. Merci particulièrement à Alexandre A., de m'avoir accordé de son temps et de s'être penché sur mes programmes. Je remercie également Mehdi A. d'avoir été mon meilleur voisin de bureau et même beaucoup plus, d'avoir relu et relu certains de mes travaux, et d'avoir toujours cherché des solutions à chacun de mes problèmes.

Je remercie tous les professeurs que j'ai pu avoir durant ma scolarité, et particulièrement mon professeur de maths durant toutes mes années de lycée et mon professeur de maths en Maths Sup, qui ont réussi à me communiquer leur amusement pour cette discipline. Je remercie également grandement Christian Robert mon professeur et responsable de M2, de m'avoir introduit au milieu

de la recherche à travers sa passion pour les méthodes de Monte-Carlo et d'avoir toujours répondu à chacune de mes sollicitations.

Je remercie tous les membres du LAMSADE de m'avoir accueilli chaleureusement dans ce laboratoire, ainsi que tout le personnel administratif. Je remercie particulièrement tous mes frères et soeurs, doctorants, qui ont créé cet univers multi-national si agréable dans les bureaux du 6e. Merci pour tous ces moments passés ensemble, pour toutes ces discussions passionnantes, ainsi que pour m'avoir fait ressentir pour la première fois que je me trouvais au bon endroit.

Je me dois de remercier infiniment l'ADEM qui fut mon refuge tant physique que psychologique pendants toutes mes années de thèses. Merci pour toutes ces magnifiques rencontres, auparavant, je n'avais jamais vu autant de belles âmes dans si peu de mètres carré. Merci à tous de m'avoir permis de m'élever ! Plus généralement, je remercie tout le personnel et les étudiants de l'Université qui créent à Dauphine un véritable environnement d'épanouissement et d'enrichissement personnel.

Je remercie également mes amis. Merci à mes zootopiettes et à la team Willy toujours au rendez-vous, merci à tous les docteurs et futurs docteurs pour toutes nos sorties, jeux, aventures... Et également à l'immortel doctorant Olivier. Merci à mon complémentaire, Loubna, d'être ce qu'elle est. Et un grand merci à tous les Khringos, élu meilleur groupe de tous les temps, vous avez réussi à me faire rire comme jamais !

Par ailleurs, je remercie ma famille, merci d'être là quoiqu'il en soit. Un grand merci particulièrement à mes neveux pour leur amour inconditionnel et à ma petite minette pour avoir été mon épaule pour pleurer, mais également pour rire quand il le fallait. Merci de m'avoir écoutée et d'avoir été la plus grande supportrice de mes idées mégalomanes. Je remercie également profondément mes parents à qui je dédie cette thèse. Merci, à mon père d'être mon pilier, c'est toi qui m'as appris à être forte, sans toi, je n'aurais jamais pu entreprendre et mener cette thèse. Merci, à ma mère d'être mon abri, c'est toi qui m'autorises à être faible, parfois... Sans toi, je n'aurais jamais pu achever cette thèse.

Je ne peux malheureusement pas citer toutes les personnes qui ont contribué à ce travail, mais je remercie toutes les personnes qui ont pu croiser ma route, et encore plus toutes celles qui l'ont partagées. Je vous suis extrêmement reconnaissante, c'est vous qui avez fait de moi ce que je suis aujourd'hui et c'est à vous que je dois cet accomplissement.

Enfin, je ne pourrais terminer sans remercier Dieu d'avoir aligné l'univers pour rendre tout cela possible et de m'avoir donné le courage, la force et la patience d'accomplir ce modeste travail.

Abstract

In the counterfactual framework (also named Rubin framework), which considers the causal inference as a missing data problem, this thesis proposes an approach based on density estimation. The aim is to infer the effect of a treatment on an outcome by estimating the probability distribution of four causal populations, defined by the outcomes with and without treatment: responders who display the expected outcome only when treated, doomed and survivors who respectively never and consistently display the expected outcome, and anti-responders who display the expected effect only when not treated. This classification enables the estimation of the individual treatment effect and the establishment of a treatment assignment policy for new individuals. The fundamental problem is that the two outcomes are not simultaneously observable. In this thesis, two models based on constraints from partial information are proposed. The constraints are built from the observed outcome and the assigned treatment, which allow the exclusion of two causal populations and consequently enforce their probability distribution at zero. For example, if a treated individual produces the expected outcome, it necessarily belongs to the responder or survivor population.

First, a parametric approach based on an adjustment of the EM algorithm is proposed. The parameters of the causal populations distributions are estimated under the outlined constraints. The algorithm is presented and implemented on a mixture of Gaussian distributions and then on a mixture of independent Gaussian and Multinomial distributions, in order to improve the results in the presence of categorical variables. Second, a non-parametric approach is introduced. The model uses an Auto-Encoder neural network enhanced by a causal prior, materialized by a mask introduced in the intermediate layer of the network. The features are reconstructed after being reduced to the hidden latent space, representing the probability distribution of causal populations. The number of optimal units on the latent variables manifold are discussed and experiments are carried out on synthetic and real-life datasets. The limitations of each approach are discussed and alternative models are proposed. An extension to multi-treatments and the open-ended question of non-compliance conclude this work.

Résumé

En se plaçant dans un cadre contrefactuel (également appelé cadre de Rubin), pour lequel l'inférence causale est considérée comme un problème à données manquantes, cette thèse propose une approche basée sur une estimation de densité. L'objectif est d'inférer l'effet d'un traitement sur un résultat en estimant la distribution de probabilité de quatre populations causales, définies par les résultats observés avec et sans traitement : les répondants qui présentent le résultat attendu uniquement lorsqu'ils sont traités, les condamnés et les survivants qui ne présentent respectivement jamais et toujours le résultat attendu, et les anti-répondants qui présentent l'effet attendu uniquement lorsqu'ils ne sont pas traités. Cette classification permet d'estimer l'effet individuel du traitement et d'établir une politique d'affectation pour de nouveaux individus. Le problème fondamental est que les deux résultats ne peuvent pas être simultanément observables. Dans cette thèse, deux modèles basés sur des contraintes sont proposés. Les contraintes sont construites sur le résultat observé et le traitement assigné, qui permettent d'exclure deux populations causales et, par conséquent, de forcer leur distribution de probabilité à zéro. Par exemple, si un individu traité présente le résultat attendu, il appartient nécessairement à la population des répondants ou des survivants.

Tout d'abord, une approche paramétrique basée sur une adaptation de l'algorithme EM est proposée. Les paramètres des distributions des populations causales sont estimés sous les contraintes définies. L'algorithme est présenté et implémenté sur un mélange de distributions Gaussiennes, puis sur un mélange de distributions Gaussiennes et Multinomiales indépendantes afin d'améliorer les résultats en présence de variables catégorielles. Ensuite, une approche non-paramétrique est introduite. Le modèle utilise un Auto-Encoder amélioré par un apriori causal, matérialisé par un masque introduit dans la couche intermédiaire du réseau. Les caractéristiques sont reconstruites après avoir été réduites à l'espace latent, qui est assimilé à la distribution de probabilité des populations causales. Des expérimentations sont menées sur des données synthétiques et réelles, les limites des approches sont discutées et des modèles alternatifs sont proposés. Enfin, une extension en multi-traitements et la question ouverte de la non-conformité concluent ce travail.

List of abbreviation

R/D/S/A	Responders/Doomed/Survivors/Anti-responders
AUUC	Area Under the Uplift Curve
ATE	Average Treatment Effect
CAE	Causal-Auto-Encoder
CAR	Causal Accuracy Rank
CART	Classification And Regression Trees
CATE	Conditional Average Treatment Effect
DAG	Direct Acyclic Graph
ECM	Expectation-Causality-Maximization
EM	Expectation-Maximization
GAN	Generative Adversarial Net
IPM	Integral Probability Metrics
IHDP	Infant Health and Development Program
ITE	Individual Treatment effect
MCMC	Markov Chain Monte Carlo
MMD	Maximum Mean Discrepancy
MSE	Mean Squared Error
PEHE	Precision in estimation of heterogeneous effects
RCM	Rubin Causal Model
RCT	Randomised Controlled Trials
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic
SCM	Structural Causal Model
SSL	Semi-Supervised Learning
SUTVA	Stable Unit Treatment Value Assumption

List of Symbols

$\mathbb{N}$	Set of natural integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\mathbb{R}^{n \times d}$	Set of $n \times d$ real-valued matrices
$\mathcal{X}$	Set of covariates or features (multi-dimensional variables which can influence the outcome)
$\mathcal{Y}$	Set of outcomes (objects of interest)
$\mathcal{D}$	Dataset
$\mathbf{M}$	An arbitrary matrix
V	An arbitrary vector
x	An arbitrary scalar
$\|A\|$	Determinant of a matrix A ($det(A)$ or $detA$)
$\mathbb{P}(X)$	Probability of a arbitrary variable X
$\mathbb{P}(X \mid Y)$	Conditional probability of a arbitrary variable X given a subset of values Y
$\mathbb{E}[X]$	Expected value of an arbitrary variable X
$\mathbb{E}[X \mid Y]$	Conditional expectation of an arbitrary variable X given a subset of values Y
$\mathcal{L}$	Likelihood function
$\mathbb{1}$	Indicator function
L	Lagrangian function
$\hat{\theta}$	Estimator of a given quantity θ
$\mathcal{N}(\mu, \sigma^2)$	Normal (or Gaussian) distribution of mean μ and std σ

Contents

List of Figures

List of Tables

xx

1 Introduction

Contents

1.1 Introduction to causal inference

From our youngest age, we learn the rules of cause and effect through the observations of our repeated experiences. If a baby pushes an object off the table, the object will fall to the ground and break. Although we are all familiar with this concept, defining causality remains a challenging task. Causality has been defined philosophically by Plato as "everything which becomes must of necessity become owing to some Cause; for without a cause it is impossible for anything to attain becoming"(28a) [Plato, 1925]. However, causality is not just a relationship between a cause and an effect, it is a phenomenon of explanation, mainly concerned with answering a "Why" question [Pearl and Mackenzie, 2018].

1.1.1 Causality as an essential resource for machine learning

Causality have been directly applied in many fields, such as healthcare, economics, social science [Holland, 1986], but it is crucial to highlight this current and growing impact in machine learning. Currently, machine learning methods are achieving outstanding performances, in particular, due to the advances in deep neural networks. However, as they become more complex, machine learning models become less interpretable and explainable. Besides, these models tackle by construction correlation questions and do not help answering the fundamental how and what if questions. Causality can bring solutions to these concerns and widen the scope of these models allowing them to gain in predictive and explainability abilities. In the following, the machine learning tasks where causality is key are highlighted.

Interpretability More and more machine learning models are seen as black boxes with unexplained decisions for humans, making them hard to trust despite their high performances. In order to explain the process behind an efficient model, interpretability in the context of causality can answer questions such as "Why did this model produce this result" and "What characteristics are responsible for this result?". The aim is then to find out whether the results are consistent with human ethics and whether the data are biased and lead to an irrational decision. In addition to intrinsically interpretable algorithms (providing an interpretation at the training time) and post-hoc interpretations (generating explanations for the made decisions), causal information can explain what decisions would have been made in an alternative situation. Pearl argues in [Pearl, 2018] that the counterfactual framework is the crucial way to achieve the highest degree of interpretability. In this vein, the authors in [Harradon et al., 2018, Chattopadhyay et al., 2019] have shown how to construct a graphical causal model to generate explanations of a deep neural network predictions. Other approaches, based on post-hoc interpretability have been proposed as well [Ribeiro et al., 2016].

Invariance Causality and invariance are closely linked. Learning an invariant predictor is equivalent to finding a data representation on which the optimal model on that representation is the same for all environments. The *spurious correlations* (i.e. the relationships between variables that do not have direct causal links and the sources of noise or exogenous variability) should be eliminated in order to find an invariant representation. Indeed, such relationships are not stable across environments and would harm the model performances. Hence, determining the true causal relationship between variables is a key to properly tackle the problem [Arjovsky et al., 2019]. Note that, not only the knowledge of causal relationships allows to find relevant invariances, but reciprocally, invariance allows to infer causality links among variables [Bühlmann et al., 2020].

Domain adaptation Domain adaptation aims at predicting an outcome on a domain that differs from the training domain. The authors in [Zhang et al., 2013] explain how domain adaptation can be seen as a causal inference problem. The intervention in the causal field can be related to the perturbation of a system in the domain adaptation, which leads to a change in the initial distribution. Causal information can then be used to explain changes in the domains of the

data distribution. More recently, the authors in [Magliacane et al., 2017] propose an approach to predict the distribution of a variable from other variables observed in one or more target domains. The idea is to select a subset of features that satisfies the invariance of the prediction in each domain. The domain adaptation problem is solved by using causal inference to predict invariant conditional distributions i.e. by estimating the conditional probability of the outcome given a subset of features remains the same in the training and target domains.

Fairness Fairness, which attempts to restore impartiality to decisions made in a biased and unfair manner, based on covariates such as gender or ethnicity of an individual, can also be considered in a causality framework. The authors in [Kusner et al., 2017] study the counterfactual fairness. The outcome predicted by the model in the real world (observed environment) is compared to the outcome predicted on a *counterfactual* (defined in Section 1.2) non-observed world. If the decision in the real world is the same as the one in the counterfactual world, the model is considered fair, otherwise unfair. The goal of fairness can be seen as the intention to remove the effects of *confounding variables* (defined in Section 1.1.3) assimilated to unfair criteria on the prediction of a final decision.

Reinforcement Learning (RL) More and more work in RL are investigating the direction of causal relationships between variables to explain the behaviour of learning agents, for instance in [Madumal et al., 2020, Zhang, 2020]. The structural causal model is learned during reinforcement learning to estimate the causal relationships between variables of interest and provide a model which performs significantly better.

More generally, causality can be used for any problem where there is a shift in the distribution of data. This can be the case in a multi-agent system, where the rules of the game change [Peysakhovich et al., 2019] and in adversarial attacks for which modified test examples are not drawn from the same distribution as the training examples [Goodfellow et al., 2014]. The knowledge of the causal structure improves the robustness of the predictive models with respect to changes of the input statistics and explain decisions and actions of the system.

1.1.2 Two main causality approaches

Since the work of Jerzy Neyman in 1923, edited a few years later in [Splawa-Neyman et al., 1990], modelling causality has been the subject of numerous studies in the fields of statistics and probability theory. Since them, two main schools are now dominant: Judea Pearl with the development of the Structural Causal Model (SCM) [Pearl, 1995, Pearl et al., 2000, Pearl, 2009] and Donald Rubin, who gave his name to the Rubin Causal Model (RCM), also known as the Neyman-Rubin Causal Model, based on a counterfactual approach [Rubin, 1974, Rubin, 1975, Rubin, 2003].

An SCM, described thoroughly in [Pearl et al., 2016, Section 1.5], is defined by structural equations that relate on endogenous variables (observed and internal to the model) and exogenous variables (unobserved variables, external to the model and imposed by the environment). An

SCM can be associated to a graphical causal model obtained with Directed Acyclic Graphs (DAGs). Graphical properties can also be used to estimate the causal effect.

The RCM, detailed in the following section, focuses on the causal relationship between a treatment variable and an outcome. In order to remove the effect of unconsidered external variables, the problem is addressed as a missing data problem. This model is based on the estimation of what would have happened in a hypothetical different situation i.e. under other treatment occurrences.

1.1.3 Concepts and definition of causal inference

The aim of this section is to define the causal inference in a statistical and probability framework. We refer to *causal inference* as the process by which a causal relationship can be established between a cause and its effects.

Causation and prediction

Prediction and causation are distinct, but complementary. Prediction is useful for forecasting and then planning future events. It answers questions such as "What will happen?", "Which individuals will be affected by a variable" and "Knowing X, is it more likely to have Y?". Causation has the objective of explaining an event and respectively answers to "What will happen under some changes?", "Why individuals will be affected by a variable" and "If we changed X, how would it change Y?". Prediction is defined in terms of a joint distribution of statistic conditions. The models of interest predict a variable Y after observing $X = x$ and thus estimates $\mathbb{P}(Y \mid X = x)$. To the opposite, causation definition requires dynamic conditions.

Definition 1.1 (Causal inference)**.** [Pearl et al., 2016, Section 1.5] *Causal inference is the process to make a prediction about the change of an outcome Y under the change of an input X, i.e. models predict Y after setting* $X = x$*. The object of interest is* $\mathbb{P}(Y \mid \text{set } X = x)$.

The relationship between X and Y can be illustrated by a DAG. If X causes Y, then X is called the *cause* and Y the *effect* (see Figure 1.1).

An effect can have more than one cause and none of which alone can explain it in its entirety. Moreover, causation has the property to be:

- Transitive: If X is a cause of Y and Y a cause of Z, then X is a cause of Z.
- Irreflexive: X cannot cause itself.
- Antisymmetric: If X is a cause of Y, then Y is not a cause of X.

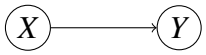


Figure 1.1: DAG of the causal relationship between the cause X and the effect Y.

Causal effect estimation

The effect of a variable is seldom due to a single cause. To estimate the causal inference of a variable of interest T on an outcome Y, we must consider all variables that affect Y. These

variables are of two categories: covariates and confounders.

Definition 1.2 (Covariates)**.** [Rubin, 2003, Subsection 5] *Covariates are the observed features of an individual (or a unit) potentially influencing the outcome Y and unaffected by the variable of interest T.*

Definition 1.3 (Confounder)**.** [Pearl et al., 2016, Section 3.4] *A confounder or confounding variable is an unobserved variable (exogenous to the model) which affects both, the variable of interest T and the outcome Y, and explains part of the causal effect of T on Y.*

Consider we study the effect of aspirin on headaches. The headache is the outcome variable. The age and the weight of patients can be considered as covariates. Taking an aspirin does not affect these features, but these variables may alter the patient's reaction. These variable could be observed and considered in the effect of aspirin. However, other unobserved variables may also influence the effect, such as level of fatigue or degree of hydration. These variables, which are not detectable, are confounding variables.

To estimate the causal effect of a variable T on an outcome Y, without the effect of the confounders, the idea is to compare a variation of the outcome relative to a baseline. The variations of Y for given values of T knowing covariates X should be observed. Because for each value of T the effect of confounders is assumed to be constant, the comparison of different values of T cancels out therefore the influence these variables.

The effect of aspirin on headache is estimated by observing the variation in headache with and without aspirin for a given patient age and weight. As the level of fatigue or the degree of hydration is considered constant with and without treatment for a given patient, the study of these two conditions reduces the influence of these variables.

A manipulation, called *intervention* must then be required (see [Pearl et al., 2016, Section 3.1]). The idea of action is formalized by Pearl with *do-calculus* introduced in [Pearl, 1995]. The distribution created by the intervention $T = t$ is given by the conditional probability of Y, knowing $X = x$ and setting: $\mathbb{P}(Y \mid X = x, do(T = t))$ or $\mathbb{P}(Y \mid X = x, \text{set } T = t)$, and the variable T is called *treatment*.

Definition 1.4 (Treatment)**.** *[Rubin, 2003, Subsection 1] The variable, for which effects are to be investigated, is called* treatment*. The intervention then consists of forcing each of the potential values of the treatment, in order to observe the ensuing outcomes.*

The intervention is to be distinguished from a simple observation. While $\mathbb{P}(Y \mid X = x, T = t)$ is an observed probability, $\mathbb{P}(Y \mid X = x, do(T = t))$ is estimated by the intervention of forcing the variable T to the value t. In addition, note that forcing the variable T to the given value t assumes there is an alternative treatment value t' to which the individual could have been exposed at the same time.

Taking or not an aspirin are the different values of the treatment. A patient decides to take the aspirin, but it can decide at the same time not to take it (or to take an alternative medicine).

1.2 Counterfactual approach

1.2.1 Potential, factual and counterfactual outcome

The counterfactual approach is based on the *potential outcome*, defined by the combination of the outcome on all potential treatment values.

In the RCM, the effect of a new drug is estimated from the outcome with and without treatment. If the treatment is given to a patient, the question is "what would have happened without treatment?". And if the treatment is not given, the question is "what would have happened with treatment?"

To formalize the problem, let us consider several observed instances $i \in \{1, .., n\}$, with $n \in \mathbb{N}$. They are individuals (or units or examples) in the database. An individual $i \in \{1, .., n\}$ can be a single person, place or thing at a particular time. Thus, if the same individual was observed at two different times, it would be considered as two different individuals. Let $\mathcal{X} \subset \mathbb{R}^d$ be an d-dimensional continuous or discrete space and $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ an i.i.d. sample, where each $\mathbf{x}_i \in \mathcal{X}$ are the continuous or discrete features. We note $y_{i,obs} \in \mathcal{Y}$ the observed outcome and t_i the treatment assignment of the individual $i \in \{1, .., n\}$, and respectively $\mathbf{y}_{obs}$ and $\mathbf{t}$ are the sets for all individuals. The treatment variable can be binary, discrete or continuous.

Prescribing or not prescribing a given medical treatment is a case of binary treatment. Choosing between several medications is an example of multiple discrete treatment. And, when a doctor assigns a specific dosage of a drug, this is a continuous case.

When there are L potential treatments to assign, for all $i \in \{1, .., n\}$, $t_i \in \{t^0, t^1, .., t^{L-1}\}$. Note that, in most cases, for the sake of convenience, the assumption of a binary treatment is usually made in many works. In binary treatment, where an individual $i \in \{1, .., n\}$ is exposed or not to a treatment, treated units ($t_i = 1$) constitute the *treatment* group, and untreated individuals ($t_i = 0$) the *control* group.

For all individuals $i \in \{1, .., n\}$ and treatments $l \in \{0, .., L-1\}$, $y_{i,l}$ is defined as the outcome of the individual i corresponding to the assigned treatment l. In binary treatment, we note $y_{i,0}$ and $y_{i,1}$ the real and hypothetical outcomes, relative to potential treatments $\{0, 1\}$.

Definition 1.5 (Factual, counterfactual and potential outcomes)**.** [Imbens and Rubin, 2015, Section 1.3] *We call a* factual outcome*, noted $\mathbf{y}_{i,obs}$ for an individual $i \in \{1, .., n\}$, the observed outcome vector associated with the assigned treatment. The* counterfactual outcomes *are the unobserved outcomes $(y_{i,0}, y_{i,1}, .., y_{i,L}) \setminus (y_{i,obs})$ associated with the non-assigned treatments. The* potential outcome*, noted $(y_{i,0}, y_{i,1}, .., y_{i,L})$, is the set of factual and counterfactual outcomes.*

In binary treatment, the potential outcome is the couple $(y_{i,0}, y_{i,1})$. One of the components is observed $y_{i,obs} = t_i\, y_{i,1} + (1 - t_i)\, y_{i,0}$, while the other is missing $y_{i,miss} = (1 - t_i)\, y_{i,1} + t_i\, y_{i,0}$. The observed and hypothetical outcomes are illustrated in Table 1.1. For an individual $i \in \{1, .., n\}$, if the treatment is assigned ($t_i = 1$), then $y_{i,1}$ is observed and $y_{i,0}$ is missing, otherwise $y_{i,0}$ is observed and $y_{i,1}$ is missing.

Observed data			
X	**t**	$\mathbf{y}_0$	$\mathbf{y}_1$
$\mathbf{x}_1$	1	.	$y_{1,1}$
$\mathbf{x}_2$	1	.	$y_{2,1}$
$\mathbf{x}_3$	0	$y_{3,0}$	.
$\mathbf{x}_4$	1	.	$y_{4,1}$
...	...	...	...
$\mathbf{x}_n$	0	$y_{n,0}$	.

Table 1.1: Observed an missing data in binary treatment. For a given individual $i \in \{1, .., n\}$, $\mathbf{x}_i$ is the covariate, $\mathbf{t}$ is the treatment assignment and $(\mathbf{y}_0, \mathbf{y}_1)$ the potential outcomes. "." denoted unobserved values.

Where the effectiveness of a medical treatment is evaluated, an individual is a patient with a specific pathology. For a given individual $i \in \{1, .., n\}$, the treatment is the action to take ($t_i = 1$) or not ($t_i = 0$) a given medication. The outcome is the patient's reaction, i.e. the evolution of the pathology. If the patient takes the treatment, the reaction with treatment $y_{i,1}$ is observed (and named the factual outcome), and its hypothetical reaction without treatment $y_{i,0}$ is missing and must be estimated (it's the counterfactual outcome). In multi-treatment, the effect of two different medical treatments may be compared. The patient's reaction with the first treatment ($t_i = 1$), the second treatment ($t_i = 2$) and without treatment ($t_i = 0$) should be compared. If the first treatment is given to the patient, then $y_{i,1}$ is the factual outcome, $(y_{i,0}, y_{i,2})$ are the counterfactual outcomes, and $(y_{i,0}, y_{i,1}, y_{i,2})$ is the potential outcome.

Causal inference can therefore be deduced from the knowledge of the counterfactual outcomes. However, the fundamental problem is that only the *factual outcome* $\mathbf{y}_{obs}$ is observed while the *counterfactual* outcomes $(\mathbf{y}_0, \mathbf{y}_1, .., \mathbf{y}_L) \setminus (\mathbf{y}_{obs})$ remains unknown. The issue consists in estimating the counterfactual outcome.

1.2.2 Strong ignorability assumption

To solve the problem of the missing counterfactual outcome and allow the estimation of the causal effect, some assumptions are necessary. The counterfactual approach is based on the Stable Unit Treatment Value Assumption (SUTVA), introduced in [Rubin, 1978, Rosenbaum and Rubin, 1983]. It consists of both:

- *Non-interference*: There is no interference among individuals. The potential outcome of any individual is not affected by the treatment assignment to the other individuals (note that this assumption can be violated when the general balance is disrupted).

  If we consider a couple living in the same house, we assume that if one partner changes the way he or she cooks in response to the treatment, there will be no impact on the other partner's response to the treatment.

- *No variation in treatment*: The treatments for all individuals are comparable. For example, in binary treatment, there is only one treatment and control state.

  > *A treatment variation can occur when multiple doses of a drug are given while only considering whether or not to take the treatment.*

To obtain an unbiased estimator of causal inference, two more assumptions must be made [Rosenbaum and Rubin, 1983, Imbens and Rubin, 2015]:

- *Unconfoundedness (or Ignorability)* assumption consists of the conditional independence on covariates between potential outcome and the treatment assignment: $(\mathbf{y}_0, \mathbf{y}_1, .., \mathbf{y}_L) \perp\!\!\!\perp \mathbf{t}|\mathbf{X}$. It means that all variables that jointly affect the potential outcomes and the treatment are observed (within the covariates).

  > *If two individuals have exactly the same features, their potential outcomes would be the same, irrespective of the assigned treatment.*

  In this way, factual and counterfactual outcomes play a similar role in the model. All the outcomes that compose the potential outcome will be considered in the same way and therefore the observed outcome is not distinguished from the counterfactual outcomes. Note that in practice, it is not possible to confirm this assumption because the counterfactual outcomes are unobservable. The available data does not allow to deduce a conditional independence of counterfactual outcomes.

- *Overlap (or common support or positivity)* assumption provides a non-zero probability of receiving the treatment, i.e. $0 < \mathbb{P}(T = 1 \mid X = \mathbf{x}) < 1$.

  > *Every patient has the possibility to receive the treatment, for example without financial constraints.*

The strong ignorability assumption consists in SUTVA, unconfoundedness and overloap assumptions.

1.3 Challenges and main contributions of the thesis

In this thesis, we consider the couterfactual outcome (RCM) framework, for which the causal inference is seen as a missing data problem. We focus on the causal relationship between two variables, a treatment variable and a observed outcome, conditional on given covariates. We restrict to binary treatment, which is a standard assumption due to the complexity of the problem. The aim is to estimate the treatment effect, estimated by the difference between the outcomes from the treatment and control groups, from observed data. As one outcome is observed and the other is missing, an intuitive way is to divide data into groups according to the treatment assignment and create estimators for each group. The problem is that, in most cases, these groups are not homogeneous and therefore not comparable.

1.3.1 The challenge of selection bias and observational data

The datasets can be collected from two types of studies: *experimental* and *observational*. Experimental studies have the advantage of being controlled by humans. This is an active approach, which sets up an experimental framework to control the assignment of the treatment. In *randomised controlled trials* (RCT), where the treatment is randomly assign and with a large number of individuals, two homogeneous treatment and control groups can be constructed and compared to the estimation of treatment effect. In this context, the correlation between the data is significant and "true" conclusions can be made. However, this kind of study is expensive, time-consuming to execute, or often difficult to implement in real-life for many reasons, such as ethical concerns. In practice, an *assignment policy* is established, allowing a decision on the treatment state to which an individual should be exposed. Individuals receiving the treatment are selected according to an apriori of the outcome based on their features, leading to a *selection bias* in the data. The challenge is therefore to estimate *heterogeneous treatment effects*.

> *In epidemiological research, to estimate the effect of a drug, an experimental trial is conducted. The drug may be assigned according to a policy based on a patient's characteristics, such as gender, age or medical history.*

The second type of study is observational. It is a passive approach, in which data are extracted from past observations. There is a growing demand to use this data because of its abundance and ease of collection. However, they are more delicate to use because of the unknown framework in which they were generated. Since the treatment assignment is not known, a strong confounding bias may exist in the data and produces false conclusions or paradoxes. The challenge is also to estimate the treatment effect from data with selection bias.

1.3.2 From average to individual treatment effect

For many years, the treatment effect has been studied through the investigation of the Average Treatment Effect (ATE), defined as the difference of expectation of individual potential outcomes [Morgan and Winship, 2015, Section 2.4]:

$$\tau_0 := \mathbb{E}[Y_1 - Y_0] \tag{1.1}$$

Under SUTVA assumption, it can be expressed by averaging over the distribution of X:

$$\begin{aligned} \tau_0 &= \mathbb{E}[\mathbb{E}[Y \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0]] && (1.2)\\ &= \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] && (1.3)\\ &= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] && (1.4) \end{aligned}$$

The objective is to reduce the selection bias by creating homogeneous groups or sub-groups, and estimate the treatment effect in average in each group, with matching, stratification or weighting methods, to cite a few (see Appendix B).

With the growing interest in personalisation like in personalised medicine, personalised marketing and personalised policy, the estimation the Individual Treatment Effect (ITE) or unit-level causal effects of an individual $i \in \{1, .., n\}$, is required. The ITE is defined as the difference between the outcome with and without treatment [Morgan and Winship, 2015, Section 2.2]:

$$ITE := y_{i,1} - y_{i,0} \tag{1.5}$$

Remark: While ITE seem to be preferred in the medical context, the term *uplift* is more common within the marketing community. Other terms are also used, such as *true lift* [Lo, 2002], *incremental value modeling*, *association* [Wasserman, 2013, chap. 19], or *c-specific treatment effects* [Shpitser and Pearl, 2006].

Estimating ITE and ATE are two different approaches. ITE is the effect of a treatment for each individual, while ATE is the average effect of the treatment on the whole population.

In epidemiology, to assess the effectiveness of a medical treatment, A/B tests are usually conducted and relate to the ATE estimation, as they study the average effect on the whole population. The outcome of a group of patients is observed with (A) and without (B) treatment. If on average, the efficacy reaches a given threshold, the drug may be introduced to the market (or goes to the next phase of trial). If we now consider the individual level, the focus is not on the efficacy of the treatment on average, but rather on the efficacy on a given individual. Regardless of the ATE, the importance is to target patients who have a decent ITE. Then, a treatment could be rejected because of a poor ATE but accepted with respect to the ITE.

ITE is not directly identifiable because for each unit in the training dataset only the outcome of the assignment treatment is observed, the other is unknown. A way to overcome this is to rely on the Conditional Average Treatment Effect (CATE), defined as the expected difference between the two potential outcomes (expectation of the ITE) [Morgan and Winship, 2015, Section 2.7.1]:

$$\tau(\mathbf{x}) := \mathbb{E}[Y_1 - Y_0 | X = \mathbf{x}] \tag{1.6}$$

With the conditional independence of Y on X:

$$\tau(\mathbf{x}) = \mathbb{E}[Y_1 \mid X = x] - \mathbb{E}[Y_0 \mid X = \mathbf{x}] \tag{1.7}$$

Moreveover, the consistency assumption implies that:

$$\tau(\mathbf{x}) = \mathbb{E}[Y_1 \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y_0 \mid X = \mathbf{x}, T = 0] \tag{1.8}$$

Thus, under strong ignorability assumption:

$$\tau(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0] \tag{1.9}$$

It can also be expressed with the *do*-calculus of Pearl's notation:

$$\tau(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 1)] - \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 0)] \tag{1.10}$$

The authors in [Künzel et al., 2019] demonstrate that the best estimator for the CATE is the best estimator for the ITE in terms of the mean squared error (MSE). The strong ignorability assumption allows to identify the causal effect without hidden confounders effect and the consistency of estimators. It is a satisfactory condition for the identifiability of the CATE function, although weaker versions of this assumption are sufficient. [Imbens and Wooldridge, 2009, Pearl, 2017, Shalit et al., 2017]. When an assignment policy is adopted, $\mathbb{E}[Y \mid T = 1] \neq \mathbb{E}[Y_1]$ and the CATE is an unbiased estimate that cannot be obtained by directly comparing the treatment groups. This issue is currently the subject of a considerable number of works (see Chapter 2), of which this thesis is a part.

1.3.3 A causal population perspective

In binary treatment ($t_i \in \{0, 1\}$) and binary outcome ($(y_{i,0}, y_{i,1}) \in \{0, 1\}^2$), a classification of individuals according to the intersection of the two potential outcomes, produces four *causal populations* [Wasserman, 2013, Section 19.1]:

- *Responders* (R) display the expected reaction only when they are treated: $y_{i,0} = 0$ and $y_{i,1} = 1$. They benefit from the treatment.
- *Doomed* (D) never display the expected effect: $y_{i,0} = 0$ and $y_{i,1} = 0$. The treatment is useless to them.
- *Survivors* (S), named also *always healthy*, always display the expected effect: $y_{i,0} = 1$ and $y_{i,1} = 1$. The treatment is useless to them.
- *Anti-responders* (A), named also *hurt*, display the expected effect only when they are not treated: $y_{i,0} = 1$ and $y_{i,1} = 0$. The treatment is noxious for them.

In order to identify the ITE, we propose in this thesis a perspective based on the structure of causal populations and we study the following issue:

How to estimate the probability distributions of causal populations?

The counterfactual outcome is therefore predicted by assuming that individuals belong to the group with the highest probability. Moreover, we prove in Section 3.1.2, that this approach allows to estimate the ITE. In contrast to related work that directly estimate the ITE, this approach provides additional information on the existing causal inference between the outcome and the treatment, which tackle a more general problem. Knowing the distribution of causal populations allows to classify individuals and predict the optimal treatments to assign. In addition to a simple prediction of counterfactual outcomes, this modelling provides a policy to support the treatments assignment for current individuals. Thus we can target individuals for whom the treatment is beneficial, those for whom the treatment has no effect and those for whom it is noxious.

1.3.4 Main contributions

To estimate the probability distribution of causal groups, we propose two approaches: a parametric and a non-parametric approach.

Expectation-Causality-Maximization (ECM) algorithm

The first one is based on a new variant of the Expectation-Maximization algorithm. We reformulate our causal problem as a missing data problem and introduce constraints depending on the characteristics of causal groups. We put a prior on the data distribution and estimate the distribution parameters by maximizing a suitably designed likelihood. We introduce a partial information based on knowledge about the treatment and the observed factual outcome, which excludes for each individual two of the four potential causal groups. The proposed algorithm, coined Expectation-Causality-Maximization (ECM), is illustrated on a Gaussian mixture for which experiments on synthetic and real datasets are carried out to prove its efficiency compared to the state-of-the-art. However, we additionally conduct experiments on datasets mixing discrete and continuous variables that are more realistic in real-life, by using a mixed normal-multinomial distribution. One of the advantages of this algorithm is its extension to any distribution. It can be adapted to any distribution for which the log-likelihood has a close form. If that is not the case, we also devise an extension with variational methods that can support such distributions.

Causal Auto-Encoder (CAE)

The second approach, consists in estimating the probability distribution of the causal groups using an Auto-Encoder. The covariates, given in input, are re-constructed with a deep neural network. The latent space is assimilated to the probability distributions of causal groups, through a mask that forces some hidden neurons to zero. The mask is designed with causal groups partial information extracted from the treatment and the observed factual outcome. The size of the latent space is discussed, and the efficiency of our approach is demonstrated in experiments on synthetic and real-life datasets. This approach has the significant advantage of processing large datasets and being easily scalable for multi-treatments. In addition, it is suitable for non-parametric distributions.

Outline of the thesis

In order to present these contributions, the thesis is organised as follows. In Chapter 2, we present the most relevant works relating to ours and the ITE estimation. Although this problem has been studied for many years using statistical models, we will focus here on recent models based on advances in machine learning. Then, Chapters 3 and 4 are devoted to a thorough presentation and results obtained from the two proposed contributions. Finally, in Chapter 5, we discuss the perspectives and future work. We first present an extension of the causal population approach in multi-treatments and propose an adaption of our models. We then consider the question of how to address the challenge of non-compliance.

2 Related work and model evaluation

Contents

In this chapter, we present a non-exhaustive but representative overview of existing related work. We focus on work that addresses the estimation of ITE and CATE. Due to the extensive literature, finding a classification is a challenging task. The authors in [Künzel et al., 2019] used a meta-learner classification. The first category, named T-learners, refers to algorithms with "two" separate models for the treatment and control group. This is a two steps approach. It estimate the ITE with an intermediate step that estimates the treatment effect through two response functions:

$$\begin{aligned} \mu_0(\mathbf{x}) &= \mathbb{E}[Y_0 \mid X = \mathbf{x}] \\ \mu_1(\mathbf{x}) &= \mathbb{E}[Y_1 \mid X = \mathbf{x}] \end{aligned} \tag{2.1}$$

The two response functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are estimated with two arbitrary functions: one for the treatment group f_1 and one for the control group f_0, and then give the CATE estimator $\hat{\tau}(\mathbf{x})$ with the difference.

T-learners

Estimation of treatment response functions:

$$\begin{aligned} \hat{\mu}_0(\mathbf{x}) &= f_0(\mathbf{x}) \\ \hat{\mu}_1(\mathbf{x}) &= f_1(\mathbf{x}) \end{aligned} \tag{2.2}$$

Estimation of the CATE:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}) \tag{2.3}$$

Any estimator can be used as base learner. In RCT, where the treatment is randomized, this simple methods has the advantage to be easy interpretable. However, selection bias between the treatment and control groups or errors in the two models could add up and distort the ITE estimate.

In the same line, X-learners algorithm estimate two functions for each treatment $\hat{\mu_0}$ and $\hat{\mu_1}$, with standard base learners. However, they do not directly estimate the CATE by the difference of the estimators, but they impute the treatment effects for each individual $i \in \{1, .., n\}$ in the database, such as:

$$\tilde{D}_i^1 := y_{i,obs} - \hat{\mu}_0(\mathbf{x}_i) \text{ if } t_i = 1 \tag{2.4}$$

$$\tilde{D}_i^0 := \hat{\mu}_1(\mathbf{x}_i) - y_{i,obs} \text{ if } t_i = 0 \tag{2.5}$$

$\tau_1(\mathbf{x}) = \mathbb{E}[\tilde{D}^1 \mid X = \mathbf{x}]$ and $\tau_0(\mathbf{x}) = \mathbb{E}[\tilde{D}^0 \mid X = \mathbf{x}]$ are then estimated and noted respectively $\hat{\tau}_1$ and $\hat{\tau}_0$. In a other step, the CATE is computed by a weighted average of the two estimates:

$$\hat{\tau}(\mathbf{x}) = w(\mathbf{x})\,\hat{\tau}_0(\mathbf{x}) + (1 - w(\mathbf{x}))\,\hat{\tau}_1(\mathbf{x}) \tag{2.6}$$

where $w \in [0, 1]$ is a weight function.

In [Stadie et al., 2018], an extension of the X-learners approach using a neural network is proposed and named Y-learners. Two neural networks are used, one to obtain the estimates $\hat{\mu}_0$ and $\hat{\mu}_1$, and the

other to estimate the CATE. The use of neural network allows to capture the dependence between the two estimation steps. The two estimates can be jointly optimised by back-propagation and networks can be trained on the same data.

An other approach consists to estimate an unique response function, named S-learners for "Single". The selection bias aims to be reduce by sharing information across the treatment and control estimator. The function takes as argument the covariates $\mathbf{x}$ and treatment assignment t. The two responses surfaces are given by the function for each treatment values. This function is then used with the two values of treatment assignment to estimate the CATE $\hat{\tau}(\mathbf{x})$, as the difference.

S-learners

Estimation of treatment response functions:

$$\begin{aligned} \hat{\mu}_0(\mathbf{x}) &= f(\mathbf{x}, t = 0) \\ \hat{\mu}_1(\mathbf{x}) &= f(\mathbf{x}, t = 1) \end{aligned} \tag{2.7}$$

Estimation of the CATE:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}) \tag{2.8}$$

Many of models proposed in recent years belong to this category, and in particular those with neural network architecture for which the treatment is given as input. We overview these models, by distinguishing in Section 2.1 the models based on two different distributions between the treatment and control groups with the objective of reducing the bias between them and in Section 2.2 the models based on the estimation of one distribution for both groups.

Then in Section 2.3, we focus on indirect approaches that estimate directly the ITE. Unlike the previous meta-learners, they do not have an intermediate step of estimating the response surface, but directly estimate the ITE. However, the boundary between these two approaches is sometimes extremely narrow, and some models fall into both categories at the same time.

In Section 2.4, we introduce datasets and metrics that can be uses to evaluate the ITE, in the challenging context of unknown true causal effect. Finally, we position our work in relation to the state-of-the-art presented and to closely related work in Section 2.5.

2.1 Models based on two distributions

In this section, we present methods that use the idea of estimating two different distributions for each treatment group, but reduce the bias by creating subgroups (Section 2.1.1), minimizing the distance between the distributions (Section 2.1.2), introducing a function with two outputs (Section 2.1.3) and by using the propensity score (Section 2.1.4).

2.1.1 Subgroup analysis

A "Virtual Twins" method is introduced in [Foster et al., 2011]. The aim is to partition the covariate space into two subgroups A and A^C, defined respectively as the group for which the treatment effect is better than the average treatment effect, and its complementary. The two responses functions are estimated with (two separate or not) random forest, introduced by [Breiman, 2001]. When an individual $i \in \{1, .., n\}$ is assigned to the group $t_i \in \{0, 1\}$, its response on the group t_i is given from the out-of-bag estimate (i.e. using sub-sampling with replacement to create training sets) from the random forest and its response on the group $1 - t_i$ is given by applying the random forest with the treatment group switched. Then, the ITE is estimated by the difference. The aim is to identify which covariate describe the treatment effect variation. Two alternative methods are proposed:

- A regression procedure VT(R): A regression tree is used to estimate $\hat{\tau}(\mathbf{x})$ for each individual. If the obtained value of $\hat{\tau}(\mathbf{x})$ is greater than some cutoff c, the individuals are defined to be in the group $\hat{A}$ (defined as the estimation subgroups A).
- A classification procedure VT(C): A new binary variable is defined as equal to 1 if $\hat{\tau}(\mathbf{x}) > c$ and 0 else. This variable is used as a outcome in the construction of the classification tree. The individuals are then classified being in $\hat{A}$ or not.

An extension of the VT model, with two separate random forests combined with synthetic forests (see [Ishwaran and Malley, 2014]), is compared in [Lu et al., 2018] with other methods. Experiments show that this adaptation outperforms all compared methods in large sample sizes.

2.1.2 Distance between distributions

The authors in [Shalit et al., 2017] give a theoretical analysis and family of algorithms, based on *representation learning*. The idea is to learn jointly the distribution of treatment and control, while learning a representation that minimises the total factual loss. This allows to learn a representation making the distribution of control and treatment groups balanced, i.e. with a similar distribution of learned representations for different treatment populations. The function f_0 and f_1 are learned under a constraint that encourages better generalization across the treated and control groups. This method has the advantage to provide theoretical guarantees.

In order to reduce the differences of the treatment and control distributions $p(\mathbf{x} \mid t = 0)$ and $p(\mathbf{x} \mid t = 1)$, they propose to use *Integral Probability Metrics* (IPM) [Müller, 1997] to measure distances between distributions:

$$IPM_G := \sup_{g \neq G} \left| \int_{\mathcal{X}} g(\mathbf{x}) \left(p(\mathbf{x} \mid t = 1) - p(\mathbf{x} \mid t = 0) \right) d\mathbf{x} \right| \tag{2.9}$$

where G is the family of functions $g : \mathcal{X} \to \mathbb{R}$. They show that the expected loss of the model, expressed as the expected Precision in Estimation of Heterogeneous Effect:

$$\epsilon_{PEHE} = \int_{\mathcal{X}} \left(\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x}) \right)^2 p(\mathbf{x}) \, d\mathbf{x} \tag{2.10}$$

is upper bounded by the error of counterfactual learning plus a IPM term. The error of counterfactual learning is given by the sum of the expected factual and counterfactual losses:

$$\epsilon_F(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} l_{h,\Phi}(\mathbf{x}, t)\, p(\mathbf{x}, t)\, d\mathbf{x}dt \tag{2.11}$$

$$\epsilon_{CF}(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} l_{h,\Phi}(\mathbf{x}, 1-t)\, p(\mathbf{x}, t)\, d\mathbf{x}dt. \tag{2.12}$$

where $\Phi : \mathcal{X} \to \mathcal{R}$ is the representation function, $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ is the potential outcome defined over the new representation space $\mathcal{R}$ and $l_{h,\Phi}(\mathbf{x}, t) = \int_{\mathcal{Y}} L(y_t, h(\Phi(\mathbf{x}), t))\, p(y_t \mid \mathbf{x})\, dy_t$ is the expected loss ($L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function). The expected factual loss ϵ_F represents the standard machine learning loss and the expected counterfactual loss ϵ_{CF} represents the expected loss, relative to the distribution when the treatment assignment is reversed. Note that, in RCT experiments, the IPM term is equal to 0 and the problem is reduced to the estimation of two separate estimators.

This result leads to the construction of the Counterfactual Regression (CFR) algorithm. It is an end-to-end regularization minimization procedure which simultaneously learns the balanced representation of the data and the counterfactual outcome. Two jointly deep neural networks are simultaneously trained to estimate the representation $\Phi(\mathbf{x})$ and the hypothesis of the outcome $h(\Phi(\mathbf{x}), \mathbf{x})$ (see Figure 2.1).

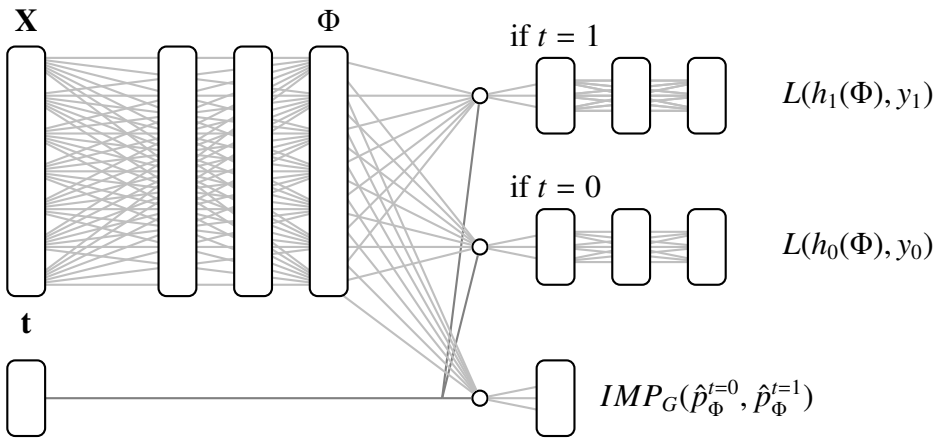


Figure 2.1: CFR architecture

$h_0(\Phi)$ and $h_1(\Phi)$ parameterize respectively the control and the treatment predicted outcomes, but only one head, corresponding to the assignment treatment, is updated to each sample. To reduce the bias induced by the imbalanced distributions, an objective function is minimized under h and Φ. The algorithm uses a gradient descent and an error back-propagation. It has been implemented with both the Wassertein distance (CFR_{WASS}) and the MMD metric (CFR_{MMD}).

The Treatment-Agnostic Representation Network (TARNet) algorithm was proposed at the same time as a reference method to compare the CFR algorithm. It is a variant of the previous algorithm

without balance regularization. It has been used as a baseline in many other papers presenting similar approaches [Shi et al., 2019, Schwab et al., 2018].

2.1.3 Gaussian Processes

The Causal Multi-task Gaussian Processes (CMGPs), introduced in [Alaa and van der Schaar, 2017], aims at modeling potential outcomes with a regression model:

$$\begin{cases} y_{i,0} = f_0(\mathbf{x}_i) + \epsilon_{i,0} \\ y_{i,1} = f_1(\mathbf{x}_i) + \epsilon_{i,1} \end{cases} \tag{2.13}$$

where $\epsilon_{i,0} \sim \mathcal{N}(0, \sigma_0^2)$ and $\epsilon_{i,1} \sim \mathcal{N}(0, \sigma_1^2)$ are Gaussian noises. The model uses a *potential outcomes* function $f(.) : \mathcal{X} \to \mathbb{R}^2$ with $\mathbf{x}$ as input and two outputs, one for each outcome of the treatment groups. The ITE is estimated by $\hat{\tau}(\mathbf{x}) = \hat{f}^T(\mathbf{x})\mathbf{e}$ where $\mathbf{e} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$. f can be modelled with a vector-valued Reproducing Kernel Hilbert Space (vvRKHS), such as $f \sim \mathcal{GP}(0, \mathbf{K}_\theta)$, where $\mathbf{K}_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$ is a reproducing kernel, which is a covariance (symmetric positive semi-definite matrix-valued) function, describing the coupling between the two responses surfaces f_0 and f_1 as follows:

$$K_\theta(\mathbf{x}, \mathbf{x}') = A_0\, k_0(\mathbf{x}, \mathbf{x}') + A_1\, k_1(\mathbf{x}, \mathbf{x}') \tag{2.14}$$

$k_w(\mathbf{x}, \mathbf{x}')$ for $w \in \{0, 1\}$ is the radial basis function and $(A_0\ ,A_1)$ are two parameters determined by the variances $(\beta_{ij}^2)_{ij}$ and correlations $(\rho_i)_i$ of the two response surfaces f_0 and f_1, such that:

$$A_0 = \begin{bmatrix} \beta_{00}^2 & \rho_0 \\ \rho_0 & \beta_{01}^2 \end{bmatrix} \qquad A_1 = \begin{bmatrix} \beta_{10}^2 & \rho_1 \\ \rho_1 & \beta_{11}^2 \end{bmatrix} \tag{2.15}$$

The introduced linear model of coregionalization kernel allows a degree of freedom that makes it possible to have different covariance functions.

The model is a multi-task learning algorithm, which aims to train multiple tasks jointly to increase efficiency and reduce over-fitting of each task. It consists of two alternate learning tasks. Once the potential functions are estimated, a loss function is minimized to reduce the selection bias. It is given by:

$$\mathcal{L}(\hat{f}) = \int_{\mathbf{x}\in\mathcal{X}} (\hat{f}^T(\mathbf{x})\mathbf{e} - \tau(\mathbf{x}))^2\, p(\mathbf{x}) d\mathbf{x} \tag{2.16}$$

The selection bias is reduced by jointly minimizing the empirical error in factual outcomes and the uncertainty in counterfactual outcomes (posterior counterfactual variance). In addition, credible intervals are obtained from the error function of the normal distribution and allow a measure of confidence in the estimation.

2.1.4 Propensity-dropout

The authors in [Alaa et al., 2017] propose a deep counterfactual network using a propensity-dropout (DCN-PD), where the *dropout probability* is expressed as:

$$DP(\mathbf{x}) = 1 - \frac{\gamma}{2} - \frac{1}{2} H(\tilde{p}_s(\mathbf{x})) \tag{2.17}$$

where $0 \leq \gamma \leq 1$ is an offset hyper-parameter, $H(p_s) = -p_s \log(p_s) - (1 - p_s) \log(1 - p_s)$ is the Shannon entropy and $\tilde{p}_s$ is the estimated propensity score for an individual $\mathbf{x}$, defined as the probability to receive the treatment $p_s(\mathbf{x}) = \mathbb{P}(T = 1 \mid X = \mathbf{x})$. The aim is to add a propensity-dropout regularization scheme to reduce the bias of observed data that reflects the underlying policy. The dropout probability imposes a larger penalties on the counterfactual estimation when the propensity score is low.

The architecture of the proposed model is also a multi-task learning which is composed of two networks (see Figure 2.2):

- The first is a network of potential outcomes designed to estimate counterfactual outcomes for the treatment and control groups. It is composed of specific layers (*idiosyncratic layers*) for each treatment group and common layers (*shared layers*) for both groups. The estimation of potential outcomes is considered as two separate learning tasks on the treatment and control groups, but linked by common layers. In practice, the neural network is trained in alternating phases. The data is divided into two task-specific batches of data: one for treated individuals and one for untreated. Only the idiosyncratic layers corresponding to the treatment assignment are updated in a given epoch. The weights of the shared layers are updated in all epochs and capture the commonalities between the two learning tasks.
- The second is a standard feed-forward network used to estimate the propensity score $p_s(x_i)$ via the samples $(\mathbf{x}_i, t_i)$, for all individuals $i \in \{1, .., n\}$. It reduces the selection bias between the treatment and control group. It is used to regularize the potential outcomes network via the dropout scheme. Shared and idiosyncratic layers weights are updated according to the estimated propensity score $\tilde{p}_s$. This regularisation allows the model to focus more on individuals with a high propensity score.

This model has the advantage to capture both the propensity score and the outcomes estimation. It is conceptually analog to propensity weighting [Abadie and Imbens, 2016].

The Perfect Match method proposed in [Schwab et al., 2018] estimate the counterfactual by taking into account the assignment policy. In the vein of TARNet model (previously described in Section 2.1.2), it uses separate heads for each treatment to maintain the effect of the treatment variable. However, the closest matches from the other treatments in term of propensity score, are used to train the other heads. It constructs virtually randomised mini-batches by adding for each individual the unobserved counterfactual outcome given by the outcomes of nearest neighbours. This algorithm is close to DCN-PD algorithms since it implements the balanced scores to dynamically adjust the dropout regularisation strength for each observed sample, depending on its treatment propensity score.

2.2 Models based on a single distribution

In this section, we present S-learner models based on one shared estimators for the two responses functions. The estimator can be modeled with a regression taking the treatment variable as

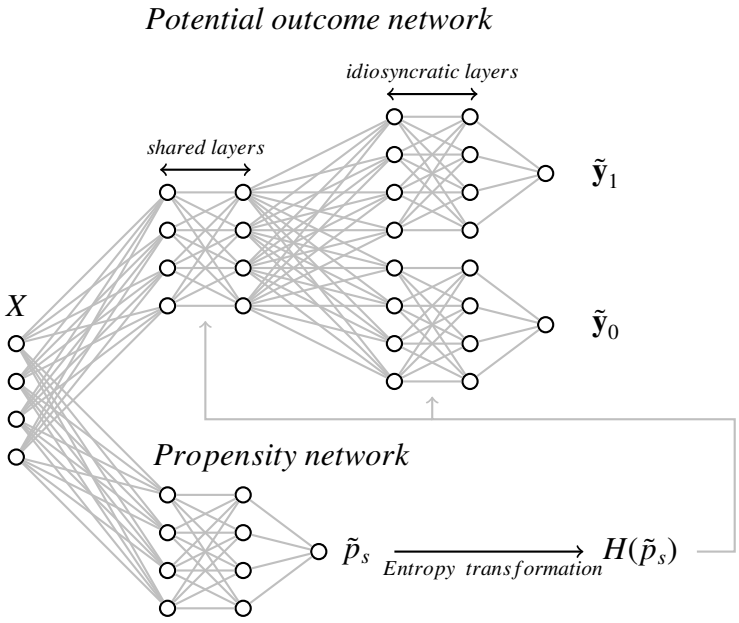


Figure 2.2: DCN-PD architecture

covariate (Section 2.2.1), with a additive regression trees (Section 2.2.2), by using a learning representation (Section 2.2.3) or a GAN architecture (Section 2.2.4).

2.2.1 Treatment as covariate

The first intuitive approach is to consider the treatment variable as a covariate in the model:

$$y_{t_i,i} = b_0 + b_1 \mathbf{x}_i + b_2 t_i + \epsilon_i \tag{2.18}$$

where b_0, b_1 and b_2 are the regression coefficients and ϵ_i the random error term for the individual $i \in \{1, .., n\}$.

This naive approach is easily interpretable and could use any machine learning model. But it performs poorly if there are a large number of covariates because the influence of the treatment variable is lost among the other ones [Künzel et al., 2019].

2.2.2 Bayesian additive regression trees (BART)

The authors in [Hill, 2011] propose an approach for modelling the potential outcomes conditional on treatment assignment and covariates. It uses a Bayesian Additive Regression Tree (BART) model developed in [Chipman et al., 2007, Chipman et al., 2010]. Two elements are combined: a sum-of-trees model and a regularization prior. The inference is modeled by an unknown function f that predicts an outcome:

$$y = f(t, \mathbf{x}) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{2.19}$$

The mean of y, $f(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}]$, is approximated by a sum of m regression trees:

$$f(\mathbf{x}) \approx g_1(t, \mathbf{x}) + g_2(t, \mathbf{x}) + .. + g_m(t, \mathbf{x}) \tag{2.20}$$

where each g_i denotes a binary regression tree. This sum-of-trees model is more flexible than a simple single tree model because it can easily incorporate additive effects. It can take into account high order interaction effects. After the construction of the sum-of-trees, a second step use a prior to regularise the parameters. Coherent posterior intervals are computed using a Markov Chain Monte Carlo (MCMC) sampling. The CATE is then estimated as the difference between the response surface for each treatment value.

BART is very competitive in practice since it can handle a large number of predictors, continuous variables and missing data. However, this method has some limitations. In some circumstances, the default prior specification is inadequate to the problem and no theoretical guarantees have been provided in this first version of the model. The authors in [Hill et al., 2020] propose modifications to the prior to accommodate high-dimensions and smooth regression functions. They also attempt to characterise the theoretical properties that would explain the performance of this model by studying convergence rate.

2.2.3 Learning Representations

The authors in [Johansson et al., 2016] propose a learning representation, which was a precursor to the work presented in Section 2.1.2. They reformulate the problem as a domain adaptation problem, and more specifically as a covariate shift problem. This framework relise in between direct and indirect estimation of ITE. The model process in two steps: a representation $\Phi : \mathcal{X} \rightarrow \mathbb{R}^d$ and the surface response function $f : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}$ are learned and then a regularized squared loss objective on the factual data are minimized. The objective is threefold: (i) minimize the error between the observed outcome and the factual learned representation over a training set, (ii) minimize the error between unobserved counterfactual outcomes and the counterfactual learned representation by adding a penalty that encourages counterfactual predictions to be closed to the nearest observed outcome, and (iii) balance the treated and untreated distributions by minimizing the distance between the learned factual and counterfactual representations. The objective function, to optimize over representation ϕ and hypotheses $h \in \mathcal{H}$, is expressed as:

$$\mathcal{B}(\phi, h) = \frac{1}{n}\sum_{i=1}^{n} |h(\phi(\mathbf{x}_i), t_i) - y_{i,obs}| + \frac{\gamma}{n}\sum_{i=1}^{n} |h(\phi(\mathbf{x}_i), 1 - t_i) - y_{j(i),obs}| + disc(\hat{P}_\phi^F, \hat{P}_\phi^{FC}) \tag{2.21}$$

where α, γ are two positive hyperparameters, $disc$ is the discrepancy measure, $j(i)$ is the nearest neighbor of $\mathbf{x}_i$ among the group that received the opposite treatment and $\hat{P}_\phi^F$ (respectively $\hat{P}_\phi^{FC}$) is the empirical factual (respectively counterfactual) distribution.

This framework leads to the creation of two models. The first, the Balancing Linear Regression (BLR), is a regression model that include the products of the treatment variables with each of the covariate. As these variables can be highly predictive of the outcome, the covariates are here weighted. A weighting diagonal matrix W is applied to the representation $\Phi(\mathbf{x}) = W\mathbf{x}$. It

determines the influence of features. A smaller weight is applied on features that differ a lot between treatment and control groups. By using an alternating sub-gradient descent, minimizing the overall objective maintains a trade-off between maximizing balance and predictive accuracy. The second model, the Balancing Neural Network (BNN), is an architecture composed of hidden layers used to learn the $\Phi(\mathbf{x})$ representation and other layers used to compute the discrepancy measure (see Figure 2.3).

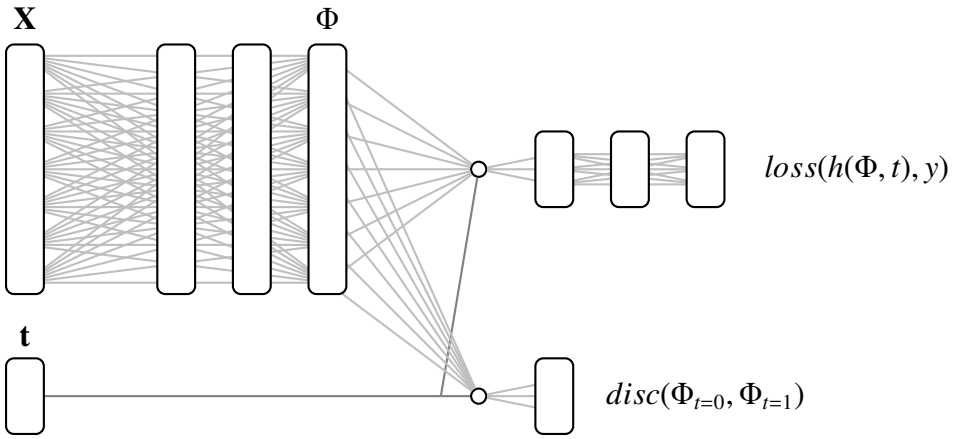


Figure 2.3: BNN architecture

This framework has the advantage to provide theoretical guarantees by deriving an upper bound on the relative counterfactual generalization error. The relative error is bounded by comparing the fitting of a ridge-regression on the factual outcomes and on the distribution with reverse treatment assignment. The limitation of this framework is that the relative error bound does not inform on the absolute quality of the representation. A new framework is then developed based on the reduction of the counterfactual error term.

2.2.4 GAN architecture

Generative Adversarial Networks for inference of Individualized Treatment Effects (GANITE) is a method to estimate the ITE in multi-treatments, by generating the potential outcomes using Generative Adversarial Networks (GANs) [Yoon et al., 2018]. The model consists of two GAN blocks: a *counterfactual block* which aims to build a complete dataset containing the value of the potential outcomes on each of the treatment; and an *ITE block* which estimates the ITE and allows to provide confidence intervals (see overview in Figure 2.4).

The counterfactual generator G generate proxies of the counterfactual outcomes for each sample of the dataset. It takes as inputs the feature vector $\mathbf{x} \in \mathcal{X}$, the factual outcome y_{obs}, a assigned treatment vector $\mathbf{t}^{(k)} = (t^1, .., t^k) \in \{0, 1\}^k$, where k is the number of treatment, and a randomly generated vector $\mathbf{z}_g \sim \mathcal{U}((-1, 1)^{k-1})$. The goal is to generate a potential vector $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_k)$ for all potential treatments by fitting a function $g : \mathcal{X} \times \{0, 1\}^k \times \mathcal{Y} \times [-1, 1]^{k-1} \rightarrow \mathcal{Y}^K$ such that

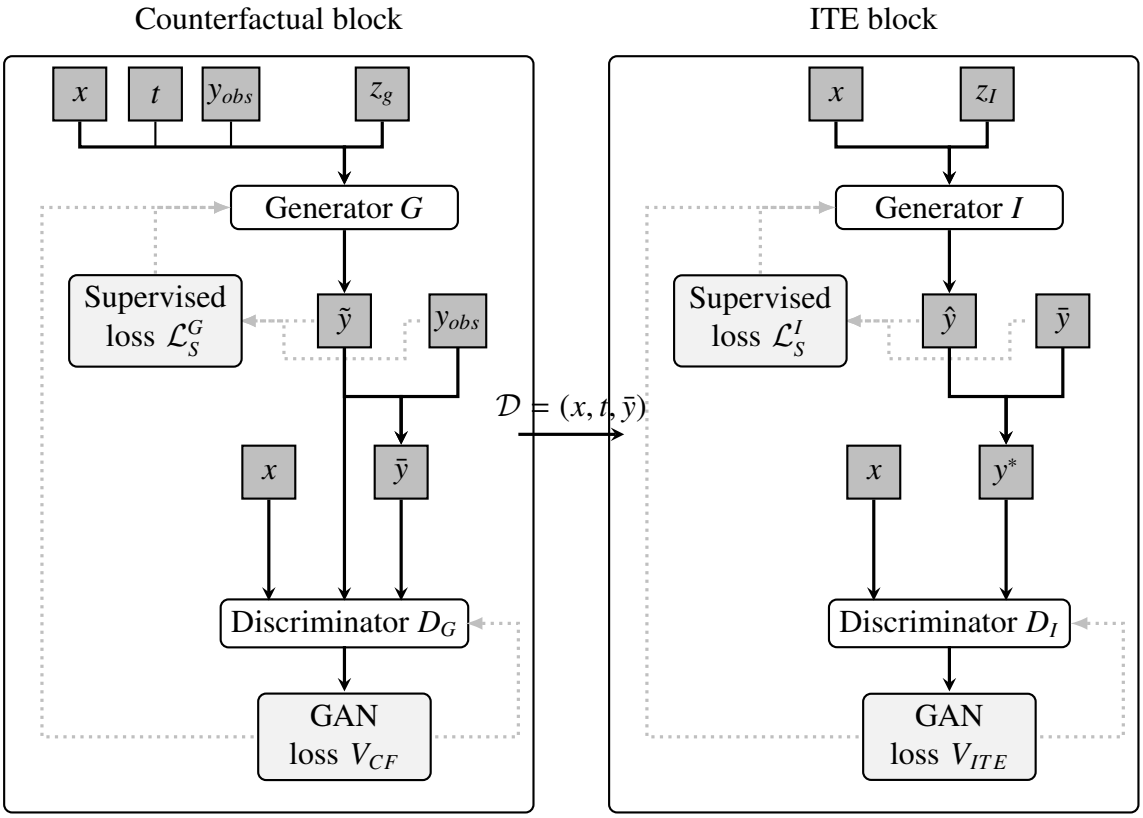


Figure 2.4: GANITE architecture [Yoon et al., 2018]

$G(\mathbf{x}, \mathbf{t}, \mathbf{y}_{obs}) = g(\mathbf{x}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{z}_g)$ and that minimizes:

$$\mathcal{L}_S^G(\mathbf{y}_{obs}, \tilde{\mathbf{y}}(\mathbf{t}^{(k)})) = (\mathbf{y}_{obs} - \tilde{\mathbf{y}}(\mathbf{t}^{(k)}))^2 \tag{2.22}$$

where $\tilde{\mathbf{y}}(\mathbf{t})$ is the predicted potential outcome corresponding to the treatment assignment. After generating a prediction of the potential outcomes $\tilde{\mathbf{y}}$, the vector is modified by replacing the value of the predicted potential outcome of the assigned treatment $\tilde{y}(\mathbf{t})$ by its true observed value y_{obs}. The vector is then denoted $\bar{\mathbf{y}}$.

The counterfactual discriminator D_G attempts to determine which input is the factual and the counterfactual outcome. It uses both $\bar{\mathbf{y}}$ and $\tilde{\mathbf{y}}$ conditional on $\mathbf{x}$ as input and determine which components came from which distribution. It is different from a standard GAN framework, which takes as input a single sample from one of two distributions and attempts to determine which distribution it came from. The minimax problem solved in this block is therefore:

$$\min_G \max_{D_G} \mathbb{E}_{(\mathbf{x},\mathbf{t}^{(k)},\mathbf{y}_f)\sim\mu_f} \left(\mathbb{E}_{\mathbf{z}_g\sim\mathcal{U}((-1,1)^k)} \left({\mathbf{t}^{(k)}}^T \log D_G(\mathbf{x}, \tilde{\mathbf{y}}) + (1 - \mathbf{t}^{(k)})^T \log(1 - D_G(\mathbf{x}, \tilde{\mathbf{y}})) \right) \right) \tag{2.23}$$

where μ_f is the joint distribution of $(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs})$.

Once the dataset $\mathcal{D}$ is completed with the predicted potential outcomes $\bar{\mathbf{y}}$, the ITE block train the ITE estimation function in a supervised way. It includes a generator I and a discriminator

D_I. The generator takes as input a feature vector $\mathbf{x}$ and a random vector $\mathbf{z}_I \sim \mathcal{U}((-1,1)^{k-1})$. Let $h : \mathcal{X} \times [-1,1]^k \to \mathcal{Y}^k$, the objective is to generate a potential outcome vector $\hat{\mathbf{y}}$ by finding a function h such that $I(\mathbf{x}) \sim \mu_Y(\mathbf{x})$, where $\mu_Y(\mathbf{x})$ is the conditional distribution of Y given $X = \mathbf{x}$. The loss to minimize is defined as:

$$\mathcal{L}_S^I(\bar{\mathbf{y}}, \hat{\mathbf{y}}) = \begin{cases} ((\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_0) - (\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0))^2 & \text{for } k = 2 \text{ (binary treatment)} \\ \|\bar{\mathbf{y}} - \hat{\mathbf{y}}\|_2^2 & \text{for } k > 2 \end{cases} \tag{2.24}$$

The discriminator D_I is built as a standard GAN generator. It takes a pair $(\mathbf{x}, \mathbf{y}^*)$, where y^* is randomly the potential outcome vector generated by the generator $\hat{\mathbf{y}}$ or the potential outcome vector $\bar{\mathbf{y}}$ given as input of the block. The discriminator gives the probability that $\mathbf{y}^*$ come from the true dataset $\mathcal{D}$. The block is trained using the following adversarial minimax optimisation:

$$\min_I \max_{D_I} \mathbb{E}_{\mathbf{x}\sim\mu_\mathbf{x}} \left(\mathbb{E}_{\mathbf{y}*\sim\mu_Y(\mathbf{x})} (\log D_I(\mathbf{x}, \mathbf{y}^*)) + \mathbb{E}_{y^*\sim I(\mathbf{x})} (\log(1 - D_I(\mathbf{x}, \mathbf{y}^*))) \right) \tag{2.25}$$

The authors provide empirical loss functions for both the counterfactual and the ITE block and produce empirical justification on the use of a combination of this two blocks.

2.3 Direct estimation

The direct estimation approach aims at estimating the ITE without estimating the potential outcomes. However, note that this method sometimes requires modelling the estimation of counterfactual outcomes. We present in this section some of this approaches. They can be based on a transformation of the outcome of interest (Section 2.3.1), on direct estimation of the impact of the effect (Section 2.3.2), on analysis of the treatment effect in subgroups (Section 2.1.1), on adapting neighbors metrics (Section 2.3.3), or by transformation of the loss function (Section 2.3.5).

2.3.1 Outcome transformation

Class variable transformation

The authors in [Jaskowski and Jaroszewicz, 2012a] propose a simple class variable transformation to directly estimate the ITE for binary outcome values. It introduces a new variable $z \in \{0,1\}$ such as:

$$z = \begin{cases} 1 & \text{if } (y_{obs} = 1 \text{ and } t = 1) \text{ or } (y_{obs} = 0 \text{ and } t = 0) \\ 0 & \text{otherwise.} \end{cases} \tag{2.26}$$

The problem is addressed as the estimation of a unique probability distribution after which the problem can be tackled through a standard machine learning framework:

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}, T = 1) - \mathbb{P}(Y = 1 \mid X = \mathbf{x}, T = 0) = 2\,\mathbb{P}(Z = 1 | X = \mathbf{x}) - 1 \tag{2.27}$$

Modeling the causal effect is equivalent to modeling the conditional distribution of variable Z by a standard classifier. Note that, this method does not allow to deduce the distribution of each of the treatment groups.

Inverse probability weighting

In [Hirano et al., 2003, Hitsch and Misra, 2018], an outcome transformation based on the propensity score $p_s(\mathbf{x}) = \mathbb{P}(T = 1 \mid X = \mathbf{x})$ is proposed as:

$$y_t^{IPW} = t\,\frac{y_{obs}}{p_s(\mathbf{x})} - (1-t)\,\frac{y_{obs}}{1-p_s(\mathbf{x})} \tag{2.28}$$

The use of y^{IPW} provide an noised but unbiased estimator of the CATE.

Double Robust Estimation

Another approach is to use a double robust estimation, based on the following variable transformation [Kang et al., 2007, Knaus et al., 2021]:

$$y_t^{DR} = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + t\,\frac{y_{obs} - \mu_1(\mathbf{x})}{p_s(\mathbf{x})} - (1-t)\,\frac{y_{obs} - \mu_0(\mathbf{x})}{1-p_s(\mathbf{x})} \tag{2.29}$$

The CATE is then estimated by $\tau(\mathbf{x}) = \mathbb{E}[Y^{DR} \mid X = \mathbf{x}]$.

2.3.2 Treatment impact identification

The authors in [Zaniewicz and Jaroszewicz, 2013] introduce Uplift Support Vector Machines (USVMs). The treatment effect is defined as a function $M(\mathbf{x}) : \mathbb{R}^m \to \{-1, 0, 1\}$ which assigns to each individual respectively a positive, neutral or negative impact of the treatment.

- The treatment has a positive impact if it has a positive outcome $+1$ when it is treated and negative prediction -1 when it is not.
- The treatment has a negative impact if the individual has a negative outcome -1 when it is treated and a positive outcome $+1$ when it is not.
- If the outcome is in the same class $+1$ or -1 regardless of whether the treatment is taken or not, the treatment effect is neutral.

The control group and the treatment group are modeled by two parallel hyperplanes $H_0 : \langle t, \mathbf{x}\rangle - b_0 = 0$ and $H_1 : \langle t, \mathbf{x}\rangle - b_1 = 0$ where $b_0, b_1 \in \mathbb{R}$ are the intercepts. The model is expressed as:

$$M(\mathbf{x}) = \begin{cases} +1 & \text{if } \langle t, \mathbf{x}\rangle > b_0 \text{ and } \langle t, \mathbf{x}\rangle > b_1 \\ 0 & \text{if } \langle t, \mathbf{x}\rangle \leq b_0 \text{ and } \langle t, \mathbf{x}\rangle > b_1 \\ -1 & \text{if } \langle t, \mathbf{x}\rangle \leq b_0 \text{ and } \langle t, \mathbf{x}\rangle \leq b_1 \end{cases} \tag{2.30}$$

The two hyperplanes are used to classify the points in each class. H_0 separates positive and neutral points, and H_1 neutral and negative points. The USVM optimization problem can be found by solving:

$$\min_{,b_0,b_1} \frac{1}{2}\langle \mathbf{t}, \mathbf{t}\rangle + C_0 \sum_{D_+^{T1} \cup D_-^{T0}} \xi_{i,0} + C_1 \sum_{D_-^{T1} \cup D_+^{T0}} \xi_{i,0} + C_1 \sum_{D_+^{T1} \cup D_-^{T0}} \xi_{i,1} + C_0 \sum_{D_-^{T1} \cup D_+^{T0}} \xi_{i,1} \tag{2.31}$$

subject to the constraints:

$$\langle t_i, \mathbf{x}_i \rangle - b_0 \geq +1 - \xi_{i,0}, \forall (\mathbf{x}_i, y_i) \in D_+^{T_1} \cup D_-^{T_0} \tag{2.32}$$

$$\langle t_i, \mathbf{x}_i \rangle - b_0 \leq -1 - \xi_{i,0}, \forall (\mathbf{x}_i, y_i) \in D_-^{T_1} \cup D_+^{T_0} \tag{2.33}$$

$$\langle t_i, \mathbf{x}_i \rangle - b_1 \geq +1 - \xi_{i,1}, \forall (\mathbf{x}_i, y_i) \in D_+^{T_1} \cup D_-^{T_0} \tag{2.34}$$

$$\langle t_i, \mathbf{x}_i \rangle - b_1 \leq -1 - \xi_{i,1}, \forall (\mathbf{x}_i, y_i) \in D_-^{T_1} \cup D_+^{T_0} \tag{2.35}$$

where $D_+^T = \{(\mathbf{x}_i, y_i) \in D^T : y_i = +1\}$, $D_-^T = \{(\mathbf{x}_i, y_i) \in D^T : y_i = -1\}$, D^{T_1} and D^{T_0} denote respectively the treatment and control group samples. $\xi_{i,j}$ are variables allowing for misclassified training cases for all $i = \{1, .., n\}$ and $j \in \{1, 2\}$, and verify $\xi_{i,j} \geq 0, \forall i, j$. C_0 and C_1 are penalty parameters. C_0 has a similar function to the penalty coefficient in classical SVMs because it controls the misclassified cost with respect to the margin maximization term $\frac{1}{2}\langle \mathbf{t}, \mathbf{t} \rangle$. The quotient $\frac{C_1}{C_0}$ gives the proportion of individual classified with a positive or negative treatment effect.

2.3.3 Causal trees

The idea is to construct a random forest by using a first partition of the data, and evaluate the treatment effect with the other partition. This approach is different from the tree-based methods presented in the previous sections because rather than splitting tree nodes by maximising the variance of the node, it uses a splitting rule which maximises the difference in treatment within a node. The methods, presented in this section, have the advantage of being adaptable in high dimension by building a large number of regression trees and averaging their predictions. In addition, they provide a theoretical analysis on consistency and asymptotic normality results.

Causal Tree Ensembles

The method, introduced in [Wager and Athey, 2018], is a non-parametric causal forest, which provides an estimation for CATE and constructs asymptotic confidence intervals. The feature space is recursively split into a set of leaves L. Then, the prediction is evaluated by identifying for each leaf $L(\mathbf{x})$ the setting:

$$\hat{\mu}(x) = \frac{1}{|i : \mathbf{x}_i \in L(\mathbf{x})|} \sum_{\{i:\mathbf{x}_i \in L(\mathbf{x})\}} y_{i,obs} \tag{2.36}$$

The individual treatment effect is then estimated by:

$$\hat{\tau}(\mathbf{x}) = \frac{1}{|i : t_i = 1, \mathbf{x}_i \in L|} \sum_{\{i:t_i=1,\mathbf{x}_i \in L\}} y_{i,obs} - \frac{1}{|i : t_i = 0, \mathbf{x}_i \in L|} \sum_{\{i:t_i=0,\mathbf{x}_i \in L\}} y_{i,obs} \tag{2.37}$$

The authors provide "honest" procedures that only use $y_{i,obs}$ in the estimation of the within-leaf treatment effect τ. The learning sample is divided into a subsample for growing the tree and an other for the prediction in the leaves. The tree is defined as honest if it only uses the observed outcome to estimate the within-leaf treatment effect τ. Three properties are then proved:

- The constructed trees are consistent for $\tau(\mathbf{x})$.
- Predictions are asymptotically Gaussian and unbiased: $\frac{\hat{\tau}(\mathbf{x})-\tau(\mathbf{x})}{\sqrt{Var[\hat{\tau}(\mathbf{x})]}} \Rightarrow \mathcal{N}(0,1)$.
- Asymptotic variance can be accurately estimated for causal forests.

2.3.4 Generalized Random Forest

The use of generalized random forest allow to reconcile both the case with significant confounding effects and the case with strong treatment heterogeneity. This model is built on local maximum likelihood (local generalized method of moments) [Athey et al., 2019]. It considers a similarity weighted set of neighbors, given by:

$$\alpha_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b_i}(\mathbf{x}) \text{ with } \alpha_{b_i}(\mathbf{x}) = \frac{\mathbb{1}(\mathbf{x}_i \in L_b(\mathbf{x}))}{|L_b(\mathbf{x})|} \tag{2.38}$$

where B is the number of trees of the forest and for each tree indexed by $b = 1..B$, $L_b(\mathbf{x})$ is the set of training individuals that fall in the same leaf as $\mathbf{x}$. The estimator is then calculated as:

$$\hat{\tau}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_{b_i}(\mathbf{x})\alpha_i(\mathbf{x})) \tag{2.39}$$

where $\hat{\tau}_{b_i}(\mathbf{x})$ is the estimate corresponding to the b tree. The authors prove that this estimate is consistent and asymptotically Gaussian and they provide confidence intervals.

2.3.5 Quasi-oracle estimation (R-learner)

The R-learner algorithm, introduced in [Nie and Wager, 2020], is a two steps algorithm. First, two quantities are estimated:

- The propensity score $p_s(\mathbf{x}) = \mathbb{P}(T = 1 \mid X = \mathbf{x})$;
- The conditional mean outcome (marginal effects) $\mu(\mathbf{x}) = \mathbb{E}(Y \mid X = \mathbf{x})$.

This two estimators $\hat{p}$ and $\hat{\mu}$ are introduced in an objective function, which capture treatment effects:

$$\hat{\mathcal{L}}(\tau(.)) = \frac{1}{n}\sum_{i=1}^{n}[(y_{i,obs} - \hat{\mu}(\mathbf{x}_i)) - (t_i - \hat{p}_s(\mathbf{x}_i))\tau(\mathbf{x}_i)]^2 \tag{2.40}$$

The treatment effect is estimated by minimising this modified loss function. A regularised term on the complexity of the $\tau(.)$ function can be added. Any methods can be used for this estimation: penalized regression, kernel ridge regression, boosting, to cite a few.

The first step can be seen as the learning of the oracle objective and the second as its estimation. The performance of this model is demonstrated in practice, especially when it uses deep neural networks. It also has the advantage of providing a "quasi-oracle" error bound for non-parametric regression problems.

2.4 Models evaluation

Since the "true" causal effect and the "true" counterfactual outcomes cannot be accessed, the evaluation task is challenging. In this section, we present datasets (Section 2.4.1) and metrics (Section 2.4.2) commonly used in counterfactual estimation to assess the effectiveness of the models and compare them to each other.

2.4.1 Artificial, semi-artificial and real-life datasets

The causal effect is estimated with both factual outcome $\mathbf{y}_{obs}$, corresponding to the assigned treatment $\mathbf{t}$, and the counterfactual outcomes $(\mathbf{y}_0, \mathbf{y}_1, ..\mathbf{y}_L) \setminus \mathbf{y}_{obs}$, corresponding to the hypothetical situation of unassigned treatments. The problem is that, in real-life, it is impossible to know the counterfactual outcomes, and then to check their accuracy. In order to assess the effectiveness of counterfactual estimation, synthetic datasets can be used. They are generated according to given distributions or by reproducing certain properties of the real data. As their generation is fully controlled, the exact distribution of the data is usually known. The "true" causal effect can be obtained by generating both factual and counterfactual outcomes.

However, models need to be tested in real life, where the distribution of data is not governed by mathematical laws or defined transformations. In real life, two types of datasets can be used: semi-synthetic datasets and purely real datasets. Semi-synthetic datasets come from real-life experiments or observations that have been artificially transformed. The covariates and the factual outcome are derived from real data, while the counterfactual outcomes and the treatment assignment can be artificially generated or created from real data. Purely real datasets are based on real experiments or observations that have not been transformed. In these datasets, only the factual outcome is known, the counterfactual outcomes are not available. Below we describe the most commonly used datasets in the models presented in the state-of-the-art. These datasets are also used to compare the models proposed in this thesis.

IHDP The Infant Health and Development Program (IHDP) are randomized experiments that began in 1985 in the US to reduce the developmental and health problems of low birth weight premature infants. The aim is to study the impact of specialist visits on the cognitive development of children. It is a semi-synthetic dataset, compiled by [Hill, 2011] for causal effect estimation. It is available in open source and used in many works on estimation of the counterfactual outcome [Hill, 2011, Shalit et al., 2017, Schwab et al., 2018, Yao et al., 2018]. It contains 747 instances with 25 covariates (6 continuous, 19 binary) measuring children characteristics like birth weight and head circumference and some features related to their mothers. The control and treatment groups are created according to the specialist visits. Children who received specialist visits are in the treatment group and the others in the control group. The treatment and control groups are artificially imbalanced by removing a biased subset of the treated population (608 control, 139 treated instances).

Twins The semi-synthetic Twins dataset has been released in [Almond et al., 2005] and formatted in [Louizos et al., 2017]. It studies the impact of the weight at birth of a baby on its survival in the USA between 1989 and 1991. Only twins which are the same sex and with a weight less than 2kg, are selected to this study. The dataset contains 49 features (21 categorical, 28 continuous) related to their mothers as if they smoked cigarettes, drank alcohol or took drugs. The sibling having the bigger weight is considered as being treated. There are 32120 instances with a balance between treated and untreated. The outcome variable corresponds to the mortality of each of the twins in their first year of life. The real distribution of each causal population is unknown, but the potential outcomes is available.

Criteo Criteo is a "Large Scale Benchmark for Uplift Modeling" [Diemert et al., 2018], constructed by Criteo AI Lab from several incrementality tests. The dataset collects information on consumer behaviour after being exposed or not to an advertisement. It contains 25M units with 12 continuous anonymized features, a treatment indicator and 2 binary outcomes (visits and conversions). The treatment group is composed of 84.6% of individuals and control group of $15,4\%$. This dataset is considered to be purely real, because the counterfactual outcome is not available.

Email Email is an another purely real-life dataset. The effect of an email marketing campaign is assessed in [Hillstrom, 2008]. The dataset contains 64000 customers divided in two groups: one receives an email from a marketing campaign (67%), the treatment group, and the other does not (34%), the control group. After two weeks, customers behaviors i.e. visiting the website or not, used as outcome, were tracked according to the customers features (6 categorical, 1 continuous). Only the outcome with or without treatment is known.

Other datasets are available and described in the survey [Yao et al., 2020], such as ACIC datasets used for the Atlantic Causal Inference Conference and datasets of the National Supported Work (NSW) [Dehejia and Wahba, 1999].

2.4.2 Metrics for counterfactual and ITE evaluation

A way to evaluate the accuracy of a model is to used the *expected means squared error*, defined as the expectation of error between counterfactual outcomes and their estimators:

$$\epsilon_{MSE} = \mathbb{E}_{\mathbf{X}\sim\mu_{\mathbf{X}}}[||\mathbb{E}_{\mathbf{y}\sim\mu_Y(\mathbf{X})}[\mathbf{y}] - \mathbb{E}_{\hat{\mathbf{y}}}[\hat{\mathbf{y}}]||_2^2] \tag{2.41}$$

where $||.||_2$ denotes the standard l_2-norm. This metric has the advantage of being immediately applicable in multi-treatment, but it is unreliable in terms of the potential bias among treatment groups.

Precision in estimation of heterogeneous effects (PEHE) is the most commonly used metric for evaluating the performance of counterfactual estimation. It is the mean squared error between the true effect and the predicted effect, where the effect is calculated as the difference of outcomes

between the treatment and control group [Hill, 2011]:

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[y_{i,1} - y_{i,0} \mid \mathbf{x}_i] - [\hat{y}_{i,1} - \hat{y}_{i,0}])^2. \tag{2.42}$$

A low ϵ_{PEHE} value means a strong prediction of the model.

For multiple k-treatments, PEHE can be expended by the average between every pair of treatments [Schwab et al., 2018]:

$$\hat{\epsilon}_{PEHE} = \frac{1}{\binom{k}{2}}\sum_{t_1=0}^{k-1}\sum_{t_2=0}^{i-1}\sum_{i=1}^{n}(\mathbb{E}[y_{i,t_1} - y_{i,t_2} \mid \mathbf{x}_i] - [\hat{y}_{i,t_1} - \hat{y}_{i,t_2}])^2 \tag{2.43}$$

For purely real-life datasets, the counterfactual is unknown and this quantity cannot be computed. Furthermore, if a model does not estimate the counterfactual outcome but directly compute the ITE, like in the direct model in [Jaskowski and Jaroszewicz, 2012a], this metric is not suitable.

Area Under the Uplift Curve (AUUC) is a metric used in binary treatment for measuring the ranking evaluation of the ITE. It can be used with unknown counterfactuals. It is calculated as the area under the *uplift curve*, defined as the difference between the lift curve of the treatment group and the lift curve of the control group. The *lift curve* is the proportion of positive outcomes ordered by their ITE value, as a function of the percentage of selected individuals [Diemert et al., 2018]. For a perfect classifier, when individuals are correctly ordered according to their ITE value, responders are the first individuals to be classified. As their ITE is equal to 1 ($\forall i \in \{1, .., n\}$, $y_{i,1} - y_{i,0} = 1 - 0$), the AUUC is strictly increasing. Then, a mixture of doomed and survivor individuals should be selected. The curve is thus constant since they have an ITE equal to 0 ($\forall i \in \{1, .., n\}$, $y_{i,1} - y_{i,0} = 1 - 1$ or $y_{i,1} - y_{i,0} = 0 - 0$). Finally, the curve is decreasing, since remaining antiresponders have an ITE equal to -1 ($\forall i \in \{1, .., n\}$, $y_{i,1} - y_{i,0} = 0 - 1$). The best performing method according to this criterion are the one with the largest AUUC. The construction and the evolution is illustrated in Figure 2.5.

AUUC can be interpreted as the net gain in success rate provided that a given percentage of the population is treated according to the model. The uplift curve being very similar to a ROC curve, AUUC has similar properties to the AUC. However, unlike the ROC curve, the uplift curve is not monotone increasing since it decreases, when the fraction of the population being treated correspond to anti-reponders (which have an ITE equal to -1), and is not bounded between 0 and 1, since it reaches a maximum equal to the number of responders.

We would expect the AUUC to be maximal when the ranking of individuals is perfect and then the best performing model should be the one with the highest AUUC value. However, since the uplift curve is built on the difference between the two lift curves (respectively on the treatment and control groups), in practice this is not systematically true. For example, when on the lift curves, a survivor in the treatment group is aligned with a doomed in the control group, the difference will be equal to 1 even though this individual does not represent a responder.

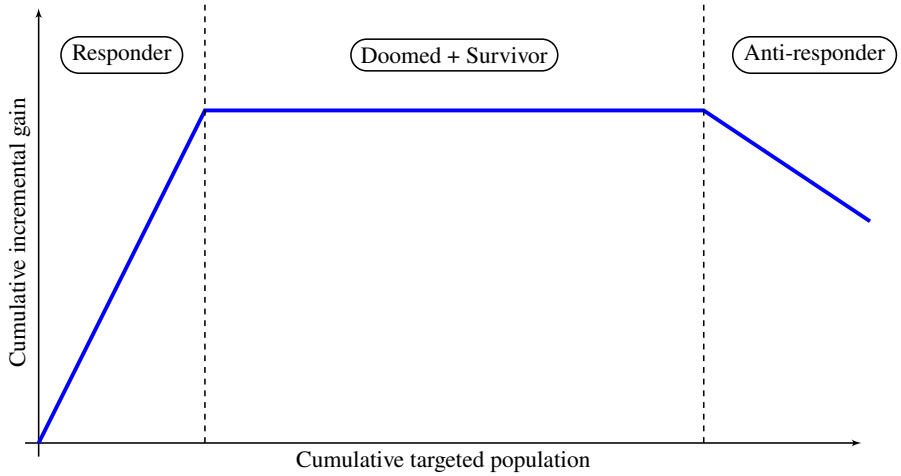


Figure 2.5: Optimal classifier - Uplift Curve

On synthetic datasets, uplift curves of predicted models are compared with the optimal classifier curve calculated using the exact data distribution. The ITE values, used for the optimal classifier, are continuous between -1 and 1 because individuals do not strictly belong to a causal group but has a probability of belonging to a group. On semi-synthetic data, the exact distribution is unknown but causal groups classification is available. Models are thus compared with the baseline constructed from the known classification. ITE values are discrete and included in $\{-1, 0, 1\}$. AUUC metric should be considered with caution because when a model predicts more positive outcomes than existing, AUUC value may be higher than the AUUC value of the optimal classifier.

The available information depends on the nature of the data considered (synthetic, semi-synthetic or purely real). The metrics used to estimate the efficiency of the models must be adapted accordingly. MSE and PEHE results are commonly used for synthetic data, while only AUUC can be used for real data. Some metrics assess the accuracy of the estimate of the counterfactual outcome, while others rather assess the ranking of individuals with respect to the ITE.

2.5 Positioning regarding the state of the art

The aim of this thesis is to contribute to the development of new methods to estimate the ITE. Our contributions are followings.

Estimation of the causal populations distributions as a hidden distribution problem

In this thesis, we aim to estimate the distribution of causal populations as a hidden distribution problem. This approach is similar to works in [Imbens and Rubin, 1997]. The authors introduce four populations based on compliance behavior: compliers, never-takers, always-takers and defiers. These populations are defined with the treatment assignment $t_i \in \{0, 1\}$ and the received

treatment $D_i(t)$ for an assignment treatment t, given an individual $i \in \{1, .., n\}$. They put the problem as a Bayesian inference and estimated the intention to treat $ITT = y_i(1, D_i(1)) - y_i(0, D_i(0))$ by using Expectation-Maximization and Data Augmentation algorithms. Although the causal population (R,D,S,A) have been mentioned in the literature [Wasserman, 2013], no work deals with there direct distribution estimation, to our knowledge. We propose two methods to solve this hidden space problem. The first propose a parametric model using an Expectation-Maximization algorithm, like in [Imbens and Rubin, 1997]. The second proposed method uses an Auto-Encoder that estimate the distribution of hidden variables, while benefiting from the efficiency of neural networks. This idea has been previously used to estimate the confounding effect. The authors in [Louizos et al., 2017] propose a Causal Effect Variational Autoencoders (CEVAE). It uses deep latent-variable models for estimating the causal effect inference. This approach differs from our model because it is in a similar framework to Pearl's work. The objective is not to estimate the causal effect as the difference between potential outcomes, but to estimate directly the effects of the confounding variables.

Beyong the estimation of the ITE

Our approach tackles a more general problem than estimating the counterfactual outcome and the individual treatment effect. We estimate the probability distribution of the causal populations, which allow to immediately deduced the counterfactual outcome and the CATE. Unlike the indirect methods presented in Section 2.1 and 2.2, we do not estimate the counterfactual outcome. We do not estimate the ITE directly either, like in Section 2.3, since we use the estimated population distribution to estimate the ITE. Thus our approach can be seen as a third area of work. The efficiency of our methods is nevertheless compared to the state-of-the-art introduced in this chapter, on the one hand by the efficiency of the estimation of the counterfactual, and on the other hand by the efficiency of the estimation of the individual treatment effect. As our approach allows to estimate the probability of belonging to a causal population, the proposed model is as attractive as the models proposing a confidence interval. Finally, in contrast to many of the methods presented, our approach has the advantage of being easily extended in multi-treatments.

3

Parametric Approach Using an Iterative Expectation-Maximization Algorithm

Contents

In this chapter, we propose to estimate the probability distributions of causal populations (R,D,S,A detailed in Section 1.3.3) with a parametric model, named Expected-Causality-Maximization (ECM) algorithm. We assume the assumptions of strong ignorability, in order to have an unbiased estimator of the ITE, and no hidden or unmeasured confounders. Binary values are assumed for the treatment assignment and the outcomes. Recall that, we have an i.i.d sample $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, and their corresponding treatment assignment $\mathbf{t} = \{t_i\}_{i=1}^n$ and observed outcome $\mathbf{y}_{obs} = \{y_{i,obs}\}_{i=1}^n$, where $t_i \in \{0,1\}$ and $y_{i,obs} \in \{0,1\}$. For a given individual $i \in \{1,..,n\}$,

the outcome with and without treatment $y_{i,1}$ and $y_{i,0}$ cannot be simultaneously observed. Only the outcome corresponding to the treatment assignment $y_{i,obs}$ is available, the other is missing.

In Section 3.1, we introduce the parametric model that projects causal constraints into the latent space of causal populations. The problem is cast as learning the probability densities for the four causal populations. In Section 3.2, we detail the algorithm implemented with a Gaussian mixture in Section 3.3 and with a hybrid mixture composed of Gaussian and multinomial distributions in Section 3.4. In both cases, an experimental study on synthetic and real datasets is conducted to confirm the efficiency of the model. Finally, in Section 3.5, we discuss the limitations and suggest an adaptation with variational Bayesian methods.

This Chapter is based on a publication at the 28th European Symposium on Artificial Neural Networks (ESANN 2020) under the title "Estimating Individual Treatment Effects through Causal Populations Identification" [Beji et al., 2020]. The material and code are available on github at https://github.com/CelineBeji/ECM.

3.1 A parametric model for causal populations

The whole population is modeled by a mixture of responders (R), doomed (D), survivors (S) and anti-responders (A). The four mutually exclusive categories of reaction to the treatment are identified in Table 3.1. For each population $k \in \{R, D, S, A\}$, let $f_k(. \mid \theta_k)$ be the prior distributions and π_k the mixing probability. Our goal is to estimate the set of parameters $\theta = \{\theta_k\}_{k\in\{R,D,S,A\}}$ and $\pi = \{\pi_k\}_{k\in\{R,D,S,A\}}$, setting the covariates distribution, expressed as:

$$p(\mathbf{X} \mid \pi, \theta) = \sum_{k=\{R,D,S,A\}} \pi_k \, f_k(\mathbf{X} \mid \theta_k) \tag{3.1}$$

Causal populations	Potential outcome	ITE	Causal Effect
Responder	$y_{i,0} = 0$ and $y_{i,1} = 1$	1	yes
Doomed	$y_{i,0} = 0$ and $y_{i,1} = 0$	0	no
Survivor	$y_{i,0} = 1$ and $y_{i,1} = 1$	0	no
Anti-responder	$y_{i,0} = 1$ and $y_{i,1} = 0$	-1	yes

Table 3.1: Causal populations in binary treatment and outcomes.

3.1.1 Causal constraints on the latent space

To apply causal constraints, we introduce a latent (or hidden) space $\mathcal{Z}$ that represents the space of causal populations. We assume there exists a function $g : \mathcal{X} \to \mathcal{Z}$ that transform the covariates $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in latent variables $\mathbf{Z} = \{z_{ik}\}_{k\in\{R,D,S,A\}}\}_{i=1}^n$. We set $\gamma(z_{ik}) := \mathbb{P}(Z = z_{ik} \mid X = \mathbf{x}_i)$ the conditional probability of $(Z = z_{ik})$ given $(X = \mathbf{x}_i)$, where $X \sim P_X$ and $Z \sim P_Z$ are random

variables. $\gamma(z_{ik})$ represents the class probability of an individual $i \in \{1, .., n\}$ to belong to the causal population $k \in \{R, D, S, A\}$ (see Table 3.2). Using Bayes' theorem, it can be expressed as:

$$\gamma(z_{ik}) = \mathbb{P}(Z = z_{ik} \mid X = \mathbf{x}_i) \tag{3.2}$$

$$= \frac{\mathbb{P}(Z = z_{ik})\mathbb{P}(X = \mathbf{x}_i \mid Z = z_{ik})}{\sum_{j=\{R,D,S,A\}} \mathbb{P}(Z = z_{ij})\mathbb{P}(X = \mathbf{x}_i \mid Z = z_{ij})} \tag{3.3}$$

$$= \frac{\pi_k f_k(\mathbf{x}_i \mid \theta_k)}{\sum_{j=\{R,D,S,A\}} \pi_j f_j(\mathbf{x}_i \mid \theta_j)} \tag{3.4}$$

Causal populations	Responder	Doomed	Survivors	Anti-responders
Class probabilities	$\gamma(z_{iR})$	$\gamma(z_{iD})$	$\gamma(z_{iS})$	$\gamma(z_{iA})$

Table 3.2: Causal class probabilities of an individual $i \in \{1, .., n\}$.

The proposed model implies to enforce several constraints on the latent space, hence excluding two causal populations according to the observed factual outcome and the assigned treatment. From the given treatment assignment $t_i \in \{0, 1\}$ and factual outcome $y_{i,obs} \in \{0, 1\}$, a partial information on the latent variable $\{\gamma(z_{ik})\}_{k\in\{R,D,S,A\}}$ can be inferred. For example, if an individual $i \in \{1, .., n\}$ is untreated ($t_i = 0$) and observed outcome is equal to zero ($y_{i,obs} = 0$), the individual cannot be by definition a survivor or an anti-responder. The probability distribution of these two populations can be forced to zero, i.e. $\gamma(z_{iS}) = \gamma(z_{iA}) = 0$, and the two other can be normalized, i.e. $\gamma(z_{iR}) + \gamma(z_{iD}) = 1$. Similar constraints occur for every value of the treatment assignment and factual outcomes. These constraints are the cornerstone of our approach. They are summarized in Table 3.3. In the same vein, initial constraints C_0^* can be added by initializing the probabilities of the two non-excluded populations to one half (see Table 3.4).

$(t_i = 0\ ,\ y_{i,obs} = 0)$	$(t_i = 0\ ,\ y_{i,obs} = 1)$	$(t_i = 1\ ,\ y_{i,obs} = 0)$	$(t_i = 1\ ,\ y_{i,obs} = 1)$
$\begin{cases}\gamma(z_{iS}) = \gamma(z_{iA}) = 0\\ \gamma(z_{iR}) + \gamma(z_{iD}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iR}) = \gamma(z_{iD}) = 0\\ \gamma(z_{iS}) + \gamma(z_{iA}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iR}) = \gamma(z_{iS}) = 0\\ \gamma(z_{iD}) + \gamma(z_{iA}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iD}) = \gamma(z_{iA}) = 0\\ \gamma(z_{iR}) + \gamma(z_{iS}) = 1\end{cases}$

Table 3.3: The causal constraints (C^*)

$(t_i = 0\ ,\ y_{i,obs} = 0)$	$(t_i = 0\ ,\ y_{i,obs} = 1)$	$(t_i = 1\ ,\ y_{i,obs} = 0)$	$(t_i = 1\ ,\ y_{i,obs} = 1)$
$\begin{cases}\gamma(z_{iS})^0 = \gamma(z_{iA})^0 = 0\\ \gamma(z_{iR})^0 = \gamma(z_{iD})^0 = \frac{1}{2}\end{cases}$	$\begin{cases}\gamma(z_{iR})^0 = \gamma(z_{iD})^0 = 0\\ \gamma(z_{iS})^0 = \gamma(z_{iA})^0 = \frac{1}{2}\end{cases}$	$\begin{cases}\gamma(z_{iR})^0 = \gamma(z_{iS})^0 = 0\\ \gamma(z_{iD})^0 = \gamma(z_{iA})^0 = \frac{1}{2}\end{cases}$	$\begin{cases}\gamma(z_{iD})^0 = \gamma(z_{iA})^0 = 0\\ \gamma(z_{iR})^0 = \gamma(z_{iS})^0 = \frac{1}{2}\end{cases}$

Table 3.4: The initial causal constraints (C_0^*)

The model is illustrated by the DAG in Figure 3.1, where each node is a random variable and the edges represent statistical dependencies between the variables. Following the usual convention, grey variables are observed while white variables are hidden (latent or missing). The dependence

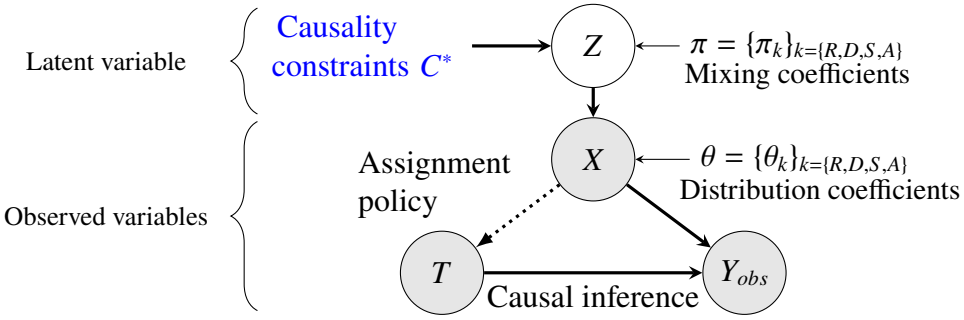


Figure 3.1: DAG model for causal inference. Observed (resp. latent) nodes are in grey (resp. white). X, Z, T and Y_{obs} are random variables corresponding respectively to covariates, latent variables, treatment assignment and observed outcome.

between node X and the treatment T is due to the treatment assignment policy. In a random assignment, the two variables are independent.

According to the graphical model in Figure 3.1, the conditional distribution of the i.i.d. latent variables $\{\mathbf{z}_i\}_{i=1}^n$, given the mixing coefficients $\pi = \{\pi_k\}_{k\in\{R,D,S,A\}}$ is written as:

$$p(\mathbf{z}_1, \cdots, \mathbf{z}_n \mid \pi) = \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} \pi_k^{z_{ik}} \tag{3.5}$$

and the conditional distribution of $\{\mathbf{x}_i\}_{i=1}^n$, given the latent variables $\{\mathbf{z}_i\}_{i=1}^n$ and distribution parameters $\theta = \{\theta_k\}_{k\in\{R,D,S,A\}}$ is expressed as:

$$p(\mathbf{x}_1, \cdots, \mathbf{x}_n \mid \mathbf{z}_1, \cdots, \mathbf{z}_n, \theta) = \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} f_k(\mathbf{x}_i \mid \theta_k)^{z_{ik}}. \tag{3.6}$$

Hence, the complete log-likelihood is written as:

$$\log p(\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{z}_1, \cdots, \mathbf{z}_n \mid \pi, \theta) = \log \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid \pi, \theta) \tag{3.7}$$

$$= \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \mathbf{z}_i, \pi, \theta)\, q(\mathbf{z}_i \mid \pi, \theta) \tag{3.8}$$

$$\tag{3.9}$$

According to the graphical model in Figure 3.1 given the dependence between variables, we have:

$$q(\mathbf{z}_i \mid \pi, \theta) = q(\mathbf{z}_i \mid \pi) \tag{3.10}$$

$$p(\mathbf{x}_i \mid \mathbf{z}_i, \pi, \theta) = p(\mathbf{x}_i \mid \mathbf{z}_i, \theta) \tag{3.11}$$

Hence, the complete log-likelihood can be expressed as:

$$\log p(\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{z}_1, \cdots, \mathbf{z}_n \mid \pi, \theta) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \mathbf{z}_i, \theta)\, q(\mathbf{z}_i \mid \pi) \tag{3.12}$$

$$= \sum_{i=1}^{n} \log p(\mathbf{x}_i \mid \mathbf{z}_i, \theta) + \log q(\mathbf{z}_i \mid \pi) \tag{3.13}$$

$$= \sum_{i=1}^{n} \log \prod_{k=\{R,D,S,A\}} f_k(\mathbf{x}_i \mid \theta_k)^{z_{ik}} + \log \prod_{k=\{R,D,S,A\}} \pi_k{}^{z_{ik}} \tag{3.14}$$

$$= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} z_{ik} \log f_k(\mathbf{x}_i \mid \theta_k) + \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} z_{ik} \log \pi_k \tag{3.15}$$

Remark: The above log-likelihood is called *complete log-likelihood* with reference to the *complete dataset* $(\mathbf{X}, \mathbf{Z})$. The observed variable $\mathbf{X}$ is denoted as *incomplete* without the corresponding value of the latent variable $\mathbf{Z}$.

The parameters of each causal populations (R,D,S,A) are estimated by maximizing the complete log-likelihood given by Eq. 3.15. The causal population class is arbitrarily assigned to the one with the highest distribution probability.

3.1.2 ITE estimation

Once the learning problem is solved, the ITE can be estimated with the probability distribution of the causal groups, as it is proved in Proposition 3.1.

Proposition 3.1. Knowing the latent distribution, the ITE can be computed as:

$$\tau(\mathbf{x}_i) = (\gamma(z_{iR}) + \gamma(z_{iS}))\mathbb{E}[\mathbb{1}_{y_{i,1}=1}] - (\gamma(z_{iS}) + \gamma(z_{iA}))\mathbb{E}[\mathbb{1}_{y_{i,0}=1}] \tag{3.16}$$

where by definition $\gamma(z_{ik}) = \frac{\pi_k f_k(\mathbf{x}_i|\theta_k)}{\sum_{l=\{R,D,S,A\}} \pi_l f_k(\mathbf{x}_i|\theta_l)}$ for $k \in \{R, D, S, A\}$.

Proof. The ITE estimated with the expected difference between the two potential outcomes:

$$\tau(\mathbf{x}_i) = \mathbb{E}[y_{i,1} \mid \mathbf{x}_i] - \mathbb{E}[y_{i,0} \mid \mathbf{x}_i] \tag{3.17}$$

The first term can be expressed by:

$$\mathbb{E}[y_{i,1} \mid \mathbf{x}_i] = \frac{\mathbb{E}[\mathbf{x}_i \mid y_{i,1} = 1]\, \mathbb{P}(y_{i,1} = 1)}{p(\mathbf{x}_i)} \tag{3.18}$$

$$= \frac{\mathbb{E}[\mathbf{x}_i \mid \mathbf{x}_i \in \{R, S\}]\, \mathbb{P}(y_{i,1} = 1)}{p(\mathbf{x}_i)} \tag{3.19}$$

$$= \frac{\sum_{k=\{R,S\}} \pi_k f_k(\mathbf{x}_i \mid \theta_k)\, \mathbb{P}(y_{i,1} = 1)}{\sum_{k=\{R,D,S,A\}} \pi_k f_k(\mathbf{x}_i \mid \theta_k)} \tag{3.20}$$

and similarly the second term can be written as:

$$\mathbb{E}[y_{i,0} \mid \mathbf{x}_i] = \frac{\sum_{k=\{S,A\}} \pi_k f_k(\mathbf{x}_i \mid \theta_k)\, \mathbb{P}(y_{i,0} = 1)}{\sum_{k=\{R,D,S,A\}} \pi_k f_k(\mathbf{x}_i \mid \theta_k)}. \tag{3.21}$$

The combination of Eq. (3.20) and (3.21) concludes the proof. □

Remark: When the probability to have an outcome equal to 1 in the treatment group is similar to the probability to have an outcome equal to 1 in the control group i.e. $\mathbb{P}(y_{i,1} = 1) = \mathbb{P}(y_{i,0} = 1)$, the individual treatment effect only depends on the pdf of the responders and the anti-responders: $\tau(\mathbf{x}_i) = \gamma(z_{iR})\, \mathbb{E}[\mathbb{1}_{y_{i,1}=1}] - \gamma(z_{iS})\, \mathbb{E}[\mathbb{1}_{y_{i,0}=1}]$.

3.2 Expectation-Causality-Maximization (ECM) algorithm

Our learning problem aims to estimate the mixing coefficients $\pi = \{\pi_k\}_{k\in\{R,D,S,A\}}$ and the distributions parameters $\theta = \{\theta_k\}_{k\in\{R,D,S,A\}}$ conditionally to the observed data $(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs})$. The Expectation-Maximization (EM) algorithm, originally introduced in [Dempster et al., 1977], is known to be an appropriate optimization algorithm for estimating the data distribution in presence of missing or hidden data, such as a mixture estimation problem. We adapt this algorithm by introducing the causal constraints defined in Section 3.1.1.

The EM algorithm is a popular tool for the estimation of statistical distributions with incomplete data such as in mixture estimation [Bishop, 2006b, McLachlan and Peel, 2004]. It is an iterative algorithm which aims to find the maximum likelihood parameters, by maximising the log-likelihood.

Among all the possible distributions parametrized by (θ, π), we find the set of values giving the distribution from which the data have been most likely sampled from. The algorithm performs two alternate optimisation steps until convergence: The Expectation (E) step and the Maximization (M) step. In the first step, the posterior distribution of the latent variables $q(\mathbf{Z} \mid \mathbf{X}, \theta, \pi)$ is estimated given the current estimation of the parameters. In the second step, the expectation of the complete log-likelihood is evaluated given the latent distribution considered as known (estimated in the E-step).

We propose an algorithm inspired of EM algorithm with extra partial information about the two unfeasible causal groups. The concept of authorized label set and partial information are not new and already defined in [Ambroise and Govaert, 2000, Côme et al., 2009]. The algorithm is enhanced with the causal constraints C^* derived from the causal structure of the problem. The latent distribution is not only estimated from the distribution of covariates $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ but also from the treatment assignment $\mathbf{t} = \{\mathbf{t}_i\}_{i=1}^n$ and the observed outcome $\mathbf{y}_{obs} = \{\mathbf{y}_{i,obs}\}_{i=1}^n$.

The proposed algorithm 1 works in three iterative steps: Expectation-Causality-Maximization (ECM). For each individual $i \in \{1, .., n\}$, the E-step estimates the probabilities of belonging to each causal groups $\{\gamma(z_{ik})\}_{k=\{R,D,S,A\}}$ according to the parameters of the four distributions

$(\theta_k, \pi_k)_{k=\{R,D,S,A\}}$. The C-step enforces the causal constraints C^* (see Table 3.3) on the latent space by forcing to zero the probabilities $\gamma(z_{ik})$ of the two unfeasible groups and normalizing the two remaining probabilities. This step can be considered as a post-processing of the E-step as we adjust the estimated probabilities $\gamma(z_{ik})$. The M-step maximises the complete log-likelihood $\mathcal{L}(\theta, \pi) = \log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta, \pi)$ to update the parameters of the distributions. For a faster convergence, the initial values of the latent variables $\gamma(z_{ik})^0$ are initialized with the C_0^* constraints that put a half probability on each of the two remaining causal populations.

Remark: Note that without information on the covariates $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, the distribution probabilities $\{\gamma(z_{ik})\}_{i=1}^n$ are uniformly distributed between the two authorized groups.

Algorithm 1: Expectation-Causality-Maximisation (ECM) algorithm

Initialisation: Initialisation of $\gamma(z_k)^0$ with the initial causal constraints C_0^* (Table 3.4) and computation of π_0 and θ_0 accordingly to M-step.

While not converged: (iterate on m)

Expected (E) step: Evaluate $q(\mathbf{Z} \mid \mathbf{X}, \theta^{m-1}, \pi^{m-1})$.

Causality (C) step: Apply the causal constraints C^* on the posterior probabilities of the latent variables $\gamma(z_k)$ according to Table 3.3.

Maximization (M) step: $(\theta^m, \pi^m) = \arg\max_{\theta,\pi}(\mathbb{E}[\log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta^{m-1}, \pi^{m-1})])$.

Check for convergence of the complete log-likelihood $\log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta^m, \pi^m)$.

End While

3.3 Implementation for a gaussian mixture model

A standard distribution is a mixture of Gaussians, defined by their parameters of mean and covariance matrix, denoted $\theta = \{\mu_k, \Sigma_k\}_{\{k=R,D,S,A\}}$. In this section, the ECM algorithm is applied by modelling each causal group (R,D,S,A) with a Gaussian distribution.

3.3.1 Algorithm

We assume that each causal population $k \in \{R, D, S, A\}$ follows a normal distribution $\mathcal{N}(\mu_k, \Sigma_k)$ and has a probability π_k in the whole set. The complete log-likelihood is expressed as:

$$\begin{aligned}
\mathcal{L}_{GM}(\pi, \mu, \Sigma) &= \log p(\mathbf{X}, \mathbf{Z} \mid \pi, \mu, \Sigma) && (3.22)\\
&= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \log \left(\pi_k \mathcal{N}(\mathbf{x}_i, \mu_k, \Sigma_k)\right)^{z_{ik}} && (3.23)\\
&= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \left(z_{ik} \log \left(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_k|}} \right) - z_{ik} \frac{(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)}{2} \right) && (3.24)
\end{aligned}$$

Maximizing the value of the complete log-likelihood is equivalent to set its partial derivatives to zero, and the estimators are given by (see proof in Appendix C):

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{\gamma}(z_{ik}) \mathbf{x}_i}{\sum_{i=1}^{n} \hat{\gamma}(z_{ik})} \tag{3.25}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} \hat{\gamma}(z_{ik})(\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T}{\sum_{i=1}^{n} \hat{\gamma}(z_{ik})} \tag{3.26}$$

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}(z_{ik}) \tag{3.27}$$

The parameters $\hat{\mu}_k$ can be interpreted as a prototypical individual of each population $k \in \{R, D, S, A\}$. They are expressed as the mean of features, weighted by the probability of belonging to one of the populations. $\hat{\Sigma}_k$ is expressed as the covariance of the features weighted by the probability of belonging to the population. Finally, the estimators $\hat{\pi}_k$ are the overall probability of observing an individual that comes from the population $k \in \{R, D, S, A\}$.

The posterior probabilities (or responsibilities) for an individual $i \in \{1, .., n\}$ to belong to the causal group k are given by:

$$\hat{\gamma}(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=\{R,D,S,A\}} \pi_l \mathcal{N}(\mathbf{x}_i | \hat{\mu}_l, \hat{\Sigma}_l)} \tag{3.28}$$

The ECM iterates by maximizing the complete log-likelihood as described in Algorithm 2.

3.3.2 Experimental Setting

The efficiency of the ECM algorithm is evaluated on synthetic and real datasets. The algorithm is compared to standard baselines according to several metrics.

Algorithm 2: ECM algorithm for a mixture of Gaussians

Initialisation: For all individual $i \in \{1, .., n\}$, set $\gamma(z_k)^0$ with the initial causal constraints:

$$\begin{cases} \text{if } (t_i = 0\ ,\ y_{i,obs} = 0), \text{ then } \gamma(z_{iS})^0 = \gamma(z_{iA})^0 = 0 \text{ and } \gamma(z_{iR})^0 = \gamma(z_{iD})^0 = \frac{1}{2} \\ \text{if } (t_i = 0\ ,\ y_{i,obs} = 1), \text{ then } \gamma(z_{iR})^0 = \gamma(z_{iD})^0 = 0 \text{ and } \gamma(z_{iS})^0 = \gamma(z_{iA})^0 = \frac{1}{2} \\ \text{if } (t_i = 1\ ,\ y_{i,obs} = 0), \text{ then } \gamma(z_{iR})^0 = \gamma(z_{iS})^0 = 0 \text{ and } \gamma(z_{iD})^0 = \gamma(z_{iA})^0 = \frac{1}{2} \\ \text{if } (t_i = 1\ ,\ y_{i,obs} = 1), \text{ then } \gamma(z_{iD})^0 = \gamma(z_{iA})^0 = 0 \text{ and } \gamma(z_{iR})^0 = \gamma(z_{iS})^0 = \frac{1}{2} \end{cases}$$

and compute $\hat{\mu}_0$, $\hat{\Sigma}_0$ and $\hat{\pi}_0$ with Eq. 3.25, 3.26 and 3.27.

While (Not Converged) do

Expectation step: Update probability distribution of latent variables. For all individual $i \in \{1, .., n\}$ and $k \in \{R, D, S, A\}$, $\hat{\gamma}(z_{ik})$ with Eq. 3.28

Causality step: Apply causal constraints for all individual $i \in \{0, n\}$:

$$\begin{cases} \text{if } (t_i = 0\ ,\ y_{i,obs} = 0), \text{ then } \gamma(z_{iS}) = \gamma(z_{iA}) = 0 \text{ and } \gamma(z_{iR}) + \gamma(z_{iD}) = 1 \\ \text{if } (t_i = 0\ ,\ y_{i,obs} = 1), \text{ then } \gamma(z_{iR}) = \gamma(z_{iD}) = 0 \text{ and } \gamma(z_{iS}) + \gamma(z_{iA}) = 1 \\ \text{if } (t_i = 1\ ,\ y_{i,obs} = 0), \text{ then } \gamma(z_{iR}) = \gamma(z_{iS}) = 0 \text{ and } \gamma(z_{iD}) + \gamma(z_{iA}) = 1 \\ \text{if } (t_i = 1\ ,\ y_{i,obs} = 1), \text{ then } \gamma(z_{iD}) = \gamma(z_{iA}) = 0 \text{ and } \gamma(z_{iR}) + \gamma(z_{iS}) = 1 \end{cases}$$

Maximization step: Update parameters of distributions for all $k \in \{R, D, S, A\}$, $\hat{\mu}_k$,$\hat{\Sigma}_k$ and $\hat{\pi}_k$ with Eq.3.25, 3.26 and 3.27.

Check for the convergence of the expected complete log-likelihood in Eq.3.24.

First, four synthetical datasets are built with known distributions. They are composed with two dimensional covariates. Each causal group (R,D,S,A) is designed with a multivariate Gaussian distribution and the number of individuals is balanced between each group. We consider four different scenarios:

- $synthetic_0$: Basic case of four distinct and linear separable distributions (Figure 3.2a).
- $synthetic_1$: Unlike the basic case, the group of anti-responder is closer to the group of responders. The distribution is therefore more difficult to isolate (Figure 3.3a).
- $synthetic_2$: The survivor and responder groups are interchanged, making the problem challenging in term of causal inference (Figure 3.4a).
- $synthetic_3$: Overlapping of the four distributions (Figure 3.5a).

In Figures 3.2a, 3.3a, 3.4a and 3.5a, the blue points, (respectively the purple, green and red ones), are the individuals belonging to the groups of responders (respectively survivors, doomed and anti-responders). Half of the samples is assigned to the treatment group and the other half to the control group. In these figures, the treatment and control groups are respectively distinguished by the crosses and the bullet points. The numerical values of means and variances are summarised in Table 3.5. The simulation is conducted with a sample of one thousand observed data ($n = 1000$).

	Responder	Doomed	Survivor	Anti-responder
$synthetic_0$	$\mu_R = (-10, 10)$ $\Sigma_R = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_D = (-10, -10)$ $\Sigma_D = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_S = (10, 10)$ $\Sigma_S = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_A = (10, -10)$ $\Sigma_A = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$
$synthetic_1$	$\mu_R = (-2, 10)$ $\Sigma_R = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_D = (-2, -10)$ $\Sigma_D = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_S = (5, 10)$ $\Sigma_S = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_A = (0, 0)$ $\Sigma_A = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$
$synthetic_2$	$\mu_R = (10, 10)$ $\Sigma_R = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_D = (-10, -10)$ $\Sigma_D = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_S = (-10, 10)$ $\Sigma_S = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$	$\mu_A = (10, -10)$ $\Sigma_A = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$
$synthetic_3$	$\mu_R = (0, 0)$ $\Sigma_R = \begin{pmatrix} 1 & 3 \\ 0 & 100 \end{pmatrix}$	$\mu_D = (-1, 1)$ $\Sigma_D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_S = (1, 2)$ $\Sigma_S = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$	$\mu_A = (4, 0)$ $\Sigma_A = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$

Table 3.5: Mean and variance values of four built synthetic datasets with two dimensional vector features distributed as a mixture of Gaussian distributions.

Then the semi-synthetic dataset IHDP, described in Section 2.4.1, is used to benchmark the proposed algorithm. It has been especially created for causal effect estimation. The underlying distribution of each causal population remains unknown, but the outcomes with and without treatment are both available by construction. The real causal group is determined with the values of the potential outcome.

The proposed algorithm is compared to simple and standard baselines that provide competitive results with respect to the state of the art [Alaa and Schaar, 2018]. Three models are used: the approach using two separate classification models (T-learner), the approach using the treatment variable as feature (S-learner)) and the model based on the class variable transformation [Jaskowski and Jaroszewicz, 2012b] (Direct approach), each method using as base classifier

a logistic regression. Note that the direct approach model is designed to estimate the causal effect and it is unable to predict the counterfactual. In addition to these baselines, a reference method is built. For synthetic datasets, it is based on the ground truth as the true distribution of the data is known. This means that it considers the probability of belonging to the causal groups for each individual and infers their causal group (with the highest probability) and their counterfactual outcome. When the data distribution is simpler, the difference between the true causal group (defined by construction) and the causal group affected by their probability distribution is insignificant. However, the more overlap between the causal groups there is, the greater the error between the actual values and the reference value is. This reference provides a view of the limits that a model can achieve. On semi-synthetic datasets, the ground truth is unknown and such a reference cannot be built. The true values of the counterfactual outcome are therefore used as reference.

The results are reported out of a sample over 20 trials and a Wilcoxon signed-rank test (with a level of 5%) is used to confirm the significance of the results. ϵ_{PEHE} and AUUC metrics, described in Section 2.4.1, are used to evaluate the effectiveness of the ECM algorithm. Although the AUUC is a metric that is subject to criticism, it does provide information on the estimated ITE ranking of individuals. We therefore use it as a complement to the ϵ_{PEHE}, which only provide the accuracy of the counterfactual prediction. A low ϵ_{PEHE} value means a good estimation of the counterfactual because it indicates that in average the predicted counterfactual outcome corresponds to the true value. This metric, which requires knowledge of the true counterfactual, can only be used on synthetic or semi-synthetic data but not in real dataset (see Section 2.4.1). The AUUC evaluates the ranking of the ITE. It is the area under the uplift curve that classifies individuals according to their ITE values. In theory, it should be increasing (due to the responders), then constant (due to a mix of doomed and survivors) and then decreasing (due to the anti-responders). An effective model would maximise the value of AUUC.

For these experiments, we introduce our own metric named *Causal Accuracy Rank* (CAR). This ranking metric inspired of the AUUC, is designed to assess the classification of each causal group. It evaluates the ranking of two individuals only if they are not in the same causal group. For example, two responders misclassified with respect to each other will not have any impact, while a responder misclassified with respect to a doomed will be penalized. We compare two by two all individuals who belong to different groups. If the difference of the predicted ITE of two individual has the same sign that the difference of the real ITE, then the two individuals are well ranked against each other. For example, let $\mathbf{x}_1$ an individual belonging to the responder group and $\mathbf{x}_2$ an individual belonging to the doomed group. The reel ITE of $\mathbf{x}_1$, respectively $\mathbf{x}_2$, is equal to 1 respectively 0. The difference is positive (equal to 1). Suppose that the predicted ITE of $\mathbf{x}_1$ is equal to 0.5 and the predicted ITE of $\mathbf{x}_2$ is equal to 0.4. According to these values, they are well ordered. The difference is positive and has the same sign that the difference of the real ITE. Now if the predicted ITE of $\mathbf{x}_1$ is equal to 0.4 and the predicted ITE of $\mathbf{x}_2$ is equal to 0.5, these two individuals are misclassified. The ITE value difference is negative and opposite to the real ITE sign. In this case, this ranking will be considered as wrong. Thus, the ranking accuracy is

expressed as:

$$CAR = \frac{1}{N} \sum_{(\mathbf{x},\mathbf{x}')\in\mathcal{X}^2 | G(\mathbf{x})\neq G(\mathbf{x}')} \mathbb{1}_{sign[\tau(\mathbf{x})-\tau(\mathbf{x}')=sign[\widehat{\tau}(\mathbf{x})-\widehat{\tau}(\mathbf{x}')]} \tag{3.29}$$

where $G(\mathbf{x})$ is the causal population assignment of the individual $\mathbf{x}$. Remark: The resulting value of this metric is between 0 and 1. A value close to 1 denotes a suitable classification, while a value close to 0 means that the model is not efficient.

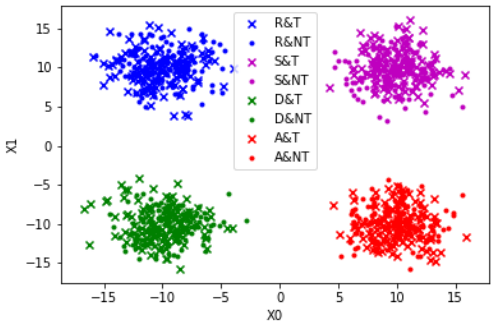
The numerical results are summarized in Table 3.6 in which the values in bold correspond to the best model among the benchmark models.

In synthetic datasets, the 2-dimension of the covariates allows to visualize the ITE on a heatmap (see Figures 3.2, 3.3, 3.4 and 3.5). This provides to analyse the distribution of the ITE according to the data distribution. The areas in red represent regions with a high proportion of individuals with a ITE close to 1, and therefore abundant in responders. In blue areas, the proportion of individual close to -1 is preponderant and therefore the regions are composed with a large number of anti-responders. The areas in white is made of individuals with ITE close to 0, i.e. doomed and survivors.

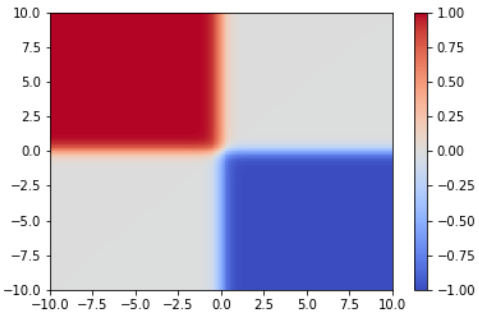
3.3.3 Results and conclusion

Figures 3.2, 3.3, 3.4 and 3.5 represent the data and the ITE distributions according to the two covariate axes $(X^{(0)}, X^{(1)})$. On the first basic case of distinct and separable distributions ($synthetic_0$), the four areas corresponding to responders, doomed, survivors and anti-responders are clearly distinct on the real ITE-heatmap. These four regions are identified by the two classifiers (T-learner) despite less distinct borders. The S-learner classifier, which divides the space linearly into two parts, accumulates the anti-responders on the inner diagonal and the responders on both sides. The direct approach model identifies the anti-responders on the upper diagonal and the responders on the lower diagonal. The ECM has the ITE-heatmap closest to the reference. This is also true in the second synthetic dataset ($synthetic_1$), the ECM is still the model closest to the reference. The S-learner and direct approach models are ITE-heatmap similar to the previous case. The T-learner model has a dissimilar shape compared to the reference, although it is closer than the other two baselines. The third data distribution ($synthetic_2$) is not linearly separable in terms of ITE, i.e. the responders (with an ITE equal to 1) and the anti-responders (with an ITE equal to -1) are not separated linearly by the doomed and the survivors (with an ITE equal to 0). For this dataset, the three models S-learner, T-learner and the direct direct approach perform poorly while the ECM ITE-heatmap remains close to the reference. Finally, when the data are overlapping ($synthetic_3$), the ECM is clearly the only model to estimate a valid distribution. The other models continue to identify areas of space according to groups without taking into account the overlap.

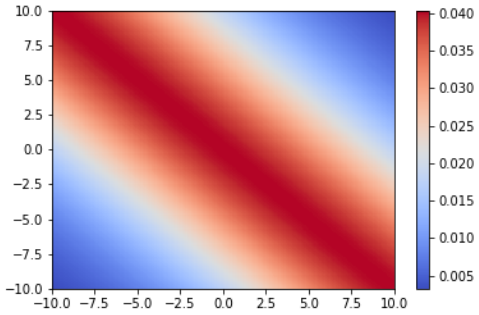
The numerical results in Table 3.6 confirm the conclusions obtained with the ITE-heatmap figures. In the first synthetic dataset, the ϵ_{PEHE} of the S-learner and ECM models are equal to zero. The counterfactual is therefore correctly estimated. The AUUC and CAR values proves



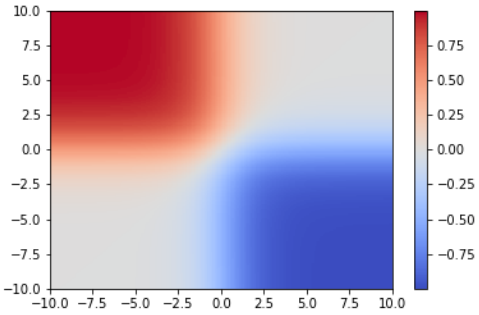
(a) Data distribution



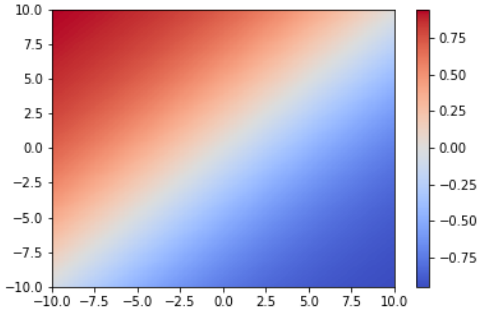
(b) Reference heatmap



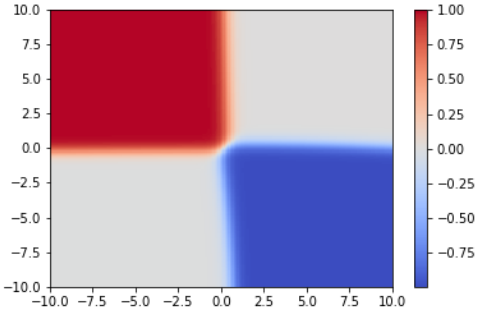
(c) S-learner heatmap



(d) T-learner heatmap

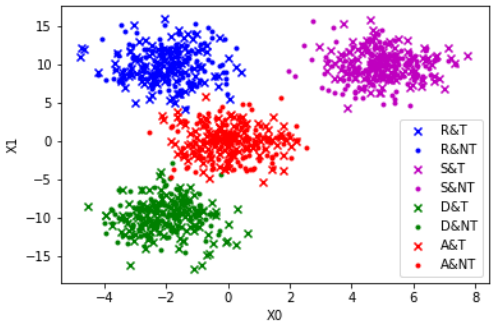


(e) Direct approach heatmap

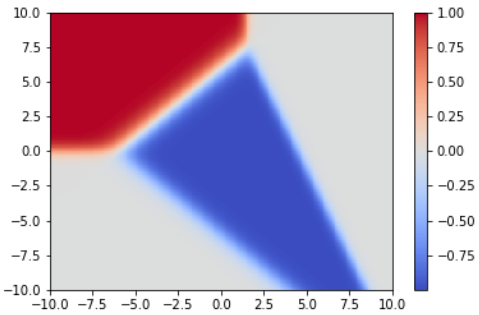


(f) ECM heatmap

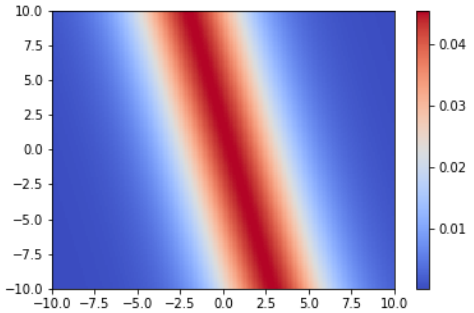
Figure 3.2: Data distribution and ITE-heatmaps results to $synthetic_0$ dataset.



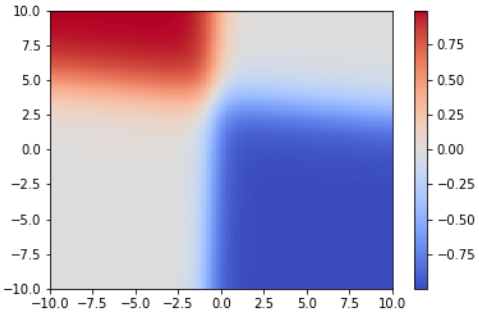
(a) Data distribution



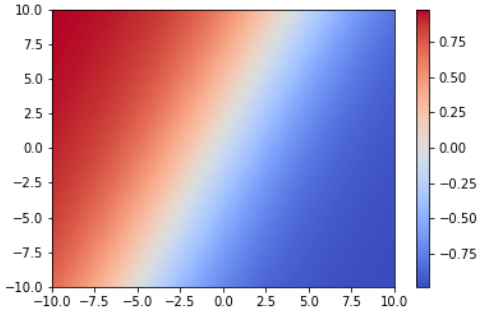
(b) Reference heatmap



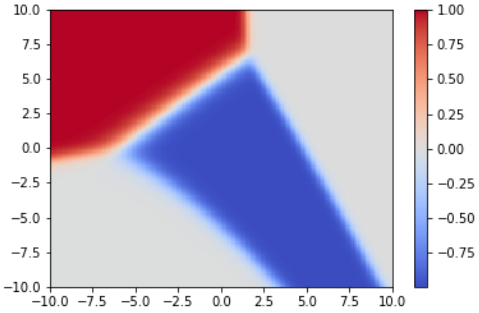
(c) S-learner heatmap



(d) T-learner heatmap

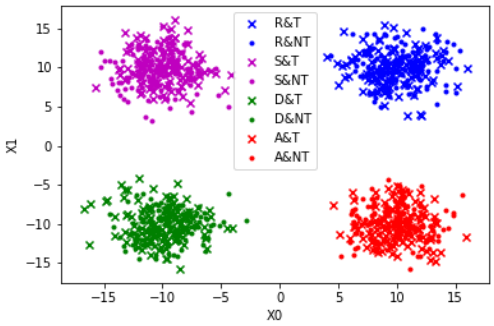


(e) Direct approach heatmap

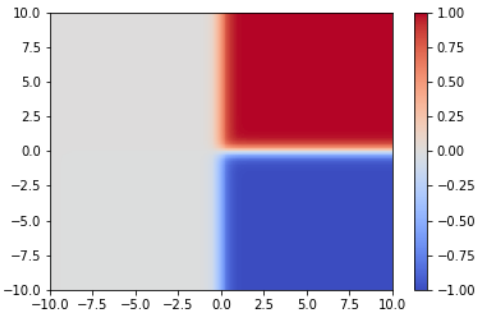


(f) ECM heatmap

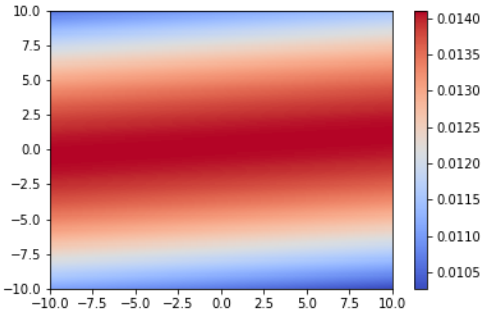
Figure 3.3: Data distribution and ITE-heatmaps results to $synthetic_1$ dataset.



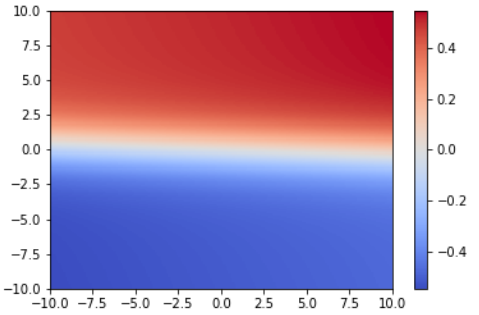
(a) Data distribution



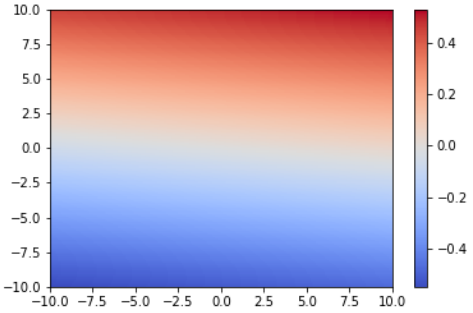
(b) Reference heatmap



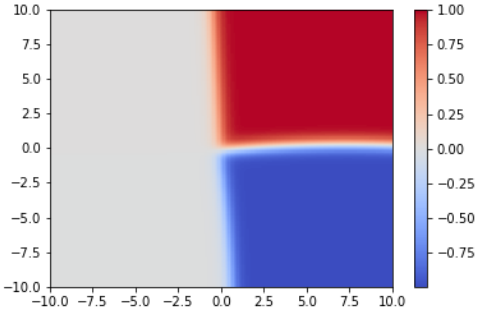
(c) S-learner heatmap



(d) T-learner heatmap

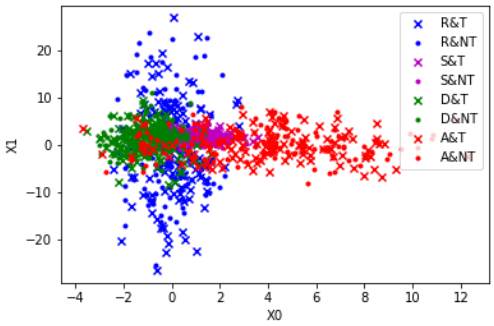


(e) Direct appraoch heatmap

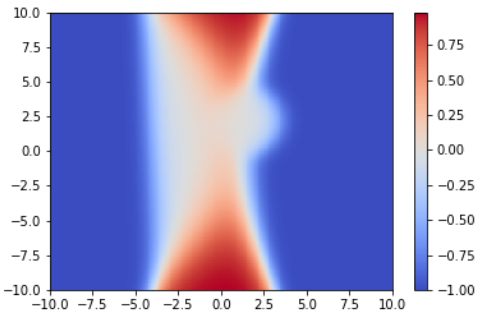


(f) ECM heatmap

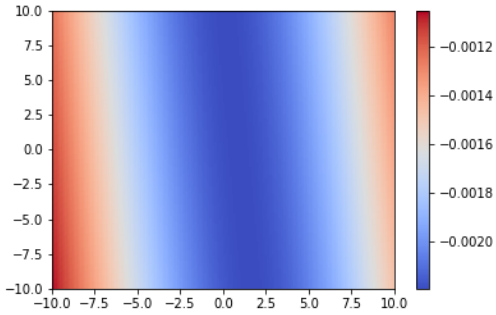
Figure 3.4: Data distribution and ITE-heatmaps results to $synthetic_2$ dataset.



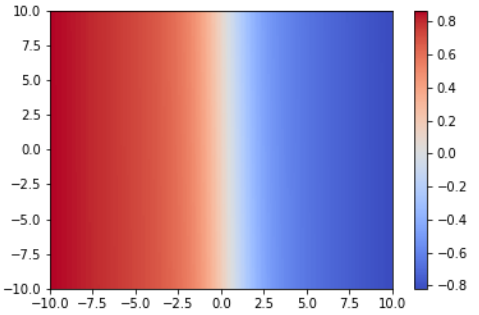
(a) Data distribution



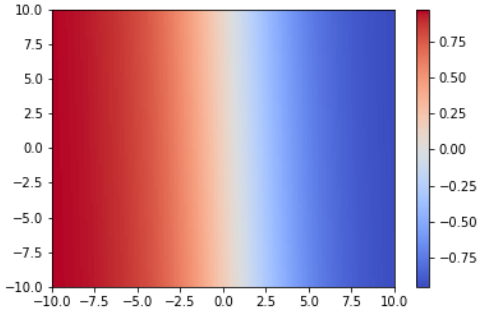
(b) Reference heatmap



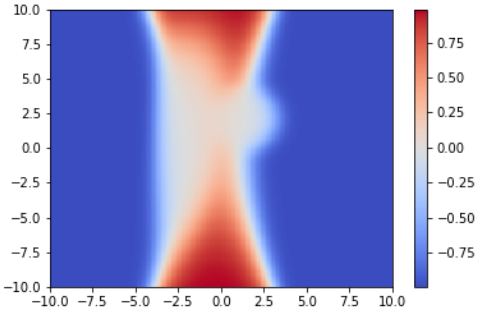
(c) S-learner heatmap



(d) T-learner heatmap



(e) Direct approach heatmap



(f) ECM heatmap

Figure 3.5: Data distribution and ITE-heatmaps results to $synthetic_3$ dataset.

	ϵ_{PEHE}	AUUC	CAR
synthetic$_0$			
Reference	0.00	1784	0.66
S-learner	0.53 +/- 0.05	283 +/- 153	0.29 +/- 0.02
T-learner	0.00 +/- 0.00	1819 +/- 228	0.63 +/- 0.02
Direct approach	.	1813 +/- 246	0.62 +/- 0.02
ECM	**0.00 +/- 0.00**	**1865 +/- 247**	**0.67 +/- 0.02**
synthetic$_1$			
Reference	0.02	1836	0.62
S-learner	0.51 +/- 0.05	255 +/- 142	0.30 +/- 0.02
T-learner	0.11 +/- 0.02	1814 +/- 317	0.60 +/- 0.01
Direct approach	.	1086 +/- 303	0.44 +/- 0.03
ECM	**0.02 +/- 0.01**	**1848 +/- 234**	**0.61 +/- 0.02**
synthetic$_2$			
Reference	0.00	1850	0.66
S-learner	0.49 +/- 0.03	440 +/- 215	0.33 +/- 0.06
T-learner	0.53 +/- 0.12	1317 +/- 238	0.50 +/- 0.02
Direct approach	.	1301 +/- 254	0.50 +/- 0.02
ECM	**0.00 +/- 0.00**	**1869 +/- 242**	**0.67 +/- 0.02**
synthetic$_3$			
Reference	0.24	1488	0.55
S-learner	0.57 +/- 0.08	742 +/- 175	0.24 +/- 0.10
T-learner	0.79 +/- 0.08	943 +/- 206	0.44 +/- 0.02
Direct approach	.	939 +/- 208	0.44 +/- 0.02
ECM	**0.27 +/- 0.04**	**1512 +/- 203**	**0.55 +/- 0.02**
IHDP			
Reference	0.00 +/- 0.00	3149	.
S-learner	0.66 +/- 0.08	2202 +/- 625	0.21 +/- 0.04
T-learner	0.67 +/- 0.07	2168 +/- 618	0.21 +/- 0.04
Direct approach	.	2191 +/- 558	0.22 +/- 0.04
ECM	**0.59 +/- 0.09**	2226 +/- 580	**0.24 +/- 0.04**

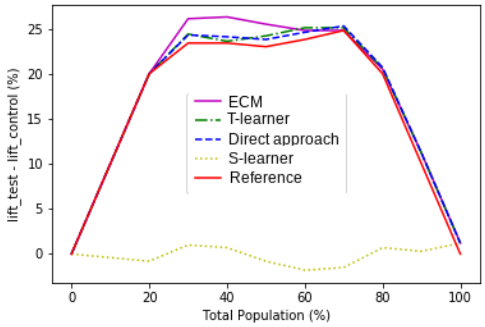
Table 3.6: Experimental results on synthetic and real datasets. Models are compared with a reference which is a optimal classifier for synthetic data and the real classification for real dataset. The values in bold correspond to the model with the highest efficiency compared to the considered models. Note that direct approach is designed to estimate the causal effect and it is unable to predict the counterfactual outcome, the ϵ_{PEHE} is then not computable for this model. As CAR metric compares the ITE predicted and the real ITE, it is then not available for the reference model of semi-synthetic datasets.

that the ECM performs most efficiently in term of causal group recognition, closely followed by the S-leaner and the direct approach. The S-leaner model shows a low efficiency with all the metrics used. In the second dataset, the performance gap is increased between ECM, S-learner and the direct approach. The ECM prediction remains close to reference values. For S-learner, the estimation according to the AUUC and CAR metrics remain also close to the reference but the prediction of the counterfactual (estimated by the ϵ_{PEHE}) is weaker. The results of the direct approach on this dataset show weak performances, even according to the AUUC and CAR metrics. On the third synthetic dataset, ECM is the only model with good results. The other models unperformed according to all metrics of interest. In the last, most complex case of overlap between causal groups, the ECM model becomes weaker in the performance of counterfactual prediction. However, it remains efficient according to AUUC and CAR measures. The other models still produce very poor results.

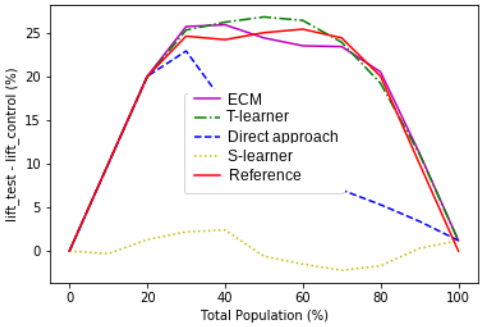
The uplifts curves, in Figure 3.6, provide more information. The flat curve relative to S-learner shows that the model does not identify the causal groups on any datasets, unlike the other models. On the simplest dataset, the S-learner, the direct approach and the ECM models have competitive results. On the second synthetic dataset, the direct approach model mixes the doomed, the survivors and the anti-responders, although it efficiently detects the responders. On the other two datasets, the S-learner and the direct approach models have the same behaviours. They do not separate doomed and survivors accurately, especially in the case of overlap where their ITE are above those of responders.

In conclusion, the results on the synthetic datasets show that the more complex the data distribution is, the more the ECM outperforms the baseline. It produces results closer to the reference model. Note that this experiment is favourable to the ECM model since the distribution of the generated data follows the same distribution as the underlying Gaussian assumption taken.

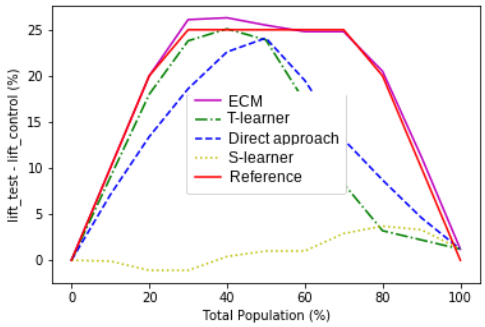
Since the IHDP dataset is not in 2-dimensional (two covariates), the ITE distribution cannot be observed via a heatmap. The numerical results in Table 3.6 are interesting to observe the accuracy of each models. The gap between the ECM and the other competitors is significant in terms of both ϵ_{PEHE} and CAR (using a 5% threshold for rejecting the null hypothesis) but not in term of AUUC. However, the data complexity produces disappointing results with all the models. It is more noticeable on the uplift curve (see Figure 3.6e). Unlike synthetic cases, this dataset has a significant imbalance in the number of individuals in the four causal populations and in the treatment and control groups. This explains why the uplift curve is not increasing, constant and then decreasing as expected. Note that the uplift curve classifies individuals according to their ITE value but it is built from the two lift curves on the treatment and control group. The distribution of data explains the difficulty of efficiently predict the distribution of causal groups. In addition, this dataset has the particularity of containing a very large number of categorical data. In the following section, we investigate whether the ECM, which is a model suitable for continuous data, could be adapted to the more common categorical data.



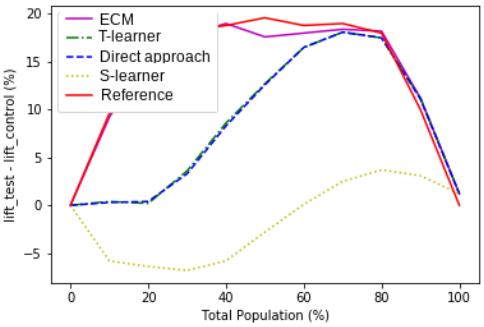
(a) *synthetic*$_0$



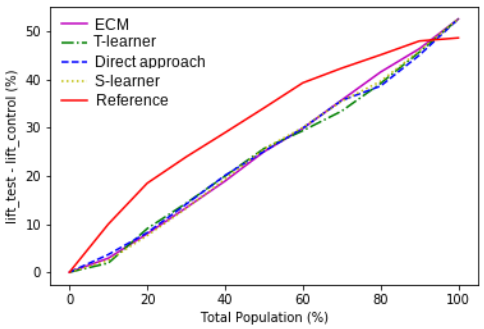
(b) *synthetic*$_1$



(c) *synthetic*$_2$



(d) *synthetic*$_3$



(e) IHDP

Figure 3.6: Uplift curves for synthetic and real datasets. They are built as the difference of the lift curves on the treatment and control groups. The reference uses the real distribution of data to estimate the ITE on synthetic datasets, and the classification of the causal groups on IHDP.

3.4 Implementation for a hybrid gaussian and independent multinomial mixture model

This section focus on data mixing discrete and continuous variables, that is more common in real life than pure continuous or discrete dataset. The importance of mixed data have recently emerged in the causal literature. For example, the authors in [Li and Shimizu, 2018] develop a mixed model where the continuous variables are modeled as a linear function of its parent variables plus a non-Gaussian noise. In [Marx and Vreeken, 2018], two scores are proposed, one for a single type distribution and an other for mixed type distribution. Based on the Minimum Description Length (MDL), the method uses classification and regression tree to model the causal dependency. In the same vein of these works, we introduce an approach that formulates the causal inference as an inference for a hybrid mixture of gaussian and multinomial covariates.

3.4.1 Algorithm

The ECM algorithm can be applied on a Gaussian mixture for continuous variables mixed with independent Multinomial distributions for each categorical variable. The covariates set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ can be decomposed in a set of $n_\mathcal{N}$ continuous variables $\mathbf{X}_\mathcal{N} = \{\mathbf{x}_{i\mathcal{N}}\}_{i=1}^n$ and a set of $n_\mathcal{M}$ categorical variables $\mathbf{X}_\mathcal{M} = \{\mathbf{x}_{i\mathcal{M}}\}_{i=1}^n$. For all $i \in \{1, .., n\}$, $\mathbf{x}_i = (\mathbf{x}_{i\mathcal{N}} \cup \mathbf{x}_{i\mathcal{M}}) \in \mathbb{R}^d$, where $d = n_\mathcal{N} + n_\mathcal{M}$. We note $\mathbf{X}_{\mathcal{M}_j}$ the j^{th} categorical variable with $j = \{1, .., n_\mathcal{M}\}$. We assume the continuous variables follow a normal distribution $\mathcal{N}(\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean and variance parameters. Each categorical variable $\mathbf{X}_{\mathcal{M}_j}$ follows an independent Multinomial distribution $\mathcal{M}(\rho_{jk})$, where n_{jk} is the number of categories ν of the variable $\mathbf{X}_{\mathcal{M}_j}$ and $\rho_{jk} = (\rho_{jk\nu})_{\nu=\{1,..,N_{jk}\}}$ is the vector given the probabilities of selecting category ν. Note that $\mathbf{x}_{\mathcal{M}_j,\nu}$ is binary with only one value equal to 1. We put $\rho = \{(\rho_{jk})_{j=\{1,..,M\}}\}_{k=\{R,D,S,A\}}$ the set of the Multinomials parameters. $\{\pi_k\}_{k=\{R,D,S,A\}}$ are the mixing parameters of the causal distributions where $\sum_{k=\{R,D,S,A\}} \pi_k = 1$.

The joint probability of covariates is given by:

$$p(\mathbf{X} \mid \mu, \Sigma, \rho, \pi) = \prod_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \pi_k \, \mathcal{N}(\mathbf{x}_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \, \mathcal{M}(\mathbf{x}_{i\mathcal{M}_1} \mid \rho_{1k}) \, \mathcal{M}(\mathbf{x}_{i\mathcal{M}_2} \mid \rho_{2k}) \cdots \mathcal{M}(\mathbf{x}_{i\mathcal{M}_M} \mid \rho_{Mk}) \tag{3.30}$$

$$= \prod_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \pi_k \, \mathcal{N}(\mathbf{x}_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \prod_{j=1}^{M} \mathcal{M}(\mathbf{x}_{i\mathcal{M}_j} \mid \rho_{jk}) \tag{3.31}$$

and the complete log-likelihood is expressed as:

$$\mathcal{L}_{MGM}(\mu, \Sigma, \rho, \pi) = \mathcal{L}_{GM}(\mu, \Sigma, \pi) + \sum_{i=1}^{n} \sum_{j=1}^{M} \sum_{\nu=1}^{n_{jk}} z_{ik} \mathbf{x}_{i\mathcal{M}_j,\nu} \log(\rho_{jk\nu}) + C \tag{3.32}$$

where $\mathcal{L}_{GM}(\mu, \Sigma, \pi)$ is defined in Eq. 3.24 and C is a constant term of μ, Σ, P and π.

The expected value of the complete log-likelihood is written as:

$$\mathbb{E}[\mathcal{L}_{MGM}(\pi,\mu,\Sigma)] = \mathbb{E}[\mathcal{L}_{GM}(\pi,\mu,\Sigma)] + \sum_{i=1}^{n}\sum_{j=1}^{M}\sum_{\nu=1}^{n_{jk}} \gamma(z_{ik})\mathbf{x}_{i\mathcal{M}_j,\nu}\log(\rho_{jk\nu}) \tag{3.33}$$

The parameters $\hat{\mu}_k$, $\hat{\Sigma}_k$ and $\hat{\pi}_k$ are respectively given in Eq. 3.25, 3.26 and 3.27.

The parameters of the Multinomials are estimated as (see proof in Appendix C):

$$\hat{\rho}_{jk\nu} = \frac{\sum_{i=1}^{n}\hat{\gamma}(z_{ik})\mathbf{x}_{i\mathcal{M}_j,\nu}}{n_{jk}\sum_{i=1}^{n}\hat{\gamma}(z_{ik})} \tag{3.34}$$

Finally, the posterior probabilities for an individual $i \in \{1,..,n\}$ to belong to the causal group k is given by:

$$\hat{\gamma}(z_{ik}) = \frac{\hat{\pi}_k\,\mathcal{N}(\mathbf{x}_{i\mathcal{N}} \mid \hat{\mu}_k,\hat{\Sigma}_k)\prod_{j=1}^{M}\prod_{\nu=1}^{n_{jk}}\mathcal{M}(\mathbf{x}_{i\mathcal{M}_j,\nu} \mid \hat{\rho}_{jk\nu})}{\sum_{l=\{R,D,S,A\}}\pi_l\,\mathcal{N}(\mathbf{x}_{i\mathcal{N}}|\hat{\mu}_l,\hat{\Sigma}_l)\prod_{j=1}^{M}\prod_{\nu=1}^{n_{jl}}\mathcal{M}(\mathbf{x}_{i\mathcal{M}_j,\nu} \mid \hat{\rho}_{jl\nu})} \tag{3.35}$$

3.4.2 Experimental Setting

We conduce an experimental study to evaluate the efficiency of the ECM algorithm with a hybrid distribution composed of Gaussian distributions for continuous data and independent Multinomial distributions for categorical data. In order to understand and explain the difference between the two proposed prior distributions (simple Gaussians or hybrid distribution), we use synthetic datasets. Samples of 1000 units are generated with a balance between the treated group and the untreated group. The causal groups are overloaped such that the posterior probabilities of belonging to a causal group do not just take the value 0 or 1, but are spread between $[0, 1]$. The three following dataset are used:

- A mixture of Gaussians. This dataset contains two continuous features distributed as a mixture of overlapping Gaussians. This is the $synthetic_3$ dataset used in previous experiments (see Section 3.3).
- A mixture of independent Multinomials. This second dataset contains four categorical features, with respectively 2, 3, 2 and 4 modalities.
- A hybrid mixture of Gaussians and independent Multinomials. This last datasets is built from the concatenation of the first two. It contains two continuous features, distributed as a Gaussian mixture, and four categorical features, distributed as an independent Multinomials mixture.

The numerical values are given in Table 3.7.

The ECM, using a hybrid distribution, is compared to the previous one, using a simple Gaussian mixture. In addition, we consider as baseline the model composed of with two separate logistic

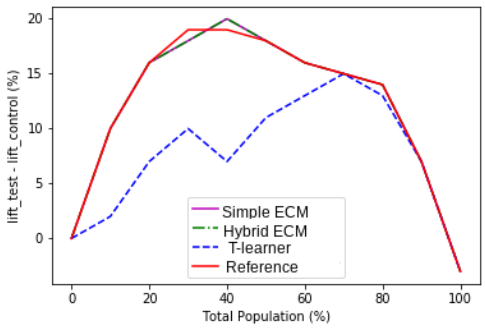
regression models (T-learner) that it is the one which demonstrated the best results of the baselines in the experiments in Section 3.3.3. This model builds one classifier for the treatment group and an other for the control group, and then estimated the counterfactual outcome for each individual accordingly. The experiments are conduced over 20 trials. The results, reported in Table 3.8, are out of sample and averaged over the trials. Two metrics are used: the percentage of misclassification of causal groups and the AUUC. A low value for the percentage of the classification error indicates an efficient prediction of the causal groups. This value is obtained by counting the number of individuals for which the predicted group is different from the real assigned group. The AUUC, which is calculated from the area under the uplift curve, is high when a valid estimate of ITE is obtained. In addition, we consider the shape of the uplift curve compared to the reference curve constructed from the true distribution of the data. More information about this metric are available in Section 2.4.1.

3.4.3 Results and conclusion

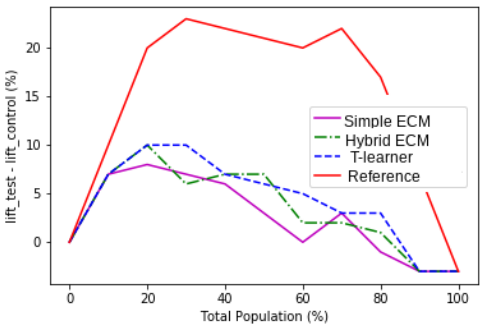
The numerical results of the experiments are reported in Table 3.8. The uplift curves are shown in Figure 3.7. For the dataset containing only continuous variables, the simple and hybrid ECM have equivalent results. This is explained by the definition of the two distributions that are similar in the absence of categorical data. These models accurately classify three quarters of the individuals. The T-learner model performs less efficiency with an error percentage of 55%. This AUUC result is also worse than the values of the ECM models. On the corresponding uplift curve (see Figure 3.7a), we observe that the responders, that could be sorted in the first individuals,

Responder	Doomed	Survivor	Anti-responder
Gaussian mixture for continuous variables			
$\mu_R = (0, 0)$	$\mu_D = (-1, 1)$	$\mu_S = (1, 2)$	$\mu_A = (4, 0)$
$\Sigma_R = \begin{pmatrix} 1 & 3 \\ 0 & 100 \end{pmatrix}$	$\Sigma_D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\Sigma_S = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$	$\Sigma_A = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$
Multinomials independents mixture for categorical variables			
$\rho_{1R} = (0.1, 0.9)$	$\rho_{1D} = (0.5, 0.5)$	$\rho_{1S} = (0.6, 0.4)$	$\rho_{1A} = (0.2, 0.8)$
$\rho_{2R} = (0.3, 0.3, 0.4)$	$\rho_{2D} = (0.7, 0.3, 0.0)$	$\rho_{2S} = (0.3, 0.4, 0.3)$	$\rho_{2A} = (0.4, 0.4, 0.2)$
$\rho_{3R} = (0.3, 0.7)$	$\rho_{3D} = (0.9, 0.1)$	$\rho_{3S} = (0.9, 0.1)$	$\rho_{3A} = (0.5, 0.5)$
$\rho_{4R} = (0.2, 0.2, 0.2, 0.4)$	$\rho_{4D} = (0.1, 0.8, 0.1, 0.0)$	$\rho_{4S} = (0.5, 0.2, 0.2, 0.1)$	$\rho_{4A} = (0.3, 0.3, 0.3, 0.1)$

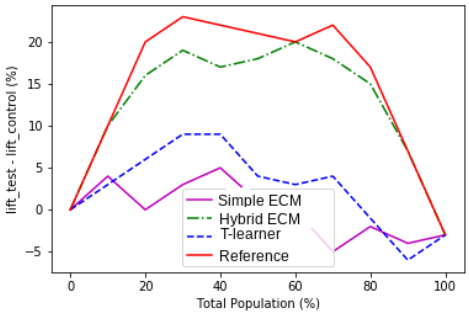
Table 3.7: Parameters of the generated Gaussians and independent Multinomials distributions.



(a) Mixture of Gaussians



(b) Mixture of independent Multinomials



(c) Hybrid mixture of Gaussians and independent Multinomilas

Figure 3.7: Uplift curves for synthetic datasets. They are built as the difference of the lift curves on the treatment and control groups. The reference uses the reel distribution of data to estimate the ITE.

are mixed with doomed and survivors by the T-learner model. The increase of the curve is then affects but the anti-responders are nevertheless correctly classified. On pure categorical dataset, the hybrid ECM has better performance in terms of percentage of classification error and AUUC. The gap is widening on the third dataset which mixes continuous and categorical variables. The hybrid ECM demonstrates efficient performances with only 16% of misclassification and an uplift curve close to reference uplift curve.

To conclude, a hybrid distribution is better suited to a dataset containing categorical data. Moreover, the closer the prior distribution is to the real distribution of the data, the better the performance of the ECM is. The real advantage benefit of ECM comes from the variety of the prior distributions that can be used. However it has limitations when the likelihood is not calculable. One solution is to use a variational adaptation of the algorithm, detailed in the next section.

3.5 Limits and variational extensions

The proposed model has limitations regarding the distribution of data. In real-life applications, the cost of calculating log-likelihood can be prohibitive because of the large scale of the latent

	Classification error	AUUC
Gaussian mixture		
T-learner	55% + / − 4%	1111 + / − 302
simple ECM	26% + / − 1%	1695 + / − 335
hybrid ECM	26% + / − 1%	1695 + / − 335
Independent Multinomials mixture		
T-learner	52% +/− 3%	657 +/− 283
simple ECM	54% +/− 6%	580 +/− 282
hybrid ECM	39% +/− 3%	715 +/− 273
Hybrid mixture		
T-learner	50% +/− 2%	807 +/− 362
simple ECM	54% +/− 5%	601 +/− 304
hybrid ECM	16% +/− 2%	1730 +/− 335

Table 3.8: Experimental results on synthetic datasets. The ECM applied to a simple Gaussians mixture (*simple ECM)*) is compared to the ECM applied to a hybrid mixture composed of Gaussians and independent Multinomials (*hybrid ECM)*). The baseline used is the T-learner model and the metrics evaluate the classification of the causal populations and the estimation of the ITE.

variables or the high complexity of the posterior distribution. In the proposed ECM algorithm, we need to evaluate the expectation of the complete log-likelihood with respect to the posterior distribution of the latent variables. Variational methods introduce a deterministic approximation based on the analytical estimation of the posterior distribution. They are thoroughly described in [Bishop, 2006b].

Given the Bayes rule and the dependence between the variables $p(\mathbf{X} \mid \theta, \pi) = p(\mathbf{X} \mid \theta)$ (see Figure 3.1), a lower bound of the log-likelihood can be expressed as:

$$\log p(\mathbf{X}, \mathbf{Z} \mid \theta, \pi) = \log p(\mathbf{Z} \mid \mathbf{X}, \theta, \pi) + \log p(\mathbf{X} \mid \theta) \tag{3.36}$$

The log marginal probability can be decomposed as:

$$\log p(\mathbf{X} \mid \theta) = \mathcal{F}\left(q(\mathbf{Z}), \theta\right) + KL(q(\mathbf{Z}) \| p(\mathbf{X} \mid \mathbf{Z}, \theta))) \tag{3.37}$$

where

$$\mathcal{F}(q(\mathbf{Z}), \theta) = \int q(Z) \log\left(\frac{p(\mathbf{X}, Z \mid \theta)}{q(Z)}\right) dZ \tag{3.38}$$

$$KL(q(\mathbf{Z}) \| p(\mathbf{X} \mid \mathbf{Z}, \theta)) = \int q(Z) \log\left(\frac{p(Z \mid \mathbf{X}, \theta)}{q(Z)}\right) dZ \tag{3.39}$$

The proposed approach does not calculate the log-likelihood but $\mathcal{F}(q(\mathbf{Z}), \theta)$, the Evidence Lower Bound (ELBO). The resulting variant is given in Algorithm 3.

Algorithm 3: Variational Expectation-Causality-Maximisation algorithm

Initialisation: Initialisation of q with the initial causal constraints.

While not converged:

E-step: Minimize $KL(q\|p)$ with (θ,π) fixed.

C-step: Apply the causal constraints C^* on q.

M-step: Maximise $\mathcal{F}(q)$ w.r.t with q fixed.

End While

In the vein of the application of Gaussians mixture, this algorithm can be applied on a variational Mixture of Gaussians [Bishop, 2006b]. We can introduce a Dirichlet distribution over the mixing parameters:

$$p(\pi) = Dir(\pi, \alpha_0) \tag{3.40}$$

and an independent Gaussian-Whishart prior governing the mean and the precision of the Gaussian distributions:

$$p(\mu, \Lambda) = p(\mu \mid \Lambda) p(\Lambda) = \prod_{k=\{R,D,S,A\}} \mathcal{N}(\mu_k \mid m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid W_0, \mu_0) \tag{3.41}$$

The choice of the Gaussian-Wishart distribution is not naive. It is one of the conjugate prior for a multivariate Gaussian. It is therefore trivial to calculate the posterior which is a Gaussian-Wishart distribution with different parameters. The posterior parameters can be found in [Gelman et al., 2004].

Like the proposed ECM algorithm, this variant determines the latent variables distribution with respect to the causal constraints, and the M-step finds the optimal solution using the maximization of the lower bound.

In contrast to the ECM, without the initial causal constraints, the algorithm does not correctly identify the obtained classes to the causal groups. Moreover, the stability of the variational algorithm is widely discussed in [Bishop, 2006a]. In addition, this algorithm could in practice suffer from problems due to the inversion of the precision matrix. An other issue of the proposed ECM algorithm is the parametric distribution that has to be chosen to implement the algorithm. Frequently, the data is not distributed according to any standard distribution but rather according to a distribution that is difficult to model. To address this issues, in the next chapter we introduce a non-parametric model applicable on more complex distributions.

4 Non-Parametric Approach Using an Auto-Encoder

Contents

In this chapter, we propose a new non-parametric approach to estimate the probability distributions of the causal populations (responder, doomed, survivors and anti-responders). We use an Auto-Encoder enhanced by a causal prior materialized by a mask and name this method the Causal Auto-Encoder (CAE). The mask introduces some partial information to the Auto-Encoder and is applied in the intermediate layer of our network. It helps with the estimation of the hidden latent representation of the observations, each dimension (or group of dimensions) being identified with the probability of belonging to a given causal population.

In Section 4.1, we present the architecture and the optimisation problem. Then, in Section 4.2, we discuss the architecture of the latent layer. The impact of the number of neurons and their possible roles according to the causal constraint is studied. An extensive numerical experiment is conducted in Section 4.3 where the proposed approach is compared to the state-of-the-art baselines on both synthetic and real-life datasets. Lastly, in Section 4.4, we discuss possible variants in the architecture of the auto-encoder to possibly boost the performance of the method.

This Chapter is based on publication at the Causal-ITALY workshop of the 20th International Conference Italian Association for Artificial Intelligence (AIxIA 2021). The material and code are available on github at https://github.com/CelineBeji.

4.1 Causal-Auto-Encoder (CAE)

Auto-encoders have long been part of the neural network environment [Rumelhart et al., 1986, Baldi and Hornik, 1989]. They were originally used for dimensionality reduction or feature learning in [Bourlard and Kamp, 1988, Hinton and Zemel, 1994]. They are feedforward neural networks which aim at reconstructing the input after compression into a lower-dimensional space and can be thought as non-linear extensions of the principal component analysis. They are composed of two parts: an encoder and a decoder. The encoder compresses the input into a hidden layer $h = f(\mathbf{x})$. The decoder reconstructs the input using the latent space as input $r = g(h)$. A loss function $l(\mathbf{x}, g(f(\mathbf{x})))$, which evaluates the error between the original input and the reconstructed input, is minimized. The model is trained via back-propagation.

We propose the Causal-Auto-Encoder (CAE) that estimates the probability distribution of the causal populations, by reconstructing the covariates. In addition to the standard structure, causal constraints are applied on the latent space. These constraints, which depend on the observed outcome and the assigned treatment, are aimed at capturing the probability distributions of each causal population.

4.1.1 Global architecture

Causal constraints, introduced in Section 3.1.1, are tailored for the case of the Auto-Encoder architecture. The constraints are specific to each individual $i \in \{1, .., n\}$ and depend on the observed outcome $y_{i,obs}$ and the assigned treatment t_i. We recall that, the causal constraints take the following form:

- If $t_i = 0$ and $y_{i,obs} = 0$, then the individual i is a responder or a doomed.
- If $t_i = 0$ and $y_{i,obs} = 1$, then the individual i is a survivor or an anti-responder.
- If $t_i = 1$ and $y_{i,obs} = 0$, then the individual i is a doomed or an anti-responder.
- If $t_i = 1$ and $y_{i,obs} = 1$, then the individual i is a responder or a survivor.

A prior information on the probabilities of the causal population can be introduced. For each

individual $i \in \{1, .., n\}$, given t_i and $y_{i,obs}$, two causal populations are by definition excluded. Their probability distribution can therefore be forced to zero.

The architecture is composed of three parts: an encoder, a mask built with the causal constraints and designed to set some units of the latent space to zero, and a decoder (see Figure 4.1 for an overview). The encoder takes as input the covariates $\mathbf{x}_i$, the i^{th} sample of the data, for $i \in \{1, .., n\}$. Each node of the initial layer represents a feature (a dimension of $\mathbf{x}_i$). After compressing $\mathbf{x}_i$ with a feedforward neural network, the latent variable $\mathbf{z}_i = \begin{pmatrix} z_{iR} & z_{iD} & z_{iS} & z_{iA} \end{pmatrix}$ is constrained by the mask $M(y_{i,obs}, t_i)$. Each node z_{ik} of the latent layer is assimilated to a causal population $k \in \{R, D, S, A\}$. z_{ik} is activated or not (put to zero) according to the mask designed from the observed outcome $y_{i,obs}$ and the assigned treatment t_i. For example, if $t_i = 0$ and $y_{i,obs} = 0$ the individual is by definition a survivor or an anti-responder. The designed mask is $M(y_{i,obs}, t_i)^T = \begin{pmatrix} 0 & 0 & 1 & 1 \end{pmatrix}$ where the first coordinate (respectively the second, the third and the fourth) corresponds to the probabilities distribution of responders (respectively doomed, survivors and anti-responders).

Remark: In Chapter 3, the considered latent variables $\{z_{ik}\}_{i=1}^n$ model the data distribution of causal populations $k \in \{R, D, S, A\}$ and the probability that a given individual $i \in \{1, ..n\}$ belongs to the causal population k is given by $\gamma(z_{ik})$. In this chapter, the considered latent space is directly the probability of belonging to a causal population. For the sake of convention, we note these variables z_{ik}, but it corresponds to $\gamma(z_{ik})$ of the previous chapter.

The constrained latent variable $\tilde{\mathbf{z}}_i$ is taken as input of the decoder. The decoder, built as a

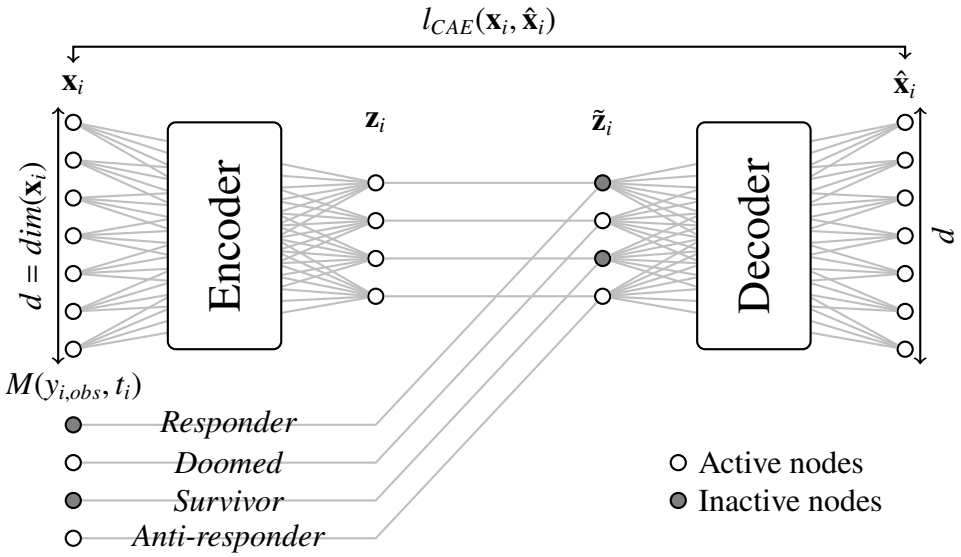


Figure 4.1: Overview of the Causal-Auto-Encoder (CAE) architecture. White nodes (respectively gray nodes) correspond to active neurons (respectively inactive neurons). $\mathbf{x}_i$ are the covariates and $\hat{\mathbf{x}}_i$ is its reconstruction by the Auto-Encoder. $\mathbf{z}_i$ is the latent variable constructed by the encoder. It is then constrained by the mask $M(y_{i,obs}, t_i)$ (some of this neurons are deactivate), from $y_{i,obs}$ the observed outcome and t_i the treatment assignment.

feedforward neural network, attempts to reconstruct the covariates. It outputs an estimated $\hat{\mathbf{x}}_i$ with the same size as the input covariates $\mathbf{x}_i$. The CAE is iteratively trained from a learning database $\{\mathbf{x}_i\}_{i=1}^{s}$ (where s is the number of sample used for learning), via back-propagation, minimizing the loss function $l_{CAE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$.

In the proposed CAE, the mask enforces some structured sparsity constraint on the hidden layer. An analogy can be seen with Sparse Auto-Encoders [Ng et al., 2011], for which some units are deactivate , or a dropout structure [Srivastava et al., 2014], which ignoring (drop out) a certain set of neurons chosen at random during the training. These structure are used to improve performance and reduce overfitting. The significant difference remains in the structure of the sparsity or dropout constraint enforced in the CAE. In the proposed architecture, the structure is known and can be directly enforced without relying on a regularization scheme.

4.1.2 Optimization

The m-multiple layers encoder is modelled by a deterministic function:

$$f_\theta(\mathbf{x}) = s_1(W_1 s_2(W_2 s_3(W_3 \cdots s_m(W_m \mathbf{x} + b_m) \cdots + b_3) + b_2) + b_1) \tag{4.1}$$

where $\theta = \{W_1, W_2, W_3, ..W_m, b_1, b_2, b_3, .., b_m\}$, $(W_1, W_2, W_3, .., W_m)$ are the weight parameters, $(b_1, b_2, b_3, .., b_m)$ the bias and $(s_1, s_2, s_3, .., s_m)$ non-linear functions. To associate the last layer of the encoder to the probabilities distribution of the causal populations, we use the *softmax(.)* activation function, defined as:

$$u(z_{ik}) = \frac{\exp(z_{ik})}{\sum_{j=1}^{p} \exp(z_{ij})} \tag{4.2}$$

where z_{ik} represents a unit of the last latent layer of the encoder and p the number of latent layer units. The hidden representation is then mapped to construct an estimation of the input $\hat{\mathbf{x}}$, using:

$$\hat{\mathbf{x}} = g_{\theta'}(\mathbf{z}) = s'_1(W'_1 s'_2(W'_2 s'_3(W'_3 \cdots s'_m(W'_m \mathbf{z} + b'_m) \cdots + b'_3) + b'_2) + b'_1) \tag{4.3}$$

where $\theta' = \{W'_1, W'_2, W'_3, ..W'_m, b'_1, b'_2, b'_3, .., b'_m\}$, are the parameters representing respectively of weight and bias and $(s'_1, s'_2, s'_3, .., s'_m)$ are non-linear functions. The parameters are optimized to minimize the loss function between the input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ over all training point.

This standard architecture is completed by a causal constraint phase, which is materialised by the application of a mask. After computing the hidden representation and rather than immediately reconstructing the input, some hidden units are set to zero. A mask $M(\mathbf{y}_{obs}, \mathbf{t})$ is designed from all of the samples and inserted in the architecture as an auxiliary input. The objective is to set specific units to zero in order to exclude incompatible populations, according to the prior information on the observed outcome and the assigned treatment. The hidden constrained representation is obtained by:

$$\tilde{\mathbf{z}} = M(y_{obs}, t)^T \mathbf{z} \tag{4.4}$$

where each of the elements of the vector $M(y_{obs}, t)$ correspond to a causal group. Assigning the first (respectively the second, the third and the fourth) element to the group of responders

(respectively doomed, survivors and anti-responders), the mask is given by:

$$M(y_{i,obs}, t_i) = \begin{pmatrix} \mathbb{1}_{t_i=y_{i,obs}} \\ 1 - y_{i,obs} \\ y_{i,obs} \\ \mathbb{1}_{t_i \neq y_{i,obs}} \end{pmatrix} \tag{4.5}$$

The set of parameters (θ, θ') are optimized jointly via back-propagation. The reconstruction error estimator for this model is expressed as:

$$l_{CAE}(\theta, \theta') = l(\mathbf{x}, g_{\theta'}(M(y_{obs}, t)^T f_\theta(\mathbf{x}))) \tag{4.6}$$

Using the standard mean square error, this optimisation problem boils down to minimize the following loss:

$$l(\mathbf{x}, \hat{\mathbf{x}}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \|g_{\theta'}(M(\mathbf{y}_{obs}, \mathbf{t})^T f_\theta(\mathbf{x})) - \mathbf{x}\|_2^2 \tag{4.7}$$

The procedure for the proposed approach is described in Algorithm 4.

Algorithm 4: Causal Auto-Encoder (CAE)

Main input: train set $\{\mathbf{x}_i\}_{i=1}^s$
Auxiliary input: train set $\{y_{i,obs}, t_i\}_{i=1}^s$

Initialisation: Initialisation of the parameters θ, θ'.

While not converged:

Feed-forward step: $\mathbf{z} = f_\theta(\mathbf{x})$.

Apply causal constraints: $\tilde{\mathbf{z}} = M(y_{obs}, t)^T \mathbf{z}$.

Compute parameters $\hat{\mathbf{x}} = g_{\theta'}(\tilde{\mathbf{z}})$

Back-propagation and check for convergence.

End While

4.1.3 Prediction

Once the model is trained, the encoder block of the model is used to encode the probability distribution of causal populations z_{ik} (corresponding to $\gamma(z_{ik})$ in the notations of Chapter 3) (see Figure 4.2). Each individual is assigned to the causal population, for which the predicted probability is the highest. Moreover, the higher the probability of belonging to a causal population is, the higher the confidence of the estimate is. Knowing the causal population predicted for a given individual, its assigned treatment and its observed outcome, the counterfactual outcome

can be directly estimated. As previously for the ECM (see Section 3.1.2), the ITE, defined as the difference between the expected outcome on the treatment group minus the expected outcome on the control group given the covariates, can also be estimated from the probability distribution of the causal population:

$$\tau(\mathbf{x}_i) = (z_{iR} + z_{iS})\mathbb{E}[\mathbb{1}_{y_{i,1}=1}] - (z_{iS} + z_{iA})\mathbb{E}[\mathbb{1}_{y_{i,0}=1}] \tag{4.8}$$

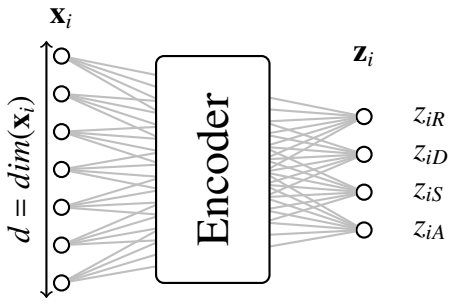


Figure 4.2: Prediction of the probability distribution of the causal populations. The encoder is set by its parameters estimated during the training. $\mathbf{x}_i$ are the covariates and the hidden variables $\mathbf{z}_{ik}$ represent the probability distribution of each causal populations $k \in \{R, D, S, A\}$.

4.2 Latent layer architecture

A prior depending on the assigned treatment and the observed outcome is integrated to the standard structure of an auto-encoder. Because of these constraints, the hidden variables are associated to the probability distribution of the causal populations. For the sake of simplicity, we have introduced this architecture with a single neuron to encode the probability of belonging to each causal group, however such a reduction in dimension may be too drastic, as it compresses the information from the input space to a constrained four dimensional space. In this section, we discuss about the architecture of the latent layer and the number of neurons used to encode a causal population and we propose strategies for enlarging the latent space.

4.2.1 Number of nodes encoding a causal population

To encode the four causal populations, we need to have at least four nodes, one for each population. However, we can also take more than one node per population. Intuitively, the more nodes there are, the more covariate information is preserved. The size of the mask depends on the number of nodes l allocated to encode a population. It constraints therefore all the nodes associated to a causal population (see Figure 4.3). The probability of belonging to a given population is calculated as the sum of the probabilities given by the nodes affected to this population.

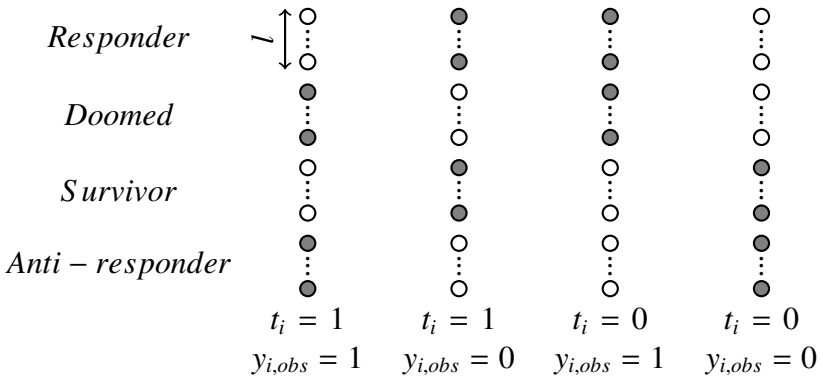


Figure 4.3: Latent space architecture when a causal population is encoded with l nodes. The causality constraint is applied for each sample i, according to the assigned treatment t_i and the observed outcome $y_{i,obs}$. White nodes (respectively grey nodes) are activated (respectively inactivated) by the mask.

4.2.2 Additional nodes

In addition to the nodes affected to each causal population, we can add unconstrained nodes. These nodes participate only to encode the information from the covariates, without being disturbed by the causal prior. Their number can be chosen as large as required, depending on the number of features. However, the number of additional nodes must be lower than the size of the covariates, so that the information is distributed both on the additional nodes and on the distribution probability nodes. The values of additional nodes are unaffected by the mask and they are not used in the estimation of the probability distributions of the causal populations. The resulting architecture is illustrated in Figure 4.4.

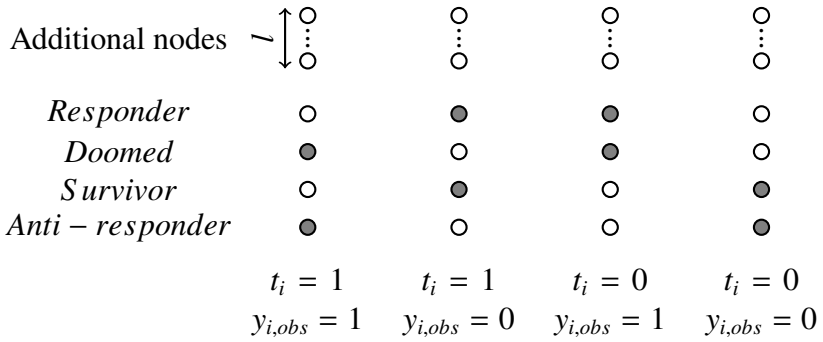


Figure 4.4: Latent space architecture when l unconstrained informative nodes are added. The causality constraints are only applied on the nodes corresponding to the causal populations. They are specific to each sample i and depend on the assigned treatment t_i and the observed outcome $y_{i,obs}$. White nodes (respectively grey nodes) are activated (respectively inactivated) by the mask.

Remark: In practice, the output of the encoder is divided into two outputs, one to encode the probabilities of the causal populations and the other to preserve the covariates information. Only

the first is subject to the action of the softmax activation function in the last layer. After the neurons encoding the probability distribution of the causal populations have passed through the mask, they are concatenated to the information neurons of the covariates to constitute the input of the decoder.

Note that it is possible to have both several nodes for encoding a population and information nodes. At this stage, there is no evidence that one architecture is preferable to another. We conduct in the next section an experimental analysis that compare the efficient of the estimation, according to various architectures.

4.3 Experiments

In this section, in order to prove the efficiency of the proposed CAE architecture, we conduct an empirical study on synthetic and real-life datasets and compared the model with the state-of-the-art baselines. The architecture and the size of the latent space is discussed on all datasets. As previously, metrics assess both the counterfactual and the ITE estimation.

4.3.1 Auto-Encoder setting

The encoder and the decoder are implemented as simple symmetric FeedForward networks. They are composed of four layers for which the dimension is halved or doubled at each layer (see Figures 4.5 and 4.6). The size of the encoder input and the decoder output are given by the dimension of covariates. The dimension of the latent representation $\mathbf{z}$ depends on the architecture of the latent layer. In this experiment, four different architectures are chosen: CAE_1, CAE_2, CAE_5 and CAE_{info5}. The first (respectively the second and the third) used one (respectively two and five) nodes to encode the probability of each causal population. No additional unconstrained nodes is added in these models. The last architecture CAE_{info5} is composed of one node to encode a causal population and five additional unconstrained nodes to capture the information of the features.

Each layer is followed by Batch Normalization, in order to stabilize the learning and to reduce the number of epochs required to train the model [Ioffe and Szegedy, 2015]. By default, the activation function Rectified linear unit (ReLU), which returns the input value if it receives a positive input and zero otherwise $f(u) = u^+ = max(0, u)$ [Glorot et al., 2011], is used except for the last layer of the encoder. The softmax is chosen as activation function at the output of the encoder in order to obtain probabilities that sum up to one. Recall, that the mean squared error loss is used as loss function (see Eq. 4.7). For all of the experiments, in order to update the parameters during the training phase, we use Adam optimizer, introduced in [Kingma and Ba, 2014], which is an adaption of the classical gradient descend. It uses the average first and second order moments of the gradient and corrected the introduced bias with parameters that control the decay rates of these moving averages.

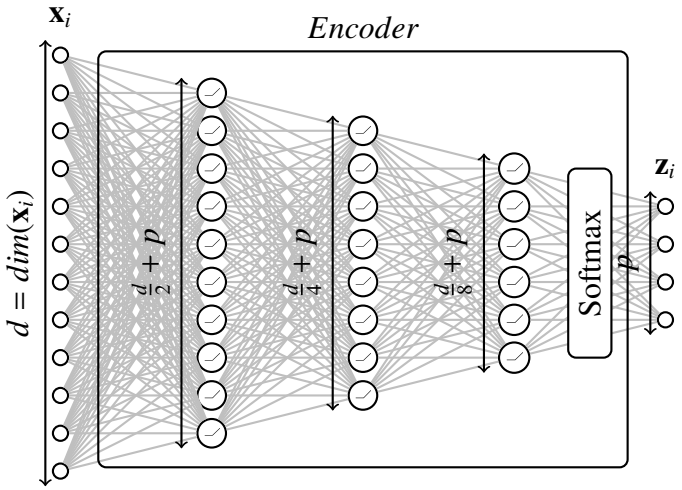


Figure 4.5: Encoder component of the Causal Auto-Encoder. It is composed of four layers for which the dimension is halved at each layer. The dimension p of the last layer of the encoder is defined by the chosen architecture of the latent space. The ReLU function is used as activation function at each layer, except for the last that used the Softmax function.

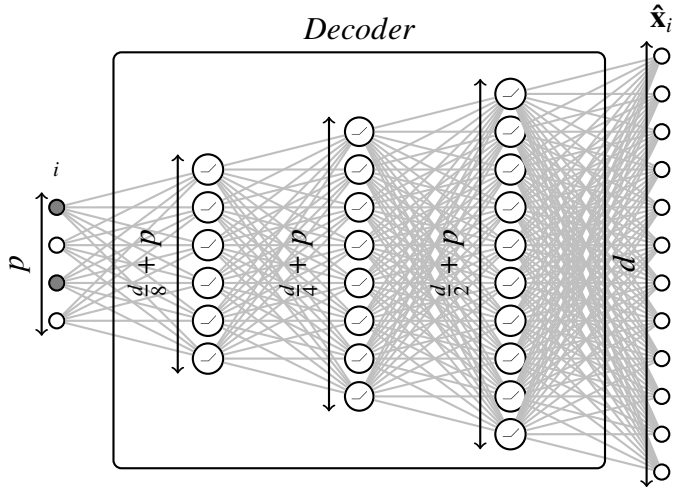


Figure 4.6: Decoder component of the Causal Auto-Encoder. It is composed of four layers for which the dimension is doubled at each layer. The dimension of the input is given by the size p fixed with the architecture of the latent space. The dimension of the output d is given by the dimension of covariates $\mathbf{x}_i$ (number of features). The ReLU function is used as activation function at each layer.

4.3.2 Experimental framework

Related work on counterfactual estimation can be categorized into three main classes: linear methods [Alaa and van der Schaar, 2017], random forest methods [Hill, 2011, Lu et al., 2018] and methods based on neural network architectures [Shalit et al., 2017, Yoon et al., 2018] (see

Chapter 2). In this section, we compare the CAE with models using two classfiers (T-learner), one for the treatment and the other for the control group. To cover each of the categories, we build three baseline models using respectively logistic regressions (T-LR), regression trees (T-RF) and multi-layer perceptron classifiers (T-MLPC). These methods do not provide confidence intervals, but they show good performances in the literature [Alaa and Schaar, 2018]. On synthetic datasets, we also compared our model with the ECM algorithm proposed in Chapter 3, which has shown its performance on a Gaussian mixture.

We use the same experimental setup as in Section 3.3.2. Recall that we evaluate out-of-sample performance over 10 trials and use a Wilcoxon signed-rank test with a level of 5% to confirm the significance of the results. We consider two metrics: expected Precision in Estimation of Heterogeneous Effect (ϵ_{PEHE}) and Area Under the Uplift Curve (AUUC) (see Section 2.4.2 for the description and limits of those metrics).

As previously, we conduce experiments on both synthetic and real-life datasets. First, several synthetic datasets with different complexities are generated. We draw a d-dimensional mixture of four distributions. Like in Chapter 3, the continuous variables are generated as multivariate Gaussians $\sum \pi_k \mathcal{N}(\mu_k, \Sigma_k)$ and the categorical variables follow each an independent multinomial distribution $\sum \mathcal{M}(\rho_k^1)...\mathcal{M}(\rho_k^{n_{\mathcal{M}}})$. For each distribution $k \in \{R, D, S, A\}$, π_k is the mixing probability and (μ_k,Σ_k) respectively the mean and the covariance of the Gaussian distribution. $\mu_k \sim \mathcal{U}((-u, u)^{(d)})$ is uniformaly distributed such as u is a parameter varying the complexity. $\Sigma_k \sim UAU^\top$ is generated as a random symmetric positive-definite matrix, where $A = (a_{ij})$ is a diagonal matrix such as $a_{ii} \sim \mathcal{U}[1, 2]$ and U is an orthogonal matrix extracted from the QR decomposition of the matrix $B^T B$, with $b_{ij} \sim \mathcal{U}[0, 1]$. The multinomial parameters ρ_k^j are randomly drawn according to a uniform distribution in the half-open interval $[0, 1[$. We conduce four experiments with different datasets with a balanced proportion of causal populations, composed with:

(i) only continuous variables, a low number of features ($d = 10$) and a low overlap ($u = 5$).

(ii) only continuous variables, a medium number of features ($d = 20$) and a low overlap ($u = 5$).

(iii) only continuous variables, a medium number of features ($d = 20$) and a significant overlap ($u = 1$).

(vi) only categorical variables, with a medium number of features ($d = 20$), a low overlap ($u = 5$).

We use in this section two common datasets IHDP and Twins (described in Section 2.4.1). Recall that, IHDP is based on a Infant Health and Development Program and was compiled by J. Hill in [Hill, 2011] for causal effect estimation. A bias between the treatment and control groups is introduced by removing a subset of the treated population. Twins, introduced in [Louizos et al., 2017], studies the impact of the weight at birth of a baby on her survival in the USA between 1989 and 1991. The sibling having the bigger weight is considered as being treated.

In addition, we consider two real-life datasets: Email and Criteo (described in Section 2.4.1). In

these datasets, the counterfactual outcome is missing and the real ITE is unavailable. Only the form of the Uplift curve and the AUUC can be used to evaluate the efficiency of these models. Email is a dataset based on an email marketing campaign [Hillstrom, 2008]. Customers are divided in two groups: one receiving an email from an advertising campaign and the other does not. The large-scale dataset Criteo was also created to investigate the impact of an advertising, but to form the control group, a random portion of the population was prevented from being targeted [Diemert et al., 2018].

4.3.3 Results and conclusion

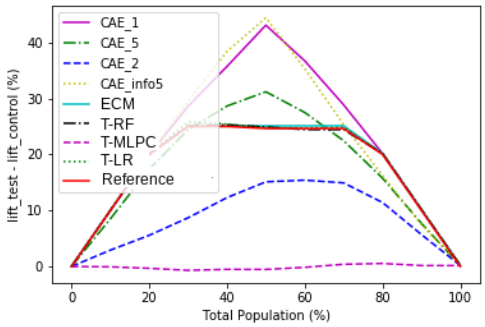
The results of the synthetic datasets are summarized in Table 4.1. According to the ϵ_{PEHE}, when the covariates are continuous (in the datasets (i), (ii) and (iii)), the baselines models T-LR, T-RF and ECM show high performances, which are better than the CAE models although there are not significant. Note, those experiments favour the ECM since its prior distribution is based on the data generation process. T-MLPC are poor results in this three first datasets. This can be explained by an overfit of the model. Moreover, when the data becomes more complex (datasets (*vi*)), it outperforms the other baseline models. Likewise, the CAE_1, CAE_5 and CAE_{info5} are the most efficient models in the case of categorical dataset. In general, on the four datasets, CAE_1, which encodes a population with one neuron, performs better than the other architectures of the Causal Auto-encoder. The gap is reduced as the complexity of the data increases. In low dimension, it seems that additional unconstrained neurons bring noise to the model. Similarly, encoding a population with more than one node seems to interfere the learning task. Overall, complex architecture models appear to improve with complexity of data. In term of AUUC, CAE models have higher significant values than the baselines. Theoretically, this should mean that they perform better. However, the comparison with the results of the ϵ_{pehe}, leads us to wonder about a deformation of the curve. Figure 4.7 displays the AUUC metric limits. An optimal classifier, based on the true distribution of data, highlights the error made by the CAE models. They predict responders and anti-responders in excess, leading to a sharp increase and decrease of the curve. The values of the AUUC are therefore higher while the models are less efficient. On the last synthetic dataset containing only categorical variables, although CAE models once again predict responders and anti-responders in excess, they are significantly better than models without neural networks predicting only doomed and survivors. Note that, due to the particular distribution of the data, the optimal classifier has not been plotted.

The numerical results on semi-synthetics datasets are given in Table 4.2. Due to the complexity of the data, ϵ_{pehe} values on IHDP dataset are high for all models compared to those on Twins dataset. On IHDP, CAE_{info5} and CAE_1 are the two most efficient models, according to both ϵ_{PEHE} and AUUC. On Twins, the baselines perform according to the ϵ_{PEHE} metric but not according to the AUUC. We can once again think about a deformation of the curve. However, when we observe the uplift curves in Figure 4.8, we notice that the CAE models capture well the distribution of the causal groups. The ranking of individuals in relation to each other is better with CAE models.

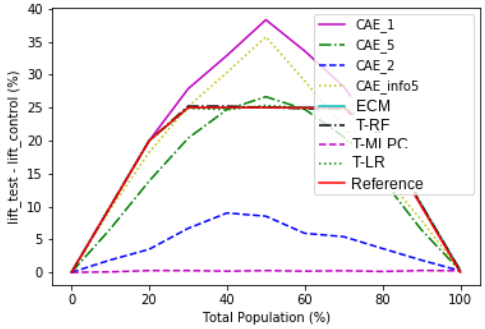
In purely real datasets, compared to the baselines, CAE models perform well. The uplift curves in Figure 4.8 have the expected shape. In particular, the larger the number of data is, the closer CAE_1

	(*i*)	(*ii*)	(*iii*)	(*iv*)
ϵ_{PEHE}				
T-LR	0.01 +/- 0.00	**0.00 +/- 0.00**	0.10 +/- 0.05	1.27 +/- 0.15
T-RF	**0.00 +/- 0.00**	**0.00 +/- 0.00**	**0.03 +/- 0.00**	1.01 +/- 0.04
T-MLPC	0.50 +/- 0.01	0.50 +/- 0.01	0.50 +/- 0.01	0.87 +/- 0.19
ECM	**0.00 +/- 0.00**	**0.00 +/- 0.00**	**0.00 +/- 0.00**	1.41 +/- 0.02
CAE_1	0.04 +/- 0.07	0.16 +/- 0.16	0.29 +/- 0.09	**0.49 +/- 0.01 (⋆)**
CAE_2	0.86 +/- 0.38	1.23 +/- 0.30	0.94 +/- 0.23	1.06 +/- 0.29
CAE_5	0.12 +/- 0.07	0.14 +/- 0.13	0.32 +/- 0.089	0.50 +/- 0.01
CAE_{info5}	0.25 +/- 0.17	0.35 +/- 0.11	0.28 +/- 0.10	0.50 +/- 0.01
AUUC				
T-LR	1864 +/- 73	1862 +/- 48	1831 +/- 52	116 +/- 75
T-RF	1851 +/- 63	1851 +/- 46	129 +/- 56	129 +/- 57
T-MLPC	85 +/- 53	70 +/- 25	70 +/- 25	116 +/- 75
ECM	1863 +/- 53	1863 +/- 60	1861.3 +/- 61	126 +/- 75
CAE_1	**2338 +/- 287 (⋆)**	**2396 +/- 185 (⋆)**	**2525 +/- 83 (⋆)**	2220 +/- 308
CAE_2	929 +/- 394	740.4 +/- 210	1436.9 +/- 461	952 +/- 894
CAE_5	1841 +/- 351	1586 +/- 383	2453.3 +/- 88	2182 +/- 368
CAE_{info5}	2271 +/- 390	1936 +/- 674	2489 +/- 64	**2525 +/- 83 (⋆)**

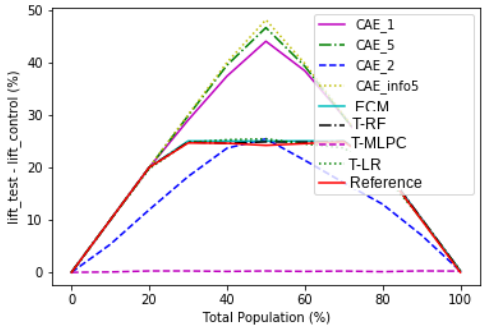
Table 4.1: ϵ_{PEHE} and AUUC results for synthetic datasets. CAE_1, CAE_2, CAE_5 are designed respectively with one, two and five nodes to encode the probabilities distribution of a causal populations, and no additional nodes. CAE_{info5} is built with one node to encode the probability of a causal populations and five additional nodes. T-LR, T-RF and T-MLPC are two classifiers models using respectively logistics regressions, regression trees and multi-layer perceptron, as classifiers. ECM is the Expectation-Causality-Maximization algorithm described in Chapter 3. Synthetics datasets are generated as a mixture of Gaussian for continuous variables and independent Multinomials for categorical variables. The studied cases are characterised by (i) a low number of features; (ii) a significant selection bias; (iii) a overleap between the causal populations; (iv) a dataset composed from only categorical variables. A (⋆) indicates a significant result using a Wilcoxon signed-rank test at level of 5% compared to second best baseline.



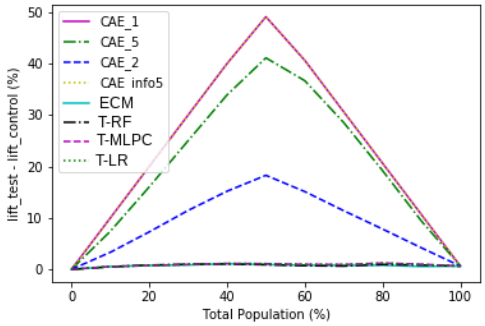
(a) Dataset (i): low number of features



(b) Dataset (ii): significant selection bias



(c) Dataset (iii): overleap between the causal populations



(d) Dataset (vi): dataset composed from only categorical variables

Figure 4.7: Uplift curves for synthetic datasets, built as the difference of the lift curves on the treatment and control groups. Synthetics datasets are generated as a mixture of Gaussian for continuous variables and independent Multinomials for categorical variables. The optimal classifier is built with the true distribution of the data.

	IHDP		Twins	
	ϵ_{PEHE}	AUUC	ϵ_{PEHE}	AUUC
T-LR	0.63 +/- 0.06	2271 +/- 442	0.04 +/- 0.01	87 +/- 45
T-RF	0.71 +/- 0.12	2438 +/- 353	0.04 +/- 0.01	86+/- 54
T-MLPC	0.68+/- 0.04	2246 +/- 365	**0.03 +/- 0.01**	97 +/- 60
CAE_1	0.47 +/- 0.10	**3535 +/- 481 ($\star$)**	0.35+/- 0.19	**293 +/- 99 ($\star$)**
CAE_2	0.69 +/- 0.24	2901 +/- 682	0.76+/- 0.20	185 +/- 159
CAE_5	0.51 +/- 0.08	3178 +/- 359	0.51 +/- 0.25	119 +/- 83
CAE_{info5}	**0.44 +/- 0.15 ($\star$)**	2989 +/- 860	0.38 +/- 0.36	244 +/- 155

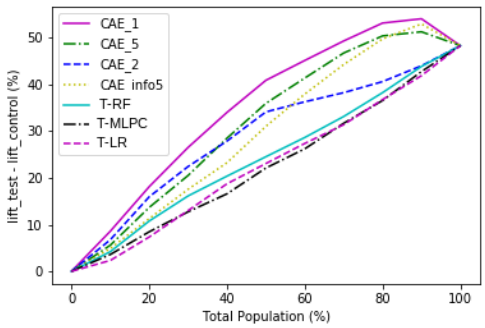
Table 4.2: Numerical results on semi-synthetic datasets. CAE_1, CAE_2, CAE_5 are designed respectively with one, two and five nodes that encode the probabilities distribution of a causal populations, and no additional nodes. CAE_{info5} is built with one node to encode the probability of a causal populations and five additional nodes. T-LR, T-RF and T-MLPC are two classifiers models using respectively logistics regressions, regression trees and multi-layer perceptron, as classifiers. ECM is the Expectation-Causality-Maximization algorithm described in Chapter 3. A ($\star$) indicates a significant result using a Wilcoxon signed-rank test at level of 5% compared to second best baseline.

curve has the perfect shape. The basic models show poor results on the Email and Criteo datasets, with misclassification on all causal populations. None of these models differ in performance from each other. Note that, the T-RF model is disqualified from the other models in large scale setups, due to its long computing time.

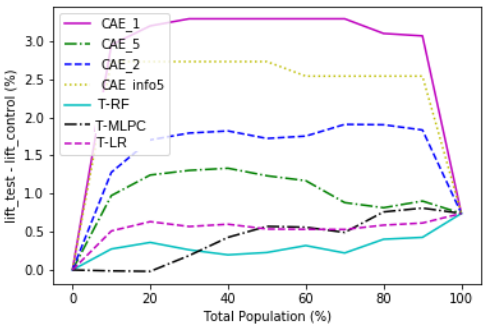
To conclude, we observe on all the datasets that CAE_2 and CAE_5 are the two architectures with the lowest performance, among the CAE architectures. We can deduce that using more than one neuron to encode the probability of a causal population, add noise. However, we observe that the unconstrained additional nodes present in model CAE_{info5} produce a better estimation of the distribution for complex data distributions.

4.4 Limits and extensions

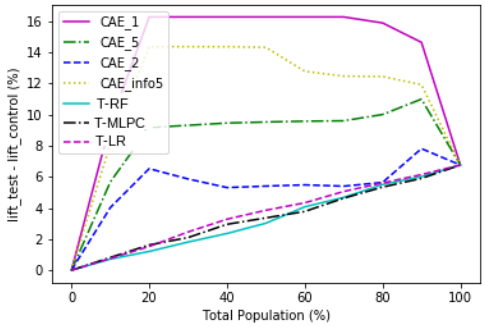
We propose an auto-encoder architecture which by means of partial information included on latent variables, estimates the probability distribution of causal populations. The proposed model is efficient on real data and has the significant advantage to be non-parametric and applicable on a large scale datasets. Compared to the ECM algorithm, it has the significant advantage that it does not need any assumptions about the prior distribution of the data. However, it has the disadvantage that little theory is currently available to justify the performance. In addition, experiments on synthetic datasets have highlighted the limitations of our model on low complexity



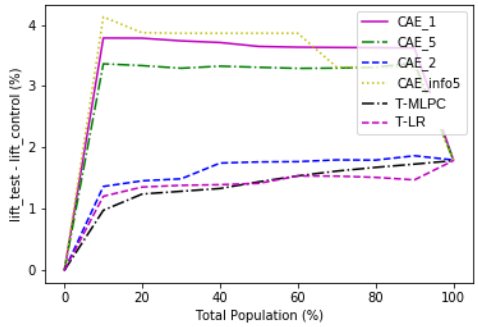
(a) IHDP



(b) Twins



(c) Email



(d) Criteo

Figure 4.8: Uplift curves for real (purely real and semi-synthetic) datasets, built as the difference of the lift curves on the treatment and control groups.

data distributions. In Section 4.4.1, we present three architectures to overcome the overfitting problem: sparse, denoising and contractive Auto-Encoders. Morevover, in contrast to the ECM, the proposed models in this chapter does not estimate the distribution of causal populations but only the probability of belonging to each of these populations. As a consequence, we lose information about the distribution of individuals and average behaviours. In Section 4.4.2, we propose a method to obtain a typical individual for each of the causal populations. Finally, we propose in Section 4.4.3 a model which is also based on a neural network architecture and which does not have the reconstruction constraint of an auto-encoder.

4.4.1 Regularized Auto-Encoder extensions

The experiments have shown that when data have a "too simple" distribution, CAE tends to overfit. To overcome this common problem in neural network models, we propose to investigate alternatives architecture regularizing the model.

Sparse Auto-Encoder

Sparse Auto-Encoder [Makhzani and Frey, 2013] is an alternative to introduce an information bottleneck. The loss function is constructed by penalizing activation within a layer. Some nodes through the network are forced to zero to selectively activate regions of the network depending on the input. This approach avoids overfit since it limits the network's capacity to memorize the input, while extracting features information. The sparsity constraint can be imposed by adding a L1 regularization term:

$$l(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \sum_i ||a_i^{(h)}|| \tag{4.9}$$

where $a_i^{(h)}$ is the vector of activations a in layer h for observation i and λ is a tunning parameter. An another way is to used the KL-divergence, which measures the difference between two probability distributions. The loss function depends in this case on a sparsity parameter, which represents the average activation of a neuron, and it is expressed as:

$$l(\mathbf{x}, \hat{\mathbf{x}}) + KL(\delta||\hat{\delta}_j) \tag{4.10}$$

where $\hat{\delta}_j = \frac{1}{m} \sum_i a_i^{(h)}(\mathbf{x})$ is the average activation on the neuron j in layer h, summing the activation for m sample. The implementation of this kind of architecture is compatible with the proposed CAE, as long as the sparsivity does not affect the last latent layer of the encoder on which causal constraints are applied.

Denoising Auto-Encoders

Another approach, to learn a generalizable encoding and decoding, is to corrupt the data with noise before they are given in input to the encoder. Denoising auto-encoder enforces the robustness by reconstructing the original input from its corrupted version. In a senses, this process artificially adds more diversity to the training data. Because the input and the output of the Auto-Encoder

are not the same, the model does not memorizes the training data. The parameters are estimated by minimizing the loss function:

$$l(\mathbf{x}, \hat{\mathbf{x}}) = l(\mathbf{x}, g(f(\tilde{\mathbf{x}}))) \tag{4.11}$$

where $\tilde{\mathbf{x}} \sim q(\tilde{\mathbf{x}} \mid \mathbf{x})$ is the corrupted input by means of a stochastic mapping by a fixed desired proportion. A common way to corrupt the input is to forced to zero a fixed number of features chosen at random [Vincent et al., 2008]. Another way is to drawn an i.i.d noise vector ϵ, such as the corrupted input is expressed as $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$. The noise distribution is specified with mean zero and variance s^2 in order to unbiased the noise. This data preprocessing does not affect the CAE model mechanism and can therefore be used without restriction [Pretorius et al., 2018].

Contractive Auto-Encoders

Contractive Auto-encoders add a penalty term that encourages the robustness of the learned representation for small variations of the input [Rifai et al., 2011]. The optimization problem aims to minimize the objective function:

$$l(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \|J_f(\mathbf{x})\|_F^2 \tag{4.12}$$

where $\|J_f(\mathbf{x})\|_F^2 = \sum_{ij} \left(\frac{\partial h_j(\mathbf{x}_i)}{\partial \mathbf{x}_i}\right)^2$ is the Frobenius norm of the Jacobian J_f and λ is a hyper-parameter controlling the strength of the regularization. In contrast to the two previous approaches, this method encourages robustness of the representation rather than robustness of the reconstruction. Both denoising Auto-Encoder and contractive Auto-Encoder can be combined to obtain robustness of the original input and the learned function. This architecture is a suitable alternative to the problems of overlearning encountered by the CAE, and will be the subject of future works.

4.4.2 Prototypical individual

Although the information required to calculate the ITE and deduce the causal population of individuals is just the probability that an individual belongs to a causal population, the approach in Chapter 3 estimates the distribution of causal populations. The advantage of the proposed ECM parametric approach is that it provides, for example for a Gaussian mixtures, the average features of a responder, a doomed, a survivor and an anti-responder and their variance, using the estimation of the parameters μ_k and Σ_k (see Section 3.2). This information is not directly obtained with the CAE approach, however we can obtain the features of a prototypical individual by using the trained decoder block of our model. If the vectors $\begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}$, $\begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix}$, $\begin{pmatrix}0 & 0 & 1 & 0\end{pmatrix}$ and $\begin{pmatrix}0 & 0 & 0 & 1\end{pmatrix}$ are given in input of the decoder, in output we obtain respectively the reconstructed features of a responder, dommed, survivor and anti-responder (see Figure 4.9).

4.4.3 Alternative neural networks

The proposed Auto-Encoder aims at estimating the probability of the causal population, but the loss function used for the training does not estimate the error related to the estimated probabilities

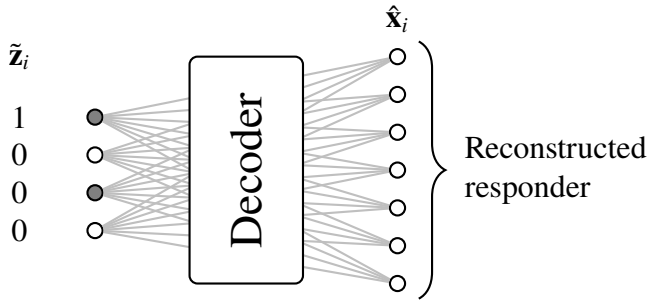


Figure 4.9: Estimation of a prototypical responder individual, using trained encoder CAE block. $\tilde{\mathbf{z}}_i$ is the prototypical latent variable and $\hat{\mathbf{x}}_i$ is the reconstructed individual. This method can be applied to reconstructed in the same way doomed, survivor and anti-responder individuals.

but the error related to the reconstruction of the input which is not then re-used. An alternative structure would be to estimate the probability distributions using a neural network and estimate the error with respect to the causal constraints, as a supervised learning problem. The error would be calculated with respect to the two probabilities, which should be set to zero according to the assigned treatment and the observed outcome. The structure is illustrated in Figure 4.10 and the loss function can be written as:

$$\sum_{k \in \{R,D,S,A\}} w_k(t_i, y_{i,obs})\, l(z_{ik}, 0) \tag{4.13}$$

where w_k is an activation function equal to 0 when t_i and $y_{i,obs}$ authorize the causal population and 1 otherwise, and l is the error of the estimated probability.

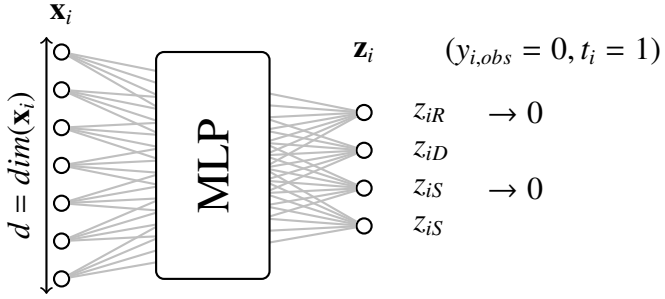


Figure 4.10: Multi-layers perceptron architecture example for a given sample x_i, where $y_{i,obs} = 0$ and $t_i = 1$.

5 Conclusion and Perspectives

Contents

5.1 Summary of the contributions

In this thesis, we propose a novel causal inference framework based on the estimation of probability distributions of causal populations. We solve the problem by estimating the hidden probability space of the causal populations, constrained with a partial information. This information, derived from the treatment assignment and the observed outcome, enforces the probability distributions of the two excluded populations to zero and normalizes the others. We propose two methods to estimate the probability distributions under causal constraints: A parametric method based on the Expectation-Maximization algorithm, and a non-parametric method using an enhanced Auto-Encoder.

The first method is an algorithm which determines the parameters of a given distribution. It has the advantage to give the exact distribution (and not only the probability distribution) of the causal groups and provide efficient results if the given prior distribution is appropriate to the data. The parameters obtained provide an interpretation of a trend in the populations, useful in practice. However, this algorithm is only applicable when the prior distribution has a closed form. We therefore propose a variational extension of the algorithm, which could be the subject of future experiments.

The second non-parametric method was proposed to overcome the prior on the distributions. In addition to have the high performance of neural network architectures, it is applicable to all

distributions. The probability distributions of causal populations are obtained by constraining the latent space of an Auto-Encoder from the observed data. Moreover, being able to use it on very high dimension is a major asset. The limitation of this architecture is that the algorithm may be subject to overfiting, when the problem is too "simple" problems, due to the number of learning parameters. Some work on the architecture adjustment would be valuable to improve the efficiency of the model.

The benefits of estimating the probability distribution of the causal populations are manifold:

- Classify individuals in causal populations.
- Estimate the counterfactual outcome and the individual treatment effect (ITE).
- Provide a treatment assignment policy for current patients.

Therefore, in this thesis we proposed a solution to the challenging task of estimating the causal effect on observational data, for which the treatment assignment procedure is unknown. We hope that this work will be useful to the scientific community and that it will contribute to advances in the field of causal inference, with a new perspective on the problem of counterfactual outcomes estimation. The main strength of this approach is that it can be adapted to multi-treatments, which will be the future issues in this field. In the following section we develop how to adapt our approach to multi-treatment, before discussing the open question of the non-compliance problem which is an unaddressed challenge in this thesis.

5.2 Multi-treatment extension

Temporary unavailable content

5.3 Open question of non-compliance

Throughout this thesis, we have assumed that the treatment assigned is the treatment given. However, in practice non-compliance is relatively common. For example, an individual might accidentally receive the wrong treatment or they may choose to not take their assigned treatment because of disinterest or fear of potential side effects. Thus, the taken treatment D is not always the same as the assigned treatment T. In the case of binary treatment and binary compliance (two compliance options: taken or not the treatment), four groups can be identified. An individual assigned to the treatment can take or not the treatment and an individual not assigned to the treatment can take or not the treatment (see Table 5.1). This problem is often referred in the literature to as an *intention-to-treat analysis*.

The authors in [Imbens and Rubin, 1997] introduce a Bayesian method for causal estimation in presence of noncompliance where the treatment assigned and the treatment received are both observed. The posterior distribution is estimated using EM and data augmentation algorithms. We propose to apply to our model a pre-processing method that uses this work. The posterior

	$t_i = 0$	$t_i = 1$
Complier (c)	$d_i = 0$	$d_i = 1$
Never taker (n)	$d_i = 0$	$d_i = 0$
Always taker (a)	$d_i = 1$	$d_i = 1$
Defier (d)	$d_i = 1$	$d_i = 0$

Table 5.1: Compliance behavior

distribution is expressed as:

$$\prod_{i \in \mathcal{S}(0,0)} (w_c g^i_{c0} + w_n g^i_{n0}) \prod_{i \in \mathcal{S}(0,1)} (w_a g^i_{a0} + w_d g^i_{d0}) \prod_{i \in \mathcal{S}(1,0)} (w_n g^i_{n1} + w_d g^i_{d1}) \prod_{i \in \mathcal{S}(1,1)} (w_c g^i_{c1} + w_a g^i_{a1}) \quad (5.1)$$

where $\mathcal{S}(.,.)$ are the subsets of units exhibiting each pattern of $(d_{i,obs}, t_{i,obs})$. $(w_j)_{j=\{c,n,a,d\}}$ are defined as the population probability of complier, never teaker, always taker and defier. $g^i_{js} = g(y_{i,obs} \mid \eta_{jt})$ for $j = \{c, n, a, d\}$ and $t = \{0, 1\}$ are the potential outcomes where the treatment s is given through η_{jt}, the parameter refer to the eight distributions.

The idea is to model a mixture of mixtures. We would not have 4 populations but instead $4^2 = 16$, each defined by the outcomes for each taken treatment and for each treatment assignment (see Table 5.2).

	Complier	Never taker
Responder	$(d_i = 1, t_i = 1, y_{i,obs} = 1)$ $(d_i = 0, t_i = 0, y_{i,obs} = 0)$	$(d_i = 0, t_i = 1, y_{i,obs} = 1)$ $(d_i = 0, t_i = 0, y_{i,obs} = 0)$
Doomed	$(d_i = 1, t_i = 1, y_{i,obs} = 0)$ $(d_i = 0, t_i = 0, y_{i,obs} = 0)$	$(d_i = 0, t_i = 1, y_{i,obs} = 0)$ $(d_i = 0, t_i = 0, y_{i,obs} = 0)$
Survivor	$(d_i = 1, t_i = 1, y_{i,obs} = 1)$ $(d_i = 0, t_i = 0, y_{i,obs} = 1)$	$(d_i = 0, t_i = 1, y_{i,obs} = 1)$ $(d_i = 0, t_i = 0, y_{i,obs} = 1)$
Anti-responder	$(d_i = 1, t_i = 1, y_{i,obs} = 0)$ $(d_i = 0, t_i = 0, y_{i,obs} = 1)$	$(d_i = 0, t_i = 1, y_{i,obs} = 0)$ $(d_i = 0, t_i = 0, y_{i,obs} = 1)$

	Always taker	Defier
Responder	$(d_i = 1, t_i = 1, y_{i,obs} = 1)$ $(d_i = 1, t_i = 0, y_{i,obs} = 0)$	$(d_i = 0, t_i = 1, y_{i,obs} = 1)$ $(d_i = 1, t_i = 0, y_{i,obs} = 0)$
Doomed	$(d_i = 1, t_i = 1, y_{i,obs} = 0)$ $(d_i = 1, t_i = 0, y_{i,obs} = 0)$	$(d_i = 0, t_i = 1, y_{i,obs} = 0)$ $(d_i = 1, t_i = 0, y_{i,obs} = 0)$
Survivor	$(d_i = 1, t_i = 1, y_{i,obs} = 1)$ $(d_i = 1, t_i = 0, y_{i,obs} = 1)$	$(d_i = 0, t_i = 1, y_{i,obs} = 1)$ $(d_i = 1, t_i = 0, y_{i,obs} = 1)$
Anti-responder	$(d_i = 1, t_i = 1, y_{i,obs} = 0)$ $(d_i = 1, t_i = 0, y_{i,obs} = 1)$	$(d_i = 0, t_i = 1, y_{i,obs} = 0)$ $(d_i = 1, t_i = 0, y_{i,obs} = 1)$

Table 5.2: Authorized values in each subpopulation. For an individual $i \in \{1, .., n\}$, d_i is the assigned treatment, t_i is the taken treatment and $y_{i,obs}$ the observed outcome.

Bibliography

[Abadie and Imbens, 2016] Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

[Alaa and Schaar, 2018] Alaa, A. and Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *ICML*.

[Alaa and van der Schaar, 2017] Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432.

[Alaa et al., 2017] Alaa, A. M., Weisz, M., and Van Der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*.

[Almond et al., 2005] Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

[Althauser and Rubin, 1970] Althauser, R. P. and Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology*, 76(2):325–346.

[Ambroise and Govaert, 2000] Ambroise, C. and Govaert, G. (2000). Em algorithm for partially known labels. In *Data analysis, classification, and related methods*, pages 161–166. Springer.

[Arjovsky et al., 2019] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

[Athey et al., 2019] Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

[Austin, 2011] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

[Baldi and Hornik, 1989] Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58.

[Beji et al., 2020] Beji, C., Bon, M., Yger, F., and Atif, J. (2020). Estimating individual treatment effects through causal populations identification. *arXiv preprint arXiv:2004.05013*.

[Bickel et al., 1975] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

[Bishop, 2006a] Bishop, C. (2006a). *Pattern Recognition and Machine Learning*. Springer.

[Bishop, 2006b] Bishop, C. M. (2006b). *Pattern recognition and machine learning*. springer.

[Bourlard and Kamp, 1988] Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Brookhart et al., 2006] Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156.

[Bühlmann et al., 2020] Bühlmann, P. et al. (2020). Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

[Chattopadhyay et al., 2019] Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. N. (2019). Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR.

[Chipman et al., 2007] Chipman, H. A., George, E. I., and McCulloch, R. E. (2007). Bayesian ensemble learning. In *Advances in neural information processing systems*, pages 265–272.

[Chipman et al., 2010] Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*.

[Cochran, 1968] Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313.

[Cole and Hernán, 2008] Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664.

[Côme et al., 2009] Côme, E., Oukhellou, L., Denoeux, T., and Aknin, P. (2009). Learning from partially supervised data using mixture models and belief functions. *Pattern recognition*, 42(3):334–348.

[D'Agostino Jr, 1998] D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281.

[Dehejia and Wahba, 1999] Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

[Dehejia and Wahba, 2002] Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

[Diemert et al., 2018] Diemert, E., Betlei, A., Renaudin, C., and Amini, M.-R. (2018). A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD*. ACM.

[Foster et al., 2011] Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

[Frangakis and Rubin, 2002] Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

[Funk et al., 2011] Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

[Gelman et al., 2004] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

[Gianicolo et al., 2020] Gianicolo, E. A., Eichler, M., Muensterer, O., Strauch, K., and Blettner, M. (2020). Methods for evaluating causality in observational studies: Part 27 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 117(7):101.

[Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[Gu and Rosenbaum, 1993] Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

[Hansen, 2004] Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.

[Harradon et al., 2018] Harradon, M., Druce, J., and Ruttenberg, B. (2018). Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*.

[Hernán et al., 2000] Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570.

[Hill et al., 2020] Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

[Hill, 2011] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[Hillstrom, 2008] Hillstrom, K. (2008). The minethatdata e-mail analytics and data mining challenge. *MineThatData blog*.

[Hinton and Zemel, 1994] Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10.

[Hirano et al., 2003] Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

[Hitsch and Misra, 2018] Hitsch, G. J. and Misra, S. (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*.

[Holland, 1986] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[Imai et al., 2007] Imai, K., King, G., and Stuart, E. A. (2007). Misunderstandings among experimentalists and observationalists: balance test fallacies in causal inference. *Journal of the Royal Statistical Society, Series A*.

[Imbens and Rubin, 1997] Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327.

[Imbens and Rubin, 2015] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[Imbens and Wooldridge, 2009] Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[Ishwaran and Malley, 2014] Ishwaran, H. and Malley, J. D. (2014). Synthetic learning machines. *BioData mining*, 7(1):28.

[Jaskowski and Jaroszewicz, 2012a] Jaskowski, M. and Jaroszewicz, S. (2012a). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*.

[Jaskowski and Jaroszewicz, 2012b] Jaskowski, M. and Jaroszewicz, S. (2012b). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*.

[Johansson et al., 2016] Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029.

[Julious and Mullee, 1994] Julious, S. A. and Mullee, M. A. (1994). Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481.

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

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Knaus et al., 2021] Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161.

[Künzel et al., 2019] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

[Kusner et al., 2017] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

[Lee et al., 2010] Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

[Li and Shimizu, 2018] Li, C. and Shimizu, S. (2018). Combining linear non-gaussian acyclic model with logistic regression model for estimating causal structure from mixed continuous and discrete data. *arXiv preprint arXiv:1802.05889*.

[Lo, 2002] Lo, V. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86.

[Louizos et al., 2017] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *NIPS*.

[Lu et al., 2018] Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219.

[Lunceford and Davidian, 2004] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

[Madumal et al., 2020] Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. (2020). Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2493–2500.

[Magliacane et al., 2017] Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2017). Domain adaptation by using causal inference to predict invariant conditional distributions. *arXiv preprint arXiv:1707.06422*.

[Makhzani and Frey, 2013] Makhzani, A. and Frey, B. (2013). K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

[Martin and Martin, 2015] Martin, A. and Martin, C. (2015). Simpson's paradox: Why smoking reduces the risk of dying of cardiovascular disease. *Value in Health*, 18(7):A383.

[Marx and Vreeken, 2018] Marx, A. and Vreeken, J. (2018). Causal inference on multivariate and mixed-type data. In *ECML PKDD*.

[McCaffrey et al., 2004] McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

[McLachlan and Peel, 2004] McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[Ming and Rosenbaum, 2000] Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56(1):118–124.

[Morgan and Todd, 2008] Morgan, S. L. and Todd, J. J. (2008). 6. a diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, 38(1):231–282.

[Morgan and Winship, 2015] Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

[Müller, 1997] Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

[Ng et al., 2011] Ng, A. et al. (2011). Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19.

[Nie and Wager, 2020] Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika, forthcoming*.

[Pearl, 1995] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

[Pearl, 2009] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

[Pearl, 2017] Pearl, J. (2017). Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389.

[Pearl, 2018] Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

[Pearl et al., 2000] Pearl, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*.

[Pearl et al., 2016] Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

[Pearl and Mackenzie, 2018] Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

[Peysakhovich et al., 2019] Peysakhovich, A., Kroer, C., and Lerer, A. (2019). Robust multi-agent counterfactual prediction. In *Advances in Neural Information Processing Systems*, pages 3077–3087.

[Plato, 1925] Plato (1925). *Plato in Twelve Volumes, Vol. 9 translated by W.R.M.*, volume 9. Lamb. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd.

[Pretorius et al., 2018] Pretorius, A., Kroon, S., and Kamper, H. (2018). Learning dynamics of linear denoising autoencoders. *arXiv preprint arXiv:1806.05413*.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

[Rifai et al., 2011] Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Icml*.

[Rosenbaum, 1987] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

[Rosenbaum, 2002] Rosenbaum, P. R. (2002). Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer.

[Rosenbaum and Rubin, 1983] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[Rosenbaum and Rubin, 1984] Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

[Rubin, 1974] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

[Rubin, 1975] Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, volume 233, page 239. American Statistical Association Alexandria, VA.

[Rubin, 1978] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

[Rubin, 1979] Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.

[Rubin, 2001] Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188.

[Rubin, 2003] Rubin, D. B. (2003). Basic concepts of statistical inference for causal effects in experiments and observational studies. *Course material in Quantitative Reasoning*, 33.

[Rubin and Thomas, 1996] Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[Schwab et al., 2018] Schwab, P., Linhardt, L., and Karlen, W. (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

[Setoguchi et al., 2008] Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.

[Shalit et al., 2017] Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org.

[Shi et al., 2019] Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2503–2513.

[Shpitser and Pearl, 2006] Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *UAI*.

[Splawa-Neyman et al., 1990] Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[Stadie et al., 2018] Stadie, B. C., Künzel, S. R., Vemuri, N., and Sekhon, J. S. (2018). Estimating heterogeneous treatment effects using neural networks with the y-learner.

[Vigen, 2015] Vigen, T. (2015). Spurious Correlations - website. *Tylervigen.Com*, page 189.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

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

[Wasserman, 2013] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[Yao et al., 2020] Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

[Yao et al., 2018] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31:2633–2643.

[Yoon et al., 2018] Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*.

[Zaniewicz and Jaroszewicz, 2013] Zaniewicz, Ł. and Jaroszewicz, S. (2013). Support vector machines for uplift modeling. In *ICDMW*, pages 131–138.

[Zhang, 2020] Zhang, J. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR.

[Zhang et al., 2013] Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML)*, pages 819–827.

[Zhao, 2004] Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *Review of economics and statistics*, 86(1):91–107.

A Additional explanations on causality

In this section, we introduce some concept used in causal inference.

A.1 Causation vs Correlation

Causation is to be distinguished from correlation. An association between two variables does not necessarily mean that one of those variables causes the other. For example, “ice cream sales” is proportional to the number of “sunburns”. This does not mean that one implies the other, but both “ice cream sales” and “sunburns” are more common in “hot and sunny weather” (see Figure A.1).

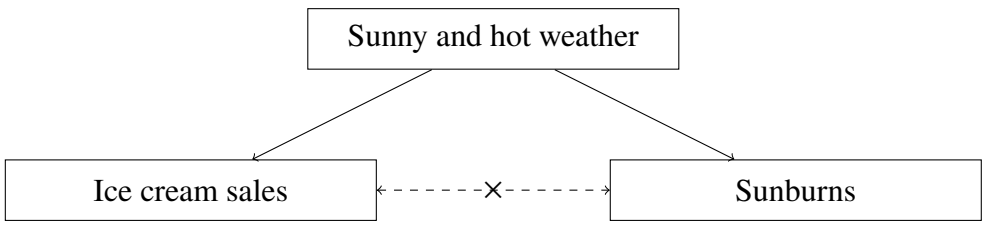


Figure A.1: Example of correlation between "Ice cream sales" and "sunburns" due to the causality coming from "Sunny and hot weather."

Causation explains the link between two variables that appear to be correlated without any explanation, as the correlation between US spending on science, space and technology with suicides by handing, strangulation and suffocation [Vigen, 2015]. The distinction between causality and correlation is clearly identifiable by DAGs (Figure A.2). If a variable Z causes two variables X and Y at once, then X and Y are correlated, but that does not necessarily mean that one causes the other.

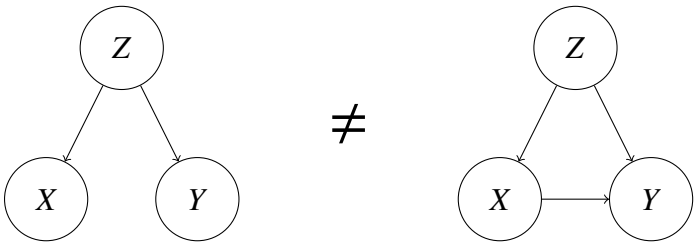


Figure A.2: Correlation vs Causation

A.2 Simpson's paradox

Simpson's paradox refers to a trend observed in several sub-populations which disappears or is reversed when these sub-populations are aggregated in the same population. A common example is the sex bias induced in graduate admissions of the University of California, Berkeley for the fall 1973 [Bickel et al., 1975]. The study of admissions in relation to the sex of candidates, appeared to discriminate against women. As show in Table A.1 the number of males actually admitted is higher than the number expected (with a equal chance of admission between men and women), while the number of females is lower than expected. This could highlight that the decision of admission is influenced by the sex of applicant.

	Observed	Expected	Difference
Men	3738	3460.7	277.3
Women	1494	1771.3	−277.3

Table A.1: Number of admissions of the University of California, Berkeley for the fall 1973 by sex of applicant (*extracted from [Bickel et al., 1975]*).

However, by subdividing admissions by type of department, the difference between the observed value and the expected value is equal to zero (see Table A.2). The observed difference in the total number of applications comes from that women tend to apply to departments with a more selective admissibility.

Others examples show this paradox. In healthcare, [Julious and Mullee, 1994] highlights an existing paradox concerning success rates in the elimination of kidney stones. When open surgery is compared to percutaneous nephrolithotomy, the second treatment appears more efficient. However, this is no true if the data is divided according to the size of the stones (see Table A.3).

This paradox results from the probability of undergoing open surgery or percutaneous nephrolithotomy varying according to the diameter of the stones. When the stone has a small diameter, percutaneous nephrolithotomy is more common. Meanwhile, when the stone has a small diameter, the rate of successful elimination is higher. A bias appears in the results.

Similarly, if all the population is considered (without subdividing into age ranges), the false conclusion that smoking reduces the risk of cardiovascular mortality can be deduced. This is

	Observed	Expected	Difference
Departement of machismatics			
Men	200	200	0
Women	100	100	0
Departement of social warfare			
Men	50	50	0
Women	150	150	0
Totals			
Men	250	229.2	20.8
Women	250	270.8	−20.8

Table A.2: Simpson's paradox in the number of admissions of the University of California, Berkeley for the fall 1973 by sex of applicant and departments (*extracted from [Bickel et al., 1975]*).

	Stone < 2cm	Stone ≥ 2cm	*Totals*
Open surgery	93%	73%	78%
Percutaneous nephrolithotomy	83%	69%	83%

Table A.3: Simpson's paradox in the success rate in elimination of kidney stones (*extracted from [Julious and Mullee, 1994]*).

explained by smoking implies a more premature mortality. In parallel, mortality from cardiovascular diseases occurs in advanced age. The population dying from cardiovascular causes is less dense with smokers [Martin and Martin, 2015]. Another famous example is the birthweight paradox [Pearl et al., 2016]. In a study on infant mortality, it was shown that low birth weight babies born to smoking mother had a lower mortality rate than those born to non-smoking mother. This would be means that it is less dangerous for a low birth weight baby to have a smoking mother. This is explained by the segregation. A mother who smokes causes a low weight to the baby without being dangerous to his survival, but a non-smoking mother, who have a low weight baby, has unaddressed factor that causes more sudden infant death.

The Simpson's paradox is due to spurious confounding factors and covariate selection. This phenomenon emerged from the observational nature of a correlation between data.

A.3 Structural causal models

Structural causal models are developed by Pearl [Pearl, 1995, Pearl, 2009, Pearl et al., 2016]. They describe the relationship between variables of interest with a set of mathematical functions. They are a set of endogenous variables V (observed variables internal to the model), exogenous variables U (unobserved variables external to the model and imposed by the environment) and functions f that express the causal relationship between U and V such as $f(U) = V$. Usually knowledge about causal relationship is not quantitative but qualitative. The expression of f in terms of a graph is then chosen rather than in terms of a mathematical formula, to provide

a more intuitive understanding of the relationship between the variables. A structural causal model can be associated to a graphical causal model obtained with a DAG. The exogenous U and endogenous V variables are represented by nodes and causal links by arcs with a associated direction representing the function f. The two following definitions can be used to determine the causality.

- X is a *direct cause* of Y, if in a graphical model a variable, X is the child of an another variable Y.
- X is a *potential cause* of Y, if in a graphical model a variable, X is the descendant of an another variable Y.

An SCM often consist of both the DAG and the structural equations:

- $U = \{X, Y, Z\}$
- $V = \{U_X, U_Y, U_Z\}$
- $f = \{f_X : X \rightarrow U_X, f_Y : Y \rightarrow U_Y, f_Z : Z \rightarrow U_Z\}$

B

Overview of propensity score methods

In this section, we introduce propensity score methods which estimate the Average Treatment Effect (ATE), defined as $\tau_0 = \mathbb{E}[Y_1 - Y_0]$. These methods differs from approach to estimate the individual treatment effect presented in Chapter 2 methods, but they were pioneers in counterfactual estimation and particularly in epidemiology to estimate the effect of a medical treatment on a patient population.

In the context of a binary treatment, an intuitive way to estimate average treatment effects is to determine separately the average treatment effect on the treatment group and on the control group, and to calculate the difference of obtained on the two groups. However, when groups are not perfectly homogeneous, i.e. constructed in a completely random way with a sufficient number of individuals to represent a population, biases may occur.

B.1 Introduction to propensity score

The *propensity score* is defined by [Rosenbaum and Rubin, 1983] as the marginal treatment probability, i.e. the probability for an individual of receiving the treatment assignmen conditioning on the observed covariates: $\forall i \in \{1, n\}$, $p_i = p(\mathbf{x}_i) := \mathbb{P}(T = 1|X = \mathbf{x}_i)$. In epidemiology, the propensity score approach is the classical method to deal with confounding effects in observational studies [Gianicolo et al., 2020]. Since the distribution of covariates $\mathbf{x}_i$ is the same for each treatment group (if $T = 0$ or $T = 1$) conditionally to the propensity score $p(\mathbf{x}_i)$, it is a balancing score. On randomized experiments, the propensity score is equiprobable on all treatment groups. In controlled experiments, it is known and defined by the specifications of the study. The true propensity score is often unknown in observational studies. Propensity score methods have the advantage of explicitly determining the balance between the control and treatment groups. Usual models, based on propensity score, consist in two steps. In the first step, the propensity score is estimated from covariates, using different methods. For example, the logistic regression model, that introduces a β parameter of estimation, is the most common parametric model used. The propensity score is expressed as $p(T = 1 \mid \mathbf{x}_i, \beta) = [1 + \exp(-\mathbf{x}_i\beta)]^{-1}$ and

β may be estimated by maximising the likelihood, in a scenario of additivity and linearity of the model [Setoguchi et al., 2008]. Other non-parametric methods can also be used, as boosted regression [McCaffrey et al., 2004], boosted CART [Lee et al., 2010], recursive partitioning and neural network [Setoguchi et al., 2008], and so on. In the second step, the counterfactual outcome is estimated according to the estimation of propensity score of each individual.

Since these methods are based on the propensity score, the relationship between the covariates and the treatment variable must be carefully considered. The aim is to use covariates that explain the outcome, without adding a confounding bias due to the treatment. Introducing variables unrelated to the outcome but related to treatment assignment significantly increases the confounding bias in the estimation of the treatment effect. In [Brookhart et al., 2006], authors show that the best model considers all variables related to the outcome, whether or not they are related to the treatment assignment. In practice, identifying covariates that affect only the treatment assignment, only the outcome, and neither the treatment nor the outcome, could be difficult. Several articles provide some guidelines for classifying covariates [Brookhart et al., 2006, Frangakis and Rubin, 2002].

B.2 Matching methods

The objective of matching methods is to reduce the treatment assignment bias by imitating a random assignment. Homogeneous treatment groups are created by associating individuals with similar properties and different treatment assignment (see Figure B.1).

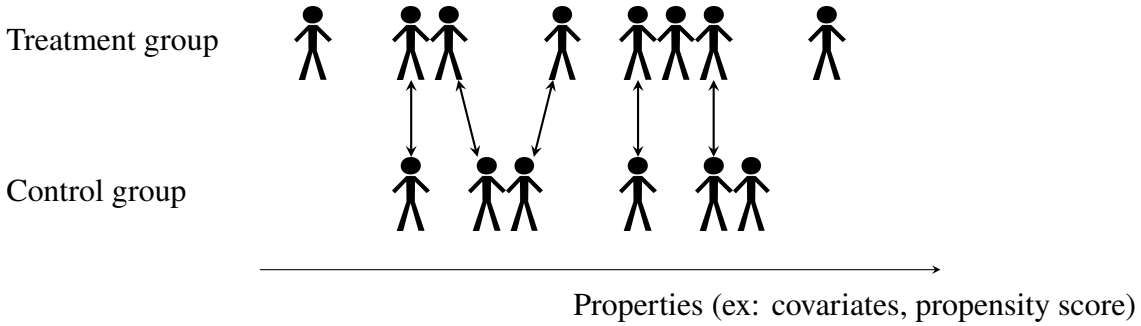


Figure B.1: One-to-one matching in binary treatment.

First matching methods have been used by matching individuals with a similar value of covariate, considering just one covariate. Two units with similar covariate, in the treatment and control group, are matched together. The size of the "reservoir" and the quality of matches are then discussed [Althauser and Rubin, 1970]. One-to-one matching, that formed pairs of treated and untreated individuals, is the most common approach. Other methods match one individual with several others. When the number of individuals to be matched together is variable, a better reduction of the bias is achieved [Ming and Rosenbaum, 2000]. A full matching is to match all the individuals at least once [Hansen, 2004]. The matching can be with or without replacement, i.e. once an individual has been selected to be matched, it can be or not reused in another match [Rosenbaum, 2002]. It is necessary to have a large number of data and a sufficient number

of units in each treatment groups to ensure the preservation of all the information. Including multiple covariates was a challenging problem. As it is difficult to find pairs with values close to all the covariates, different distances between individuals were then used:

- Exact matching [Imai et al., 2007]: $d_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{else} \end{cases}$
- Mahalanobis [Rubin, 1979, Zhao, 2004]: $d_{ij} = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}$
- Propensity score [Rosenbaum and Rubin, 1983, Rubin and Thomas, 1996]: $d_{ij} = (p_i - p_j)$, where p_k is the propensity score for individual k.
- Linear propensity score [Rubin, 2001]: $d_{ij} = |logit(p_i) - logit(p_j)|$, where $logit(p_k) = \log(\frac{p_k}{1-p_k})$ for $p_k \in]0, 1[$.

The approaches based on the difference in propensity score or in logit of propensity scores is preferred for bias reduction [Rosenbaum and Rubin, 1983, D'Agostino Jr, 1998].

In [Gu and Rosenbaum, 1993], authors propose a greedy and optimal matching algorithm based on propensity score. An untreated individual, close in term of propensity score, is randomly selected to be matched with a treated individual. The procedure is repeated until all untreated individual have been matched or when no match can be found. As it is uncommon to find units with exactly the same propensity score, a nearest neighbor matching is used to match two individuals together. The individual with the closest propensity score value is matched with a given individual from the other group. If multiple individuals have a propensity score equally close, then one individual is selected at random. This nearest neighbor matching imposes that a maximum distance must be specified. The absolute difference must thus be below a specified threshold.

Once the matched sample has been created, the treatment effect is estimated as the difference between outcomes in the matched sample. A study in a propensity score matched sample is equivalent to a randomized controlled trial experiment. Outcomes can be directly compared. The major disadvantage of the matching methods is that some individuals may not find equivalent individuals in the other group. This problem involves an information loss, but also introduces another source of selection bias. Matching methods are therefore limited by the fact that the control group must contain at least as many individuals as the treatment group.

B.3 Stratification

Stratification approaches based on propensity score (sometimes referred to as sub-classifications) aim to stratifying individuals into mutually exclusive stratum based on their estimated propensity score. Individuals are ranked from their propensity score and distributed in different strata. Standard methods operate through the following steps [Rosenbaum and Rubin, 1984, Frangakis and Rubin, 2002]:

- The propensity score $\hat{p}(\mathbf{x}_i)$ is estimated for all individuals i.

- According to their score, individuals are ranked and stratified into subsets. K strata are created according to the sample quantiles of the $\hat{p}(\mathbf{x}_i)$, where for $j \in \{1, .., K\}$, the quantile $\hat{q}_j$ is constructed such as $\hat{p}(\mathbf{x}_i) \leq \hat{q}_j$, the proportion of $\hat{q}_j$ is approximately j/K, $\hat{q}_0 = 0$ and $\hat{q}_K = 1$.

- Within each stratum, the effect of the treatment is estimated as the difference between the outcome of treated and untreated individuals.

- The global effect of the treatment is estimated by weighting each stratum according to the proportion of observations in each stratum.

Moreover, a regression model can be introduced in the third step. Within stratum, an adjustment can be added to reduce residual confusion within stratum (difference between treated and untreated groups) [D'Agostino Jr, 1998, Lunceford and Davidian, 2004].

Dividing the population by using the quantiles of a continuous confounding or of the estimated propensity score variable, eliminates already 90% of the bias due to measured confounders [Cochran, 1968, Rosenbaum and Rubin, 1984]. Stratification does not have the disadvantage of matching, which can reduce the precision of the estimated treatment effect by excluding some individuals, but stratification can introduce a greater bias with the creation of residual confounding within stratum which conduce to a minor residual imbalance [Austin, 2011]. A better balance was achieved by matching than by stratification methods.

B.4 k-Nearest Neighbors matching for stratification

The nearest neighbors (kNN) method introduced in [Dehejia and Wahba, 2002], is applied to stratification and commonly used as a baseline. The algorithm works as follow:

- A parsimonious logistic function is used to estimate the propensity score.

- All observations are ranked from lowest to highest propensity score.

- The observations are divided into strata. The strata are created to ensure that the difference in propensity score for the treated and control observations is insignificant.

- The distribution of covariates between the treatment and control groups in each stratum is tested. If the covariates are not balanced in a stratum, the stratum is divided into finer strata and re-evaluated. If some covariates remain unbalanced for all the divided strata, they are re-used to adjust the logistic function.

This method has the advantage to be efficient even when there are a few individuals in one of the treatment groups. The major difference is, after individuals are matched, the unmatched ones are directly used to estimate the propensity score.

B.5 Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weighting (IPTW) is a weighting method that reweights the units in order to make the control and treatment groups comparable. It was introduced first in [Rosenbaum, 1987] and has diverse and overlapping origins [Lunceford and Davidian, 2004, Morgan and Todd, 2008]. This method uses the propensity score to give a weight for all units i, given by:

$$w_i = \frac{t_i}{p_i} + \frac{(1 - t_i)}{1 - p_i} \tag{B.1}$$

where $t_i = \{0, 1\}$ is the treatment assignment and p_i is the propensity score. The weights are defined as the inverse of the probability of receiving the treatment that the unit actually received. Thus, a synthetic sample is created in which the distribution of covariates is independent of the treatment assignment. To estimate Average Treatment Effect (ATE), two popular estimators are used:

$$\hat{ATE}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{t_i y_{i,obs}}{\hat{p}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - t_i)y_{i,obs}}{1 - \hat{p}_i} \tag{B.2}$$

$$\hat{ATE}_2 = (\sum_{i=1}^{n}\frac{t_i}{\hat{p}_i})^{-1}\sum_{i=1}^{n}\frac{t_i y_{i,obs}}{\hat{p}_i} - (\sum_{i=1}^{n}\frac{1 - t_i}{1 - \hat{p}_i})^{-1}\sum_{i=1}^{n}\frac{(1 - t_i)y_{i,obs}}{1 - \hat{p}_i} \tag{B.3}$$

In [Lunceford and Davidian, 2004], more estimators are described and compared with stratification-based estimators. To decrease the variance of the estimate, when weights produce a bias in the treatment effect, alternative methods adapting weights can be used and are also described in this paper. Moreover, when the weights are inaccurate or unstable (units with a low probability of receiving treatment), stabilizing weights are proposed in [Hernán et al., 2000]. One of the limitations of weighting methods is that in areas of overlap between the probability of belonging to the treatment or control group, the weights of individuals can be extremely large and induce wide confidence intervals when the specifications of the weights are correct [Cole and Hernán, 2008].

B.6 Covariate adjustment

Covariate adjustment methods use a regression of the outcome variable from the treatment assignment and the propensity score. A logistic regression model can be expressed as:

$$logit(\mathbb{P}(Y = 1|X = \mathbf{x}_i, p_i)) = b_0 + b_1\mathbf{x}_i + b_2 p_i \tag{B.4}$$

The treatment effect is then determined using the coefficient regression $\beta = \{b_0, b_1, b_2\}$ from the fitted regression model. This approach requires that the relationship relating the outcome to treatment and propensity score should be correctly modeled and a regression model occurs relating outcomes to treatment assignment and covariates.

B.7 Doubly Robust estimation

The Doubly Robust (DR) estimation model, introduced in [Funk et al., 2011], estimates the causal effect using both outcome regression and propensity score. The outcome is regressed on a covariate function within each treatment group, using propensity score weighting. The estimated parameters are used to calculate the counterfactual outcome given covariates values. In addition, the treatment is modeled by a function of covariates to estimate the propensity score. Knowing the propensity score and the potential outcome, the doubly robust estimator is calculated in presence and absence of treatment for each unit. The observed outcome is combined with the estimated counterfactual outcome on each treatment group. Then, the mean of the value returned by the doubly robust estimator is calculated across the entire population on the control and treatment group, and the ratio effect measure is calculated.

B.8 Neural networks exploiting the sufficiency of propensity score

The model, proposed in [Shi et al., 2019], uses a neural network for estimating the treatment effect from observational data. This paper proposes to estimate the causal effect by predicting the counterfactual outcome by exploiting the sufficiency of propensity score, and then producing an estimator for the ATE. If the average treatment effect is identifiable, i.e. if $ATE = \mathbb{E}[\mathbb{E}[Y|X = \mathbf{x}, T = 1] - \mathbb{E}[Y|X = \mathbf{x}, T = 0]]$ then $ATE = \mathbb{E}[\mathbb{E}[Y|g(X), T = 1] - \mathbb{E}[Y|g(X), T = 0]]$, where g is the propensity score $g(x) = P(T = 1|X = \mathbf{x})$. A neural network architecture, named *Dragonnet* (see Figure B.2) using two deep networks is implemented. The first network provides a prediction of the treatment value which creates an estimator of propensity score. The second uses the penultimate layer of the first network as feature to predict the outcome.

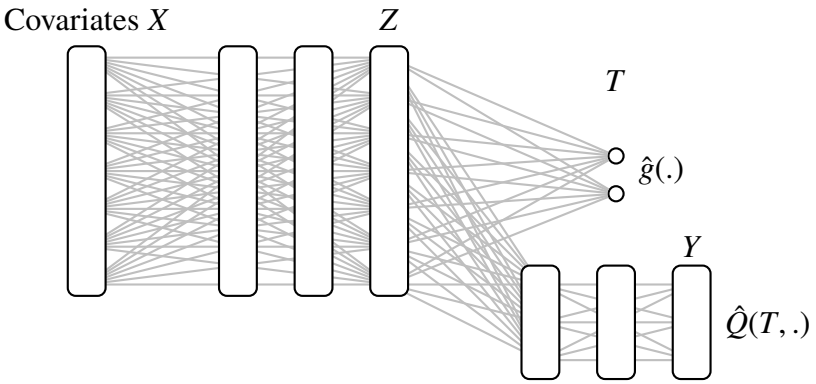


Figure B.2: Dragonnet architecture

The model ideally chooses a trade-off between the prediction of the treatment and the prediction of the conditional outcome from covariates and treatment information. After proposing estimators for the propensity score and the conditional outcome, the authors introduce a *targeted*

regularization to estimate the ATE. This procedure introduces a bias in the model in order to obtain non-parametrically optimal asymptotic properties. This modifies the objective function to guarantee desirable asymptotic robustness and efficiency properties on ATE.

Introduced propensity score methods have the advantage of being intuitive and easy to interpret. They explicitly determine the degree of balance between individuals in the treatment and control groups, and reduce the imbalance in measured characteristics. However, they use simple models, often less efficient than tree-based or neural network methods.

C

Maximum likelihood estimation

In this section, we provide the calculations of the maximum likelihood estimation used in the ECM algorithms.

C.1 Gaussian mixture model

Recall that, each causal population $k \in \{R, D, S, A\}$ follows a normal distribution $\mathcal{N}(\mu_k,\Sigma_k)$ and has a probability π_k in the whole set. The marginal distribution of the mixture model is given by:

$$p(\mathbf{x}, \mu, \Sigma, \pi) = \sum_{k=\{R,D,S,A\}} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k) \tag{C.1}$$

Similarly, the joint probability of observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=\{1,..,n\}}$ is:

$$p(\mathbf{X}, \mu, \Sigma, \pi) = \prod_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k) \tag{C.2}$$

The complete likelihood takes the form:

$$p(\mathbf{X}, \mathbf{Z} \mid \mu, \Sigma, \pi) = \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} (\pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k))^{z_{ik}} \tag{C.3}$$

$$= \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} \left(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma_k}} exp\left(-\frac{1}{2} (\mathbf{x}_i - \mu_k)^T \Sigma^{-1} (\mathbf{x}_i - \mu_k) \right) \right)^{z_{ik}} \tag{C.4}$$

and the log-likelihood is expressed as:

$$\mathcal{L}(\pi, \mu, \Sigma) = \log p(\mathbf{X}, \mathbf{Z} \mid \pi, \mu, \Sigma) \tag{C.5}$$

$$= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \log \left(\pi_k \mathcal{N}(\mathbf{x}_i, \mu_k, \Sigma_k)\right)^{z_{ik}} \tag{C.6}$$

$$= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \left(z_{ik} \log(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{det\Sigma_k}}) - z_{ik} \frac{(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)}{2} \right) \tag{C.7}$$

The expected value of the complete-data log-likelihood to maximize is equal to:

$$\mathbb{E}[\mathcal{L}(\pi, \mu, \Sigma)] = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \left(\gamma(z_{ik}) \log(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{det\Sigma_k}}) - \gamma(z_{ik}) \frac{(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)}{2} \right) \tag{C.8}$$

$$= \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \left(\gamma(z_{ik}) \log(\pi_k) - \frac{\gamma(z_{ik})}{2} \log(det\Sigma_k) - \frac{\gamma(z_{ik})}{2} (\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) + C \right) \tag{C.9}$$

where $\gamma(z_{ik}) = p(z_{ik} = 1 \mid X = \mathbf{x}_i)$ and C is a constant term of μ, Σ and π.

Maximizing the expected value of the complete log-likelihood is equivalent to set its partial derivatives to zero.

$$\frac{\partial \mathcal{L}(\mu, \Sigma, \pi)}{\partial \mu_k} = 0 \Leftrightarrow -\sum_{i=1}^{n} \gamma(z_{ik}) \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) = 0 \tag{C.10}$$

$$\Leftrightarrow -\Sigma_k^{-1} \left(\sum_{i=1}^{n} \gamma(z_{ik}) (\mathbf{x}_i - \mu_k) \right) = 0 \tag{C.11}$$

$$\Leftrightarrow \sum_{i=1}^{n} \gamma(z_{ik}) (\mathbf{x}_i - \mu_k) = 0 \tag{C.12}$$

$$\Leftrightarrow \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i - \mu_k \left(\sum_{i=1}^{n} \gamma(z_{ik}) \right) = 0 \tag{C.13}$$

$$\Leftrightarrow \mu_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i}{\sum_{i=1}^{n} \gamma(z_{ik})} \tag{C.14}$$

For a symmetric matrix $\mathbf{A}$, we have:

$$\frac{\partial}{\partial \mathbf{A}} \log(\det \mathbf{A}^{-1}) = -\mathbf{A}^{-1}$$

In addition, for any matrix $\mathbf{A}$ and vector v,

$$\frac{\partial}{\partial \mathbf{A}} v^T \mathbf{A} v = v v^T$$

Hence,

$$\frac{\partial \mathcal{L}(\mu, \Sigma, \pi)}{\partial \Sigma_k^{-1}} = 0 \Leftrightarrow \sum_{i=1}^{n} \left(\frac{\gamma(z_{ik})}{2} (\Sigma_k^{-1})^{-1} - \frac{\gamma(z_{ik})}{2} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T) \right) = 0 \tag{C.15}$$

$$\Leftrightarrow \Sigma_k \left(\sum_{i=1}^{n} \frac{\gamma(z_{ik})}{2} \right) = \sum_{i=1}^{n} \frac{\gamma(z_{ik})}{2} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T \tag{C.16}$$

$$\Leftrightarrow \Sigma_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik})(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\sum_{i=1}^{n} \gamma(z_{ik})} \tag{C.17}$$

To estimate the partial derivative of π_k, we are subject to the constraint $\sum_k \pi_k = 1$. We introduce the Lagrangian and setting its partial derivatives to zero.

$$L(\pi) = \sum_{i=1}^{n} \sum_{k} \gamma(z_{ik}) \log(\pi_k) + \lambda(1 - \sum_{k} \pi_k) + C_{\pi_k} \tag{C.18}$$

where C_{π_k} is a constant term according to π_k.

$$\frac{\partial L(\pi)}{\partial \pi_k} = 0 \Leftrightarrow \frac{\sum_{i=1}^{n} \gamma(z_{ik})}{\pi_k} - \lambda = 0 \tag{C.19}$$

$$\Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} \gamma(z_{ik})}{\pi_k} \tag{C.20}$$

$$\Leftrightarrow \pi_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik})}{\lambda} \tag{C.21}$$

Knowing that $\sum_k \pi_k = 1$,

$$\frac{\sum_{i=1}^{n} \sum_k \gamma(z_{ik})}{\lambda} = 1 \tag{C.22}$$

and then,

$$\lambda = \sum_{i=1}^{n} \sum_{k} \gamma(z_{ik}) = \sum_{i=1}^{n} 1 = n \tag{C.23}$$

Therefore,

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ik}) \tag{C.24}$$

And the posterior probabilities (or responsibilities) for a individual $i \in \{1, n\}$ to belong to the causal group k are given by:

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)}{\sum_{l=\{R,D,S,A\}} \pi_l \mathcal{N}(\mathbf{x}_i | \mu_l, \Sigma_l)} \tag{C.25}$$

C.2 Hybrid Gaussian and independent Multinomials mixture model

Recall that, the covariates $\mathbf{X}$ can be decomposed in a set of continuous variables $\mathbf{X}_{\mathcal{N}}$ and a set of categorical variables $\mathbf{X}_{\mathcal{M}}$. We note $n_{\mathcal{M}}$ the number of categorical variables and $\mathbf{X}_{\mathcal{M}_j}$ the j^{th} categorical variable with $j = 1, ..M$. We assume the continuous variables follow a normal distribution $\mathcal{N}(\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean and variance parameters. Each categorical variable $\mathbf{X}_{\mathcal{M}_j}$ follows an independent Multinomial distribution $\mathcal{M}(\rho_{jk})$, where n_{jk} is the number of categories ν of the variable $\mathbf{X}_{\mathcal{M}_j}$ and $\rho_{jk} = (p_{jk\nu})_{\nu=\{1,..,n_{jk}\}}$ is the vector given the probabilities of selecting category ν. Note that $\mathbf{x}_{i\mathcal{M}_j,\nu}$ is binary with only one value equal to 1. We put $\rho = ((\rho_{jk})_{j=\{1,..,M\}})_{k=\{R,D,S,A\}}$ the set of the Multinomials parameters. $(\pi_k)_{k=\{R,D,S,A\}}$ are the mixing parameters of the causal distributions where $\sum_{k=\{R,D,S,A\}} \pi_k = 1$.

The joint probability of covariates is given by:

$$p(\mathbf{X}, \mu, \Sigma, \rho, \pi) = \prod_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \pi_k \mathcal{N}(x_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \mathcal{M}(x_{i\mathcal{M}_1} \mid \rho_{1k}) \mathcal{M}(x_{i\mathcal{M}_2} \mid \rho_{2k}) ... \mathcal{M}(x_{i\mathcal{M}_M} \mid \rho_{Mk}) \tag{C.26}$$

$$= \prod_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \pi_k \mathcal{N}(x_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \prod_{j=1}^{M} \mathcal{M}(\mathbf{x}_{i\mathcal{M}_j} \mid \rho_{jk}) \tag{C.27}$$

The complete likelihood takes the form:

$$p(\mathbf{X}, \mathbf{Z} \mid \mu, \Sigma, \rho, \pi) = \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} \left(\pi_k \mathcal{N}(\mathbf{x}_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \prod_{j=1}^{M} \mathcal{M}(\mathbf{x}_{i\mathcal{M}_j} \mid \rho_{jk}) \right)^{z_{ik}} \tag{C.28}$$

$$= \prod_{i=1}^{n} \prod_{k=\{R,D,S,A\}} \left(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma_k}} exp\left(-\frac{1}{2} (\mathbf{x}_{i\mathcal{N}} - \mu_k)^T \Sigma^{-1} (\mathbf{x}_{i\mathcal{N}} - \mu_k) \right) \prod_{j=1}^{M} \prod_{\nu=1}^{n_{jk}} \rho_{jk\nu}^{\mathbf{x}_{i\mathcal{M}_j,\nu}} \right)^{z_{ik}} \tag{C.29}$$

and the log-likelihood is expressed as:

$$\mathcal{L}(\mu, \Sigma, \rho, \pi) = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} [z_{ik} \log(\frac{\pi_k}{(2\pi)^{\frac{d}{2}} \sqrt{det\Sigma_k}}) - z_{ik} \frac{(\mathbf{x}_{i\mathcal{N}} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_{i\mathcal{N}} - \mu_k)}{2} + \sum_{j=1}^{M} \sum_{\nu=1}^{n_{jk}} z_{ik} \mathbf{x}_{i\mathcal{M}_j,\nu} \log(\rho_{jk\nu})] \tag{C.30}$$

$$\mathcal{L}(\mu, \Sigma, \rho, \pi) = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} [z_{ik} \log(\pi_k) - \frac{z_{ik}}{2} \log(det\Sigma_k) - \frac{z_{ik}}{2} (\mathbf{x}_{i\mathcal{N}} - \mu_k)^T \Sigma_k^{-1} (x_{i\mathcal{N}} - \mu_k) + \sum_{j=1}^{M} \sum_{\nu=1}^{n_{jk}} z_{ik} \mathbf{x}_{i\mathcal{M}_j,\nu} \log(\rho_{jk\nu}) + C] \quad \text{(C.31)}$$

where C is a constant term of μ, Σ, P and π.

Maximizing the expected value of the complete log-likelihood:

$$\mathbb{E}[\mathcal{L}(\pi, \mu, \rho, \Sigma)] = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} [\gamma(z_{ik}) \log(\pi_k) - \frac{\gamma(z_{ik})}{2} \log(det\Sigma_k) - \frac{\gamma(z_{ik})}{2} (\mathbf{x}_{i\mathcal{N}} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_{i\mathcal{N}} - \mu_k) + \sum_{j=1}^{M} \sum_{\nu=1}^{n_{jk}} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu} \log(\rho_{jk\nu}) + C] \quad \text{(C.32)}$$

is equivalent to set its partial derivatives to zero.

$$\frac{\partial \mathcal{L}(\mu, \Sigma, \rho, \pi)}{\partial \mu_k} = 0 \Leftrightarrow \mu_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{N}}}{\sum_{i=1}^{n} \gamma(z_{ik})} \quad \text{(C.33)}$$

$$\frac{\partial \mathcal{L}(\mu, \Sigma, \rho, \pi)}{\partial \Sigma_k} = 0 \Leftrightarrow \Sigma_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) (\mathbf{x}_{i\mathcal{N}} - \mu_k)(\mathbf{x}_{i\mathcal{N}} - \mu_k)^T}{\sum_{i=1}^{n} \gamma(z_{ik})} \quad \text{(C.34)}$$

$$\frac{\partial \mathcal{L}(\mu, \Sigma, \rho, \pi)}{\partial \pi_k} = 0 \Leftrightarrow \pi_k = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ik}) \quad \text{(C.35)}$$

See proofs in the previous section.

The parameters of the Multinomials are subject to the constraint $\sum_\nu \rho_{jk\nu} = 1$. The Lagrangian is then introduced in the log-likelihood:

$$L(P) = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} \sum_{j=1}^{M} \sum_{\nu=1}^{n_{jk}} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu} \log(\rho_{jk\nu}) + \lambda(1 - \sum_{\nu=1}^{n_{jk}} \rho_{jk\nu}) + C_{\rho_{jk\nu}} \quad \text{(C.36)}$$

where $C_{\rho_{jk\nu}}$ is a constant term according to $\rho_{jk\nu}$.

$$\frac{\partial L(P)}{\partial \rho_{jk\nu}} = 0 \Leftrightarrow \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu}}{\rho_{jk\nu}} - \lambda = 0 \tag{C.37}$$

$$\Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu}}{\rho_{jk\nu}} \tag{C.38}$$

$$\Leftrightarrow \rho_{jk\nu} = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu}}{\lambda} \tag{C.39}$$

Knowing that $\sum_{\nu} \rho_{jk\nu} = 1$, we have:

$$\lambda = \sum_{\nu=1}^{n_{jk}} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu} = \sum_{i=1}^{n} \gamma(z_{ik}) \left(\sum_{\nu=1}^{n_{jk}} \mathbf{x}_{i\mathcal{M}_j,\nu} \right) = n_{jk} \sum_{i=1}^{n} \gamma(z_{ik}) \tag{C.40}$$

Therefore,

$$\rho_{jk\nu} = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_{i\mathcal{M}_j,\nu}}{n_{jk} \sum_{i=1}^{n} \gamma(z_{ik})} \tag{C.41}$$

Finally, the posterior probabilities for a individual $i \in \{1, n\}$ to belong to the causal group k is given by:

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_{i\mathcal{N}} \mid \mu_k, \Sigma_k) \prod_{j=1}^{M} \prod_{\nu=1}^{n_{jk}} \mathcal{M}(\mathbf{x}_{i\mathcal{M}_j,\nu} \mid \rho_{jk\nu})}{\sum_{l=\{R,D,S,A\}} \pi_l \mathcal{N}(\mathbf{x}_{i\mathcal{N}} | \mu_l, \Sigma_l) \prod_{j=1}^{M} \prod_{\nu=1}^{n_{jl}} \mathcal{M}(\mathbf{x}_{i\mathcal{M}_j,\nu} \mid \rho_{jl\nu})} \tag{C.42}$$

D Resumé de la thèse en français

Cette section a pour objectif de résumer le contenu de la thèse et les résultats obtenus, pour un public francophone.

D.1 Introduction générale à l'inférence causale

Dès notre plus jeune âge, nous apprenons les règles de cause à effet par l'observation de nos expériences répétées. Si un bébé pousse un objet de la table, l'objet tombera sur le sol et se brisera. Bien que nous soyons tous familiers avec ce concept, définir la causalité reste une tâche difficile. La causalité a été définie philosophiquement par Platon comme "tout ce qui devient doit nécessairement devenir grâce à quelque Cause ; car sans cause, il est impossible que quoi que ce soit atteigne le devenir" (28a) [Plato, 1925]. Cependant, la causalité n'est pas seulement une relation entre une cause et un effet, c'est un phénomène d'explication, qui vise principalement à répondre à la question "pourquoi" [Pearl and Mackenzie, 2018].

D.1.1 La causalité, une ressource essentielle à l'apprentissage automatique

La causalité a été directement appliquée dans de nombreux domaines, tels que la santé, l'économie et les sciences sociales [Holland, 1986], mais il est crucial de souligner son impact actuel et croissant dans l'apprentissage automatique. Actuellement, les méthodes d'apprentissage automatique atteignent des performances exceptionnelles, notamment grâce aux progrès des réseaux de neurones. Cependant, à mesure qu'ils deviennent plus complexes, les modèles d'apprentissage automatique deviennent moins faciles à interpréter et à expliquer. En outre, ces modèles s'attaquent par construction à des questions de corrélation et ne permettent pas de répondre aux questions fondamentales du "comment" et du "si". La causalité peut apporter des solutions à ces problèmes et élargir le champ d'application de ces modèles en leur permettant de gagner en capacité de prédiction et d'explication. Dans ce qui suit, les tâches d'apprentissage automatique pour lesquelles la causalité est essentielle sont mises en évidence.

Interprétabilité De plus en plus de modèles d'apprentissage automatique sont considérés comme des boîtes noires avec des décisions inexpliquées pour les humains, ce qui les rend difficiles à croire malgré leurs performances élevées. Afin d'expliquer le processus qui se cache derrière un modèle efficace, l'interprétabilité dans le contexte de la causalité peut répondre à des questions telles que : "Pourquoi ce modèle a-t-il produit ce résultat ?" et "Quelles sont les caractéristiques responsables de ce résultat ?". Il s'agit alors de savoir si les résultats sont conformes à l'éthique humaine et si les données sont biaisées et conduisent à une décision irrationnelle. Outre les algorithmes intrinsèquement interprétables (fournissant une interprétation au moment de la formation) et les interprétations post-hoc (générant des explications pour les décisions prises), les informations causales peuvent expliquer quelles décisions auraient été prises dans une situation alternative. Pearl soutient dans [Pearl, 2018] que le cadre contrefactuel est le moyen crucial d'atteindre le plus haut degré d'interprétabilité. Dans cette optique, les auteurs dans [Harradon et al., 2018, Chattopadhyay et al., 2019] ont montré comment construire un modèle causal graphique pour générer des explications des prédictions d'un réseau neuronal profond. D'autres approches, basées sur l'interprétabilité post-hoc, ont également été proposées [Ribeiro et al., 2016].

Invariance La causalité et l'invariance sont étroitement liées. L'apprentissage d'un prédicteur invariant équivaut à trouver une représentation des données sur laquelle le modèle optimal est le même pour tous les environnements. Pour trouver une représentation invariante, il faut éliminer les *corrélations fallacieuses* (c'est-à-dire les relations entre les variables qui n'ont pas de liens causaux directs et les sources de bruit ou de variabilité exogène). En effet, de telles relations ne sont pas stables à travers les environnements et nuiraient aux performances du modèle. Par conséquent, la détermination de la véritable relation causale entre les variables est une clé pour aborder correctement le problème [Arjovsky et al., 2019]. Notons que, non seulement la connaissance des relations causales permet de trouver les invariances pertinentes, mais réciproquement, l'invariance permet d'inférer les liens de causalité entre les variables [Bühlmann et al., 2020].

Adaptation de domaine L'adaptation de domaine vise à prédire un résultat sur un domaine qui diffère du domaine d'entraînement. Les auteurs dans [Zhang et al., 2013] expliquent comment l'adaptation de domaine peut être vue comme un problème d'inférence causale. La notion d'intervention en causalité peut être reliée à la perturbation d'un système, entraînant un changement dans la distribution initiale, dans le cadre de l'adaptation de domaine. L'information causale peut alors être utilisée pour expliquer les changements dans les domaines de la distribution des données. Plus récemment, les auteurs dans [Magliacane et al., 2017] proposent une approche permettant de prédire la distribution d'une variable à partir d'autres variables observées dans un ou plusieurs domaines cibles. L'idée est de sélectionner un sous-ensemble de caractéristiques qui satisfait l'invariance de la prédiction dans chaque domaine. Le problème d'adaptation de domaine est résolu en utilisant l'inférence causale pour prédire les distributions conditionnelles invariantes, c'est-à-dire en estimant la probabilité conditionnelle du résultat étant donné qu'un sous-ensemble de caractéristiques reste le même dans le domaine d'apprentissage et le domaine cible.

Fairness La fairness (l'équité), qui tente de rétablir l'impartialité des décisions prises de manière biaisée et injuste, en fonction de covariables telles que le sexe ou l'origine ethnique d'un individu, peut également être étudiée dans un cadre de la causalité. Les auteurs dans [Kusner et al., 2017] étudient la fairness contrefactuelle. Le résultat prédit par le modèle dans le monde réel (environnement observé) est comparé au résultat prédit dans un environnement *counterfactuel* non-observé (défini dans la section D.1.3). Si la décision dans le monde réel est la même que celle dans le monde contrefactuel, le modèle est considéré comme juste, sinon injuste. L'objectif de la fairness peut être considéré comme l'intention d'éliminer les effets des *facteurs de confusion* (définis dans la section D.1.3) assimilées à des critères injustes sur la prédiction d'une décision finale.

Apprentissage par renforcement (RL) De plus en plus de travaux en RL étudient la direction des relations causales entre les variables pour expliquer le comportement des agents d'apprentissage, par exemple dans [Madumal et al., 2020, Zhang, 2020]. Le modèle causal structurel est appris pendant l'apprentissage par renforcement pour estimer les relations causales entre les variables d'intérêt et fournir un modèle nettement plus performant.

Plus généralement, la causalité peut être utilisée pour tout problème où il y a un changement dans la distribution des données. Cela peut être le cas dans un système multi-agents, où les règles du jeu changent [Peysakhovich et al., 2019] et dans les attaques adverses pour lesquelles les exemples de test modifiés ne sont pas tirés de la même distribution que les exemples d'entraînement [Goodfellow et al., 2014]. La connaissance de la structure causale améliore la robustesse des modèles prédictifs par rapport aux changements des statistiques d'entrée et explique les décisions et les actions du système.

D.1.2 Les deux grandes approches de la causalité

Depuis les travaux de Jerzy Neyman en 1923, traduits et édités quelques années plus tard dans [Splawa-Neyman et al., 1990], la modélisation de la causalité a fait l'objet de nombreuses études dans les domaines de la statistique et de la théorie des probabilités. Depuis, deux écoles principales sont aujourd'hui dominantes : Judea Pearl avec le développement du modèle causal structurel (SCM) [Pearl, 1995, Pearl et al., 2000, Pearl, 2009] et Donald Rubin, qui a donné son nom au modèle causal de Rubin (RCM), également connu sous le nom de modèle causal de Neyman-Rubin, basé sur une approche contrefactuelle [Rubin, 1974, Rubin, 1975, Rubin, 2003].

Un SCM est défini par des équations structurelles qui se rapportent à des variables endogènes (observées et internes au modèle) et exogènes (variables non observées, externes au modèle et imposées par l'environnement). Un SCM peut être associé à un modèle causal graphique obtenu avec des graphes acycliques dirigés (DAG). Les propriétés graphiques peuvent également être utilisées pour estimer l'effet causal.

Le RCM, détaillé dans la section suivante, se concentre sur la relation causale entre une variable de traitement et un résultat. Afin d'éliminer l'effet des variables externes non prises en compte, le problème est traité comme un problème à données manquantes. Ce modèle est basé sur

l'estimation de ce qui se serait passé dans une situation hypothétique différente, c'est-à-dire avec d'autres occurrences de traitement.

D.1.3 Concepts et définition de l'inférence causale

L'objectif de cette section est de définir l'inférence causale dans un cadre statistique et probabiliste. Nous désignons par *inférence causale* le processus par lequel une relation causale peut être établie entre une cause et ses effets.

Causalité et prédiction

La prédiction et la causalité sont distinctes, mais complémentaires. La prédiction est utilisée pour prévoir et ensuite planifier des événements futurs. Elle répond à des questions telles que : "Que va-t-il se passer ?", "Quels individus seront affectés par une variable" et "Sachant X, est-il probable d'avoir Y ?". La causalité a pour objectif d'expliquer un événement et répond respectivement aux questions suivantes : "Que se passera-t-il sous l'effet de certains changements ?", "Pourquoi des individus seront-ils affectés par une variable ?" et "Si nous changions X, comment cela changerait-il Y ?". La prédiction est définie en termes d'une distribution conjointe de conditions statistiques. Les modèles d'intérêt prédisent une variable Y après l'observation de $X = x$ et estiment ainsi $\mathbb{P}(Y \mid X = x)$. A l'opposé, la définition de la causalité requiert des conditions dynamiques. et estime $\mathbb{P}(Y \mid \text{set } X = x)$.

Definition D.1 (Inférence causale)**.** [Pearl et al., 2016, Section 1.5] *L'inférence causale est le processus permettant de faire une prédiction sur le changement d'un résultat Y sachant le changement d'une entrée X, c'est-à-dire que les modèles prédisent Y après avoir fixé $X = x$. L'objet d'intérêt est $\mathbb{P}(Y \mid set\ X = x)$.*

La relation entre X et Y peut être illustrée par un DAG. Si X cause Y, alors X est appelé la *cause* et Y l' *effet*.

Un effet peut avoir plus d'une cause et aucune d'entre elles ne peut à elle seule l'expliquer dans sa totalité. De plus, la causalité a la propriété d'être:

- Transitive : Si X est une cause de Y et Y une cause de Z, alors X est une cause de Z.
- Irréflexive : X ne peut pas être cause de lui-même.
- Antisymétrique : Si X est une cause de Y, alors Y n'est pas une cause de X.

Estimation de l'effet causal

L'effet d'une variable n'est rarement dû qu'à une seule cause. Pour estimer l'inférence causale d'une variable d'intérêt T sur un résultat Y, nous devons considérer toutes les variables qui affectent Y. Ces variables sont de deux catégories : les covariables et les facteurs de confusion.

Definition D.2 (Covariables)**.** [Rubin, 2003, Sous-section 5] *Les covariables sont les caractéristiques observées d'un individu (ou d'une unité) pouvant influencer le résultat Y et non affectées par la variable d'intérêt T.*

Definition D.3 (Facteur de confusion)**.** [Pearl et al., 2016, Section 3.4] *Un facteur de confusion ou une variable de confusion est une variable non observée (exogène au modèle) qui affecte à la fois la variable d'intérêt T et le résultat Y, et qui explique une partie de l'effet causal de T sur Y.*

Considérons que nous étudions l'effet de l'aspirine sur les maux de tête. Le mal de tête est la variable de résultat. L'âge et le poids de l'individu peuvent être considérés comme des covariables. La prise d'une aspirine n'affecte pas ces caractéristiques, mais ces variables peuvent modifier la réaction du patient. Ces variables peuvent être observées et prises en compte dans l'effet que procure l'aspirine. Cependant, d'autres variables non observées peuvent également influencer cet effet, comme le niveau de fatigue ou le degré d'hydratation de l'individu. Ces variables, qui ne sont pas détectables, sont des variables confondantes.

Pour estimer l'effet causal d'une variable T sur un résultat Y, sans l'effet des variables confondantes, l'idée est de comparer la variation du résultat d'interêt par rapport à une base de référence. Il faut observer les variations de Y pour des valeurs de traitement T données, sachant les covariables X. Comme pour chaque valeur de T l'effet des facteurs de confusion est supposé constant, la comparaison de différentes valeurs de T annule l'influence de ces variables.

L'effet de l'aspirine sur les céphalées est estimé en observant la variation des céphalées avec et sans aspirine pour un âge et un poids donnés de l'individu. Comme le niveau de fatigue ou le degré d'hydratation sont considérés comme constants avec et sans traitement pour un individu donné, l'étude de ces deux conditions réduit l'influence de ces variables.

Une manipulation, appelée également *intervention* doit alors être requise (voir [Pearl et al., 2016, Section 3.1]). L'idée d'action est formalisée par Pearl avec l'opérateur *do-calculus* introduit dans [Pearl, 1995]. La distribution créée par l'intervention $T = t$ est donnée par la probabilité conditionnelle de Y, sachant $X = x$: $\mathbb{P}(Y \mid X = x, do(T = t))$ ou $\mathbb{P}(Y \mid X = x, \text{set } T = t)$. La variable T est alors appelée *traitement*.

Definition D.4 (Traitement)**.** *[Rubin, 2003, Subsection 1] La variable d'interêt, pour laquelle les effets sont étudiés, est appelée* traitement*. L'intervention consiste alors à forcer chacune des valeurs potentielles du traitement, afin d'observer les résultats qui en découlent.*

L'intervention est à distinguer d'une simple observation. Alors que $\mathbb{P}(Y \mid X = x, T = t)$ est une probabilité observée, $\mathbb{P}(Y \mid X = x, do(T = t))$ est estimée par l'intervention consistant à forcer la variable T à la valeur t. En outre, il faut noter que le fait de forcer la variable T à la valeur donnée t suppose qu'il existe une valeur de traitement alternative t' à laquelle l'individu aurait pu être exposé au même moment.

Prendre ou non une aspirine sont les différentes valeurs de traitement possible. Un individu décide de prendre une aspirine, mais il peut décider en même temps de ne pas la prendre (ou de prendre un autre médicament ou produit alternatif). Il y a donc une idée d'intervention.

D.2 L'approche contrefactuelle

D.2.1 Résultat potentiel, factuel et contrefactuel

L'approche contrefactuelle est basée sur le *résultat potentiel*, défini par la combinaison du résultat sur toutes les valeurs potentielles de traitement.

Dans le RCM, l'effet d'un nouveau médicament est estimé à partir du résultat avec et sans traitement. Si le traitement est administré à un patient, la question est de savoir "ce qui se serait passé sans traitement". Et si le traitement n'est pas administré, la question est "que se serait-il passé avec le traitement ?".

Pour formaliser le problème, considérons plusieurs instances observées $i \in \{1, .., n\}$, avec $n \in \mathbb{N}$. Ce sont des individus (aussi appelés unités ou exemples) de la base de données. Un individu $i \in \{1, .., n\}$ peut être une personne, un lieu ou une chose unique à un moment donné. Ainsi, si le même individu est observé à deux moments différents, il sera considéré comme deux individus différents. Soit $\mathcal{X} \subset \mathbb{R}^d$ un espace continu ou discret de dimension d et $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ un échantillon i.i.d., où chaque $\mathbf{x}_i \in \mathcal{X}$ sont les caractéristiques continues ou discrètes. Nous notons $y_{i,obs} \in \mathcal{Y}$ le résultat observé et t_i l'affectation du traitement de l'individu $i \in \{1, .., n\}$, et respectivement $\mathbf{y}_{obs}$ et $\mathbf{t}$ leur ensemble pour tous les individus. La variable de traitement peut être binaire, discrète ou continue.

Prescrire ou ne pas prescrire un traitement médical donné est un cas de traitement binaire. Choisir entre plusieurs médicaments est un exemple de traitement discret multiple. Lorsqu'un médecin attribue un dosage spécifique d'un médicament, il s'agit d'un cas continu.

Lorsqu'il existe L traitements potentiels à attribuer, pour tout $i \in \{1, .., n\}$, $t_i \in \{t^0, t^1, .., t^{L-1}\}$. Notons que, dans la plupart des cas, pour des raisons de commodité, l'hypothèse d'un traitement binaire est généralement faite dans de nombreux travaux. Pour un traitement binaire, où un individu $i \in \{1, .., n\}$ est exposé ou non à un traitement, les unités traitées ($t_i = 1$) constituent le groupe *traitement*, et les individus non traités ($t_i = 0$) le groupe *contrôle*.

Pour tous les individus $i \in \{1, .., n\}$ et les traitements $l \in \{0, .., L-1\}$, $y_{i,l}$ est défini comme le résultat de l'individu i correspondant au traitement l attribué. En cas de traitement binaire, on note $y_{i,0}$ et $y_{i,1}$ les résultats réels et hypothétiques, relatifs aux traitements potentiels $\{0, 1\}$.

Definition D.5 (Résultats réels, contrefactuels et potentiels)**.** [Imbens and Rubin, 2015, Section 1.3] *Nous appelons un* résultat factuel*, noté $\mathbf{y}_{i,obs}$ pour un individu i dans$\{1, .., n\}$, le résultat observé associé au traitement assigné. Les* résultats contrefactuels *sont les résultats non observés $(y_{i,0}, y_{i,1}, .., y_{i,L}) \setminus (y_{i,obs})$ associés aux traitements non attribués. Le* résultat potentiel*, noté $(y_{i,0}, y_{i,1}, .., y_{i,L})$, est l'ensemble des résultats factuels et contrefactuels.*

Pour un traitement binaire, le résultat potentiel est le couple $(y_{i,0}, y_{i,1})$. L'un des composants est observé $y_{i,obs} = t_i\, y_{i,1} + (1 - t_i)\, y_{i,0}$, tandis que l'autre est manquant $y_{i,miss} = (1 - t_i)\, y_{i,1} + t_i\, y_{i,0}$. Les résultats observés et hypothétiques sont illustrés dans le tableau D.1. Pour un individu $i \in \{1, .., n\}$,

si le traitement est attribué ($t_i = 1$), alors $y_{i,1}$ est observé et $y_{i,0}$ est manquant, sinon $y_{i,0}$ est observé et $y_{i,1}$ est manquant.

Données observées			
X	**t**	$\mathbf{y}_0$	$\mathbf{y}_1$
$\mathbf{x}_1$	1	.	$y_{1,1}$
$\mathbf{x}_2$	1	.	$y_{2,1}$
$\mathbf{x}_3$	0	$y_{3,0}$	.
$\mathbf{x}_4$	1	.	$y_{4,1}$
...	...	...	...
$\mathbf{x}_n$	0	$y_{n,0}$	.

Table D.1: Les données observées et manquantes dans le cas d'un traitement binaire. Pour un individu donné $i \in \{1, .., n\}$, $\mathbf{x}_i$ correspond aux covariables, $\mathbf{t}$ à l'affectation du traitement et $(\mathbf{y}_0, \mathbf{y}_1)$ aux résultats potentiels. Les valeurs non observées sont désignées par ".".

Dans l'exemple de l'évaluation de l'efficacité d'un traitement médical, un individu correspond à un patient atteint d'une pathologie spécifique. Pour un individu donné $i \in \{1, .., n\}$, le traitement est l'action de prendre ($t_i = 1$) ou non ($t_i = 0$) le médicament donné. Le résultat est la réaction du patient, c'est-à-dire l'évolution de la pathologie. Si le patient prend le traitement, la réaction avec le traitement $y_{i,1}$ est observée (et nommée le résultat factuel), et sa réaction hypothétique sans traitement $y_{i,0}$ est manquante et doit être estimée (c'est le résultat contrefactuel). Dans le cas d'un traitement multiple, l'effet de deux traitements médicaux différents peut être comparé. Il faut comparer la réaction du patient avec le premier traitement ($t_i = 1$), le deuxième traitement ($t_i = 2$) et sans traitement ($t_i = 0$). Si le premier traitement est administré au patient, alors $y_{i,1}$ est le résultat factuel, $(y_{i,0}, y_{i,2})$ sont les résultats contrefactuels, et $(y_{i,0}, y_{i,1}, y_{i,2})$ est le résultat potentiel.

L'inférence causale peut directement être déduite de la connaissance des résultats contrefactuels. Cependant, le problème fondamental est que seul le *résultat factuel* $\mathbf{y}_{obs}$ est observé alors que les *résultats contrefactuels* $(\mathbf{y}_0, \mathbf{y}_1, .., \mathbf{y}_L) \setminus (\mathbf{y}_{obs})$ sont inconnus. Ainsi, la problématique consiste à estimer le résultat contrefactuel.

D.2.2 L'hypothèse d'ignorabilité forte

Afin de permettre l'estimation des résultats contrefactuels et de l'effet causal, certaines hypothèses sont nécessaires. L'approche contrefactuelle est basée sur l'hypothèse de la valeur de traitement unitaire stable (SUTVA), introduite dans [Rubin, 1978, Rosenbaum and Rubin, 1983]. Il se compose de deux éléments :

- *Non-interférence* : Il n'y a pas d'interférence entre les individus. Le résultat potentiel d'un individu n'est pas affecté par l'affectation du traitement aux autres individus (notons que cette hypothèse peut être violée lorsque l'équilibre général est perturbé).

Si nous considérons un couple vivant dans la même maison, nous supposons que si l'un des partenaires change sa façon de cuisiner en réponse au traitement, il n'y aura pas d'impact sur la réponse de l'autre partenaire au traitement.

- *Pas de variation dans le traitement* : Les traitements de tous les individus sont comparables. Par exemple, dans un traitement binaire, il n'y a qu'un seul état de traitement et de contrôle.

Une variation de traitement peut se produire lorsque plusieurs doses d'un médicament sont administrées alors que l'on ne considère que le fait de prendre ou non le traitement.

Pour obtenir un estimateur sans biais de l'inférence causale, deux autres hypothèses doivent être faites : [Rosenbaum and Rubin, 1983, Imbens and Rubin, 2015] :

- L'hypothèse "*Unconfoundedness" (ou d'ignorabilité)* consiste en l'indépendance entre le résultat potentiel et l'affectation du traitement, conditionnellement aux covariables: $(\mathbf{y}_0, \mathbf{y}_1, .., \mathbf{y}_L) \perp\!\!\!\perp \mathbf{t}|\mathbf{X}$. Cela signifie que toutes les variables qui affectent conjointement les résultats potentiels et le traitement sont observées (au sein des covariables).

Si deux individus ont exactement les mêmes caractéristiques, leurs résultats potentiels seront les mêmes, quel que soit le traitement attribué.

Ainsi, les résultats factuels et les résulats contrefactuels jouent un rôle similaire dans la modélisation. Tous les résultats qui composent le résultat potentiel seront considérés de la même manière et, par conséquent, le résultat observé n'est pas distingué des résultats contrefactuels. Notons qu'en pratique, il n'est pas possible de confirmer cette hypothèse car les résultats contrefactuels sont inobservables. Les données disponibles ne permettent pas de déduire une indépendance conditionnelle des résultats contrefactuels.

- *L'hypothèse "Overlap" (ou de chevauchement)* fournit une probabilité non nulle de recevoir le traitement, c'est-à-dire $0 < \mathbb{P}(T = 1 \mid X = \mathbf{x}) < 1$.

Chaque patient a la possibilité de recevoir le traitement, par exemple sans contraintes financières.

L'hypothèse d'ignorabilité forte est constituée des hypothèses SUTVA, unconfoundedness et overloap.

D.3 Les challenges de la thèse

Dans cette thèse, nous nous plaçons dans le cadre contrefactuel (RCM), dans lequel l'inférence causale est vue comme un problème de données manquantes. Nous nous focalisons sur la relation causale qui peut exister entre deux variables, une variable de traitement et un résultat observé, conditionnellement à des co-variables. Nous nous limitons au cas de traitement binaire, ce qui est une hypothèse standard prise en raison de la complexité du problème. L'objectif est d'estimer l'effet du traitement, déterminé par la différence entre les résultats des groupes de traitement et de contrôle. Comme l'un des résultats est observé et que l'autre est manquant, une manière intuitive

consiste à diviser les données en groupes selon l'attribution effective du traitement et à créer des estimateurs pour chacun des groupes. La problématique réside dans le fait que, dans la plupart des cas, ces deux groupes ne sont pas homogènes et donc les résultats de sont pas comparables.

D.3.1 Le défi du biais de sélection et des données observationelles

Les données utilisées peuvent être collectés à partir de deux types d'études : *expérimentales* et *observationnelles*. Les études expérimentales ont l'avantage d'être contrôlées par des humains. Il s'agit d'une approche active, qui met en place un cadre expérimental pour contrôler l'attribution du traitement. Dans des *essais contrôlés randomisés* (RCT), lorsque l'échantillon comprend un nombre important de données et que le traitement est attribué de manière aléatoire, deux groupes homogènes de traitement et de contrôle peuvent être construits et comparés afin d'estimer l'effet du traitement. Dans ce contexte, la corrélation entre les données est significative et de "vraies" conclusions peuvent être tirées. Cependant, ce type d'études sont coûteuses, longues à réaliser et souvent difficiles à mettre en oeuvre pour de multiples raisons telles que l'étique. En pratique, une *politique d' attribution* est établie, permettant de décider l'état de traitement auquel un individu doit être exposé. Les individus recevant le traitement sont sélectionnés en fonction d'un apriori sur leur réponse au traitement en fonction de leurs caractéristiques, ce qui entraîne un *biais de sélection* dans les données. Le défi consiste donc à estimer les *effets de traitement hétérogène*.

Dans la recherche épidémiologique, pour estimer l'effet d'un médicament sur un patient, on procède à des essais cliniques expérimentaux. Dans certains cas, le médicament peut être affecté selon une politique basée sur les caractéristiques du patient, comme le sexe, l'âge ou les antécédents médicaux.

Les second types d'études sont nomées études observationnelles. Il s'agit d'une approche passive dans laquelle les données sont extraites d'observations passées. Il existe une demande croissante pour utiliser ces données en raison de leur abondance et de leur facilité de collecte. Cependant, elles sont plus délicates à utiliser en raison du cadre inconnu dans lequel elles ont été générées. L'attribution du traitement n'étant pas connue, la difficulté est d'estimer l'effet du traitement malgré le biais de sélection potentiel. De plus, un fort biais dû aux facteurs confondants peut exister, produisant des paradoxes ou conduisant à de fausses conclusions.

D.3.2 De l'effet moyen à l'effet individuel du traitement

Pendant de nombreuses années, l'effet du traitement a été étudié par l'estimation de l'effet moyen du traitement (ATE) [Morgan and Winship, 2015], défini comme la différence d'espérance des résultats potentiels :

$$\tau_0 := \mathbb{E}[Y_1 - Y_0] \tag{D.1}$$

Sous l'hypothèse SUTVA, elle peut être exprimée en faisant la moyenne sur la distribution de X :

$$\tau_0 = \mathbb{E}[\mathbb{E}[Y \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0]] \tag{D.2}$$
$$= \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \tag{D.3}$$
$$= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \tag{D.4}$$

L'objectif est de réduire le biais de sélection en créant des groupes ou sous-groupes homogènes, et d'estimer l'effet du traitement en moyenne dans chaque groupe, avec par exemple des méthodes d'appariement, de stratification ou de pondération (voir l'Appendice B).

Avec l'intérêt croissant pour la personnalisation, comme dans le cas de la médecine personnalisée, du marketing personnalisé et de la politique personnalisée, il est nécessaire d'estimer l'effet individuel du traitement (ITE) c'est-à-dire les effets causaux au niveau de l'unité d'un individu. L'ITE est défini comme la différence entre le résultat avec et le résultat sans traitement d'un individu $i \in \{1, .., n\}$ donné [Morgan and Winship, 2015, Section 2.2]:

$$ITE := y_{i,1} - y_{i,0} \tag{D.5}$$

Remarque : Alors que le terme ITE semble être préféré dans le contexte médical, le concept d'*uplift* est plus courant au sein de la communauté marketing. D'autres termes sont également utilisés, tels que *vrai lifting* [Lo, 2002], *modélisation de la valeur incrémentale*, *association* [Wasserman, 2013, chap. 19], ou *effets de traitement spécifiques* [Shpitser and Pearl, 2006].

Notons que l'estimation de l'ITE et de l'ATE sont deux approches différentes. L'ITE est l'effet d'un traitement pour chaque individu, tandis que l'ATE est l'effet moyen du traitement sur l'ensemble de la population.

En recherche épidémiologique, lorsque l'efficacité d'un traitement médical est évalué, des tests A/B sont généralement réalisés. Ils se situent dans le cadre de l'estimation de ATE, car ils étudient l'effet moyen sur l'ensemble de la population. Le résultat d'un groupe de patients est observé avec (A) et sans (B) traitement. Si, en moyenne, l'efficacité atteint un seuil donné, le médicament peut être mis sur le marché (ou passer à la phase suivante de l'essai clinique). Si nous nous concentrons maintenant au niveau de l'individu, l'accent n'est pas mis sur l'efficacité du traitement en moyenne, mais plutôt sur l'efficacité du traitement pour un individu donné. Comparé à l'estimation de l'ATE, il faut cibler les patients qui ont un ITE acceptable. Ainsi, un traitement pourrait être écarté en raison de son efficacité moyenne, mais retenu par rapport à son efficacité sur l'individu considéré.

L'ITE n'est pas directement calculable car pour chaque individu de l'ensemble des données d'entraînement, seul le résultat du traitement attribué est observé, l'autre est inconnu. Une façon de surmonter cette difficulté est de s'appuyer sur l'effet du traitement moyen conditionnel (CATE), défini comme la différence espérée entre les deux résultats potentiels (espérance de l'ITE). [Morgan and Winship, 2015, Section 2.7.1]:

$$\tau(\mathbf{x}) := \mathbb{E}[Y_1 - Y_0 | X = \mathbf{x}] \tag{D.6}$$

Avec l'indépendance conditionnelle de Y sur X :

$$\tau(\mathbf{x}) = \mathbb{E}[Y_1 \mid X = x] - \mathbb{E}[Y_0 \mid X = \mathbf{x}] \tag{D.7}$$

De plus, l'hypothèse de cohérence implique que :

$$\tau(\mathbf{x}) = \mathbb{E}[Y_1 \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y_0 \mid X = \mathbf{x}, T = 0] \tag{D.8}$$

Ainsi, sous l'hypothèse de forte ignorabilité :

$$\tau(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0] \tag{D.9}$$

Il peut également être exprimé avec le *do*-calcul de la notation de Pearl :

$$\tau(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 1)] - \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 0)] \tag{D.10}$$

Les auteurs dans [Künzel et al., 2019] démontrent que le meilleur estimateur pour le CATE est également le meilleur estimateur pour l'ITE en termes d'erreur quadratique moyenne (MSE). L'hypothèse d'ignorabilité forte permet d'identifier l'effet causal du traitement sans l'effet des facteurs confondants cachés et d'assurer la consistance de l'estimateur. C'est une condition satisfaisante pour l'identifiabilité de l'ATE, bien que des versions plus faibles de cette hypothèse soient suffisantes : [Imbens and Wooldridge, 2009, Pearl, 2017, Shalit et al., 2017]. Lorsqu'une politique d'attribution de traitement est adoptée, $\mathbb{E}[Y \mid T = 1] \neq \mathbb{E}[Y_1]$ et le CATE est une estimation non biaisée qui ne peut être obtenue en comparant directement les groupes de traitement. Cette question fait actuellement l'objet d'un nombre considérable de travaux et cette thèse en fait partie.

D.3.3 Une perspective axée sur l'identification des populations causales

Dans le cadre d'un traitement et d'un résultat binaire ($t_i \in \{0, 1\}$ et $(y_{i,0}, y_{i,1}) \in \{0, 1\}^2$), une classification des individus selon la combinaison des deux résultats potentiels, produit quatre *populations causales* [Wasserman, 2013, Section 19.1]:

- *Répondants* (R) réagissent comme on le souhaite uniquement lorsqu'ils sont traités : $y_{i,0} = 0$ et $y_{i,1} = 1$. Le traitement leur est bénéfique.
- *Condamné* (D) ne présentent jamais la réaction souhaitée : $y_{i,0} = 0$ et $y_{i,1} = 0$. Le traitement est sans effet sur eux.
- *Survivant* (S), présentent toujours la réaction souhaitée : $y_{i,0} = 1$ et $y_{i,1} = 1$. Le traitement est sans effet sur eux.
- *Anti-répondant* (A), ne présentent la réaction souhaitée que lorsqu'ils ne sont pas traités : $y_{i,0} = 1$ et $y_{i,1} = 0$. Le traitement est nocif pour eux.

Afin d'identifier l'ITE, nous proposons dans cette thèse une perspective basée sur la structure des populations causales et nous abordons la problématique suivante :

Comment estimer les distributions de probabilité des populations causales ?

Le résultat contrefactuel est ensuite prédit en supposant que les individus appartiennent au groupe causal dont la probabilité est la plus élevée. De plus, nous démontrons que cette approche permet également d'estimer l'ITE. En outre, contrairement aux travaux connexes qui estiment directement l'ITE, en abordant un problème plus large, cette approche fournit des informations supplémentaires sur l'inférence causale existante entre le résultat et le traitement. Les informations sur la distribution des populations causales permettent de classifier les individus dans chaque groupe et de prédire le traitement optimal à attribuer à chaque individu. En plus d'une simple prédiction des résultats contrefactuels, cette modélisation offre une politique d'attribution du traitement pour de nouveaux individus non traités. Les individus pour lesquels le traitement sera bénéfique, ceux pour lesquels le traitement n'a aucun effet et ceux pour lesquels il est nocif, peuvent ainsi être détectés. Ainsi, nous pouvons cibler les individus dans chacun de ces groupes.

Pour estimer la distribution de probabilité des groupes causaux, nous proposons deux approches : une approche paramétrique basée sur l'algorithme d'Espérance-Maximisation et une approche non-paramétrique utilisant une architecture d'Auto-Encoder.

D.4 Approche paramétrique utilisant un algorithme d'espérance-maximisation itératif

Cette partie s'appuie sur une publication au 28e symposium européen sur les réseaux neuronaux artificiels (ESANN 2020) "Estimating Individual Treatment Effects through Causal Populations Identification" [Beji et al., 2020]. Le matériel et le code sont disponibles sur github à l'adresse suivante : https://github.com/CelineBeji.

D.4.1 Modèle paramétrique des populations causales

La population entière est modélisée par un mélange de répondants (R), condamnés (D), survivants (S) et d'anti-répondants (A). Pour chaque population $k \in \{R, D, S, A\}$, posons $f_k(. \mid \theta_k)$ les distributions antérieures et π_k la probabilité de mélange. Notre objectif est d'estimer l'ensemble des paramètres $\theta = \{\theta_k\}_{k\in\{R,D,S,A\}}$ et $\pi = \{\pi_k\}_{k\in\{R,D,S,A\}}$, en fixant la distribution des covariables, tel que :

$$p(\mathbf{X} \mid \pi, \theta) = \sum_{k=\{R,D,S,A\}} \pi_k \, f_k(\mathbf{X} \mid \theta_k) \tag{D.11}$$

Afin d'appliquer des contraintes causales, nous introduisons un espace latent (ou caché) $\mathcal{Z}$ qui représente l'espace des populations causales. Nous supposons qu'il existe une fonction $g : \mathcal{X} \to \mathcal{Z}$ qui transforme les covariables $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ en variables latentes $\mathbf{Z} = \{z_{ik}\}_{k\in\{R,D,S,A\}}\}_{i=1}^n$. Nous définissons $\gamma(z_{ik}) := \mathbb{P}(Z = z_{ik} \mid X = \mathbf{x}_i)$ la probabilité conditionnelle de $(Z = z_{ik})$ étant donné $(X = \mathbf{x}_i)$, où $X \sim P_X$ et $Z \sim P_Z$ sont des variables aléatoires. L'expression $\gamma(z_{ik})$ représente la probabilité qu'un individu $i \in \{1, .., n\}$ appartienne à la population causale $k \in \{R, D, S, A\}$.

Par conséquent, la log-vraisemblance complète peut être exprimée comme :

$$\log p(\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{z}_1, \cdots, \mathbf{z}_n \mid \pi, \theta) = \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} z_{ik} \log f_k(\mathbf{x}_i \mid \theta_k) + \sum_{i=1}^{n} \sum_{k=\{R,D,S,A\}} z_{ik} \log \pi_k \tag{D.12}$$

Le modèle proposé implique l'application de plusieurs contraintes sur l'espace latent, excluant deux populations causales selon le résultat factuel observé et le traitement affecté. A partir de l'affectation du traitement $t_i \in \{0, 1\}$ et du résultat factuel $y_{i,obs} \in \{0, 1\}$, une information partielle sur l'espace latent $\{\gamma(z_{ik})\}_{k\in\{R,D,S,A\}}$ peut-être déduite. Par exemple, si un individu $i \in \{1, .., n\}$ n'est pas traité ($t_i = 0$) et que le résultat observé est égal à zéro ($y_{i,obs} = 0$), cet individu ne peut pas être par définition un survivant ou un anti-répondant. La distribution de probabilité de ces deux populations peut être forcée à zéro, c'est-à-dire $\gamma(z_{iS}) = \gamma(z_{iA}) = 0$, et les deux autres peuvent être normalisées, c'est-à-dire $\gamma(z_{iR}) + \gamma(z_{iD}) = 1$. Des contraintes similaires existent pour chaque valeur de l'affectation du traitement et des résultats factuels. Ces contraintes sont le fondement de notre approche. Elles sont résumées dans le tableau D.2. Dans le même ordre d'idées, des contraintes initiales peuvent être ajoutées en initialisant les probabilités des deux populations non exclues à $\frac{1}{2}$.

$(t_i = 0\ ,\ y_{i,obs} = 0)$	$(t_i = 0\ ,\ y_{i,obs} = 1)$	$(t_i = 1\ ,\ y_{i,obs} = 0)$	$(t_i = 1\ ,\ y_{i,obs} = 1)$
$\begin{cases}\gamma(z_{iS}) = \gamma(z_{iA}) = 0 \\ \gamma(z_{iR}) + \gamma(z_{iD}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iR}) = \gamma(z_{iD}) = 0 \\ \gamma(z_{iS}) + \gamma(z_{iA}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iR}) = \gamma(z_{iS}) = 0 \\ \gamma(z_{iD}) + \gamma(z_{iA}) = 1\end{cases}$	$\begin{cases}\gamma(z_{iD}) = \gamma(z_{iA}) = 0 \\ \gamma(z_{iR}) + \gamma(z_{iS}) = 1\end{cases}$

Table D.2: Contraintes de causalité(C^*)

Une fois le problème d'apprentissage résolu, l'ITE peut être estimé avec la distribution de probabilité des groupes causaux :

$$\tau(\mathbf{x}_i) = (\gamma(z_{iR}) + \gamma(z_{iS}))\mathbb{E}[\mathbb{1}_{y_{i,1}=1}] - (\gamma(z_{iS}) + \gamma(z_{iA}))\mathbb{E}[\mathbb{1}_{y_{i,0}=1}] \tag{D.13}$$

avec par définition $\gamma(z_{ik}) = \frac{\pi_k f_k(\mathbf{x}_i|\theta_k)}{\sum_{l=\{R,D,S,A\}} \pi_l f_k(\mathbf{x}_i|\theta_l)}$ pour $k \in \{R, D, S, A\}$.

D.4.2 Algorithme d'Espérance-Causalité-Maximisation (ECM)

Nous proposons un algorithme inspiré de l'algorithme EM avec des informations partielles supplémentaires sur les deux groupes causaux impossibles. L'algorithme est enrichi de contraintes causales C^* déduites de la structure causale du problème. La distribution latente est non seulement estimée à partir de la distribution des covariables $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ mais aussi à partir de l'affectation du traitement $\mathbf{t} = \{\mathbf{t}_i\}_{i=1}^n$ et du résultat observé $\mathbf{y}_{obs} = \{\mathbf{y}_{i,obs}\}_{i=1}^n$.

L'algorithme proposé 5 fonctionne en trois étapes itératives : Espérance-Causalité-Maximisation (ECM). Pour chaque individu $i \in \{1, .., n\}$, l'étape E estime les probabilités d'appartenance à chaque groupe causal $\gamma(z_{ik})\}_{k=\{R,D,S,A}$ selon les paramètres des quatre distributions $(\theta_k, \pi_k)_{k=\{R,D,S,A\}}$.

L'étape C impose les contraintes causales C^* (voir Tableau 3.3) sur l'espace latent en forçant à zéro les probabilités $\gamma(z_{ik})$ des deux groupes impossibles et en normalisant les deux probabilités restantes. Cette étape peut être considérée comme un post-traitement de l'étape E car c'est un ajustement des probabilités estimées $\gamma(z_{ik})$. L'étape M maximise la log-vraisemblance complète $\mathcal{L}(\theta, \pi) = \log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta, \pi)$ afin de mettre à jour les paramètres des distributions. Pour une convergence plus rapide, les valeurs initiales des variables latentes $\gamma(z_{ik})^0$ sont initialisées avec les contraintes initiales à $\frac{1}{2}$ sur chacune des deux populations causales possibles.

Algorithm 5: Algorithme d'Espérance-Causalité-Maximisation (ECM)

Initialisation: Initialisation de $\gamma(z_k)^0$ avec les contraintes causales initiales C_0^* et calcul de π_0 et θ_0 d'après l'étape M.

Tant qu'il n'y a pas de convergence: (itérer sur m)

Étape d'Espérance (E) : Evaluation de $q(\mathbf{Z} \mid \mathbf{X}, \theta^{m-1}, \pi^{m-1})$.

Étape de Causalité (C) : Application des contraintes de causalité C^* sur les probabilités a posteriori des variables latentes $\gamma(z_k)$ en fonction du Tableau D.2.

Étape de Maximisation (M) : $(\theta^m, \pi^m) = \arg\max_{\theta,\pi}(\mathbb{E}[\log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta^{m-1}, \pi^{m-1})])$.

Vérification de la convergence de la log-vraisemblance complète

$\log p(\mathbf{X}, \mathbf{t}, \mathbf{y}_{obs}, \mathbf{Z} \mid \theta^m, \pi^m)$.

Fin de la boucle

D.4.3 Résultats et conclusion

L'algorithme Espérance-Causalité-Maximisation (ECM) est illustré sur un mélange gaussien, pour lequel des expérimentations sur des jeux de données synthétiques et réels sont réalisées, afin de vérifier son efficacité par rapport à l'état de l'art. Les résultats sur les jeux de données synthétiques montrent que plus la distribution des données est complexe, plus l'ECM surpasse les modèles de référence. Notons que les jeux de données synthétiques générés sont particulièrement favorables au modèle ECM puisque qu'ils sont distribués en mélange de gaussiennes de moyenne et variance variable. Sur le jeu de données semi-synthétique utilisé (IHDP), l'écart de performance entre l'ECM et les autres modèles est significatif par rapport à la métrique estimant l'erreur quadratique moyenne entre le vrai résultat contrefactuel et son estimation, mais pas en utilisant la métrique estimant le rang des individus selon leur ITE. La distribution des données explique la difficulté de prédire efficacement la distribution des groupes causaux. En effet, cet ensemble de données a la particularité de contenir un très grand nombre de données catégorielles.

Ainsi, nous cherchons à savoir si l'ECM peut être adpaté aux données catégorielles, plus courantes dans des jeux de données réelles. Pour cela, nous utilisons une distribution mixant des gaussiennes et des multinomiales indépendantes, comme apriori de distribution. Ce nouvel algorithme,

nommé ECM hybride est alors comparé au premier que nous qualifions d'ECM simple. Sur des données contenant uniquement des variables continues, l'ECM simple et l'ECM hybride ont des résultats équivalents. Cela s'explique par la définition des deux distributions qui sont similaires en l'absence de données catégorielles. Sur un jeu de données catégoriel pur, l'ECM hybride a de meilleures performances. L'écart se creuse sur les jeux de données pour lesquels nous mélangeons des variables continues et catégorielles. Les expérimentations montrent qu'une distribution hybride est mieux adaptée à un jeu de données contenant à la fois des données continues mais également des données catégorielles.

L'un des avantages de cet algorithme est son extension à n'importe quelle distribution. Il peut être adapté à toute distribution pour laquelle la log-vraisemblance a une forme fermée. Plus la distribution a priori est proche de la distribution réelle des données, meilleures sont les performances de l'ECM. Cependant, il a des limites lorsque la vraisemblance n'est pas calculable. Une solution consiste à utiliser une adaptation variationnelle de l'algorithme et une autre consiste à étudier une approche non paramétrique. C'est cette deuxième option que nous développons dans la section suivante.

D.5 Approche non paramétrique utilisant un Auto-Encoder

Cette partie est basé sur une publication au workshop Causal-ITALY de la 20e conférence internationale de l'Association italienne pour l'intelligence artificielle (AIxIA 2021). Le matériel et le code sont disponibles sur github https://github.com/CelineBeji.

Nous proposons une architecture nommée Auto-Encoder Causal (CAE) dont le but est d'estimer la distribution de probabilité des populations causales, en reconstruisant les covariables. En plus de la structure standard, des contraintes causales sont appliquées sur l'espace latent. Ces contraintes, qui dépendent du résultat observé et du traitement attribué, visent à capturer les distributions de probabilité de chaque population causale.

D.5.1 Architecture globale

L'architecture est composée de trois parties : un encodeur, un masque construit en accord avec les contraintes causales de façon à ce que certaines unités de l'espace latent soient ramenées à zéro, et un décodeur (voir la Figure D.1 pour une vue d'ensemble). L'encodeur prend en entrée les covariables $\mathbf{x}_i$, correspondant au i^{ime} échantillon des données, pour $i \in \{1, .., n\}$. Chaque nœud de la couche initiale représente une caractéristique (une dimension de $\mathbf{x}_i$). Après avoir comprimé $\mathbf{x}_i$ grâce à un réseau neuronal feedforward, la variable latente $\mathbf{z}_i = \begin{pmatrix} z_{iR} & z_{iD} & z_{iS} & z_{iA} \end{pmatrix}$ est contrainte par le masque $M(y_{i,obs}, t_i)$. Chaque nœud z_{ik} de la couche latente est assimilé à une population causale $k \in \{R, D, S, A\}$. z_{ik} est activé ou non (mis à zéro) en fonction du masque établi à partir du résultat observé $y_{i,obs}$ et du traitement assigné t_i. Par exemple, si $t_i = 0$ et $y_{i,obs} = 0$, l'individu est par définition un survivant ou un anti-répondant. Le masque conçu prend la forme $M(y_{i,obs}, t_i)^T = \begin{pmatrix} 0 & 0 & 1 & 1 \end{pmatrix}$ où la première coordonnée (respectivement la deuxième, la troisième et la quatrième) correspond à la probabilité d'être un répondeur (respectivement un

condamné, un survivant et un anti-répondant). La variable latente contrainte $\tilde{\mathbf{z}}_i$ est prise comme entrée du décodeur. Le décodeur, qui repose sur un réseau neuronal feedforward, a pour but de reconstruire les covariables. Il produit une estimation de $\hat{\mathbf{x}}_i$, dont la de taille est identique à celle des covariables d'entrée $\mathbf{x}_i$. Le CAE est entraîné de manière itérative à partir d'une base de données d'apprentissage $\{\mathbf{x}_i\}_{i=1}^{s}$ (où s est le nombre d'échantillons utilisés pour l'apprentissage), par rétropropagation, en minimisant la fonction de perte $l_{CAE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$.

D.5.2 Prédiction

Une fois le modèle entraîné, le bloc encodeur du modèle est utilisé pour encoder la probabilité des populations causales z_{ik} (correspondant à $\gamma(z_{ik})$ dans les notations du section D.4). Chaque individu est affecté à la population causale pour laquelle la probabilité prédite est la plus élevée. Plus la probabilité d'appartenir à une population causale est élevée, plus la certitude de l'estimation est élevée. Connaissant la population causale prédite pour un individu donné, le traitement qui lui a été affecté et son résultat face au traitement, le résultat contrefactuel peut directement être estimé. Comme précédemment pour l'ECM, l'ITE (défini comme la différence entre le résultat attendu sur le groupe de traitement moins le résultat attendu sur le groupe de contrôle étant donné les covariables) peut également être estimé à partir de la distribution de probabilité de la population causale.

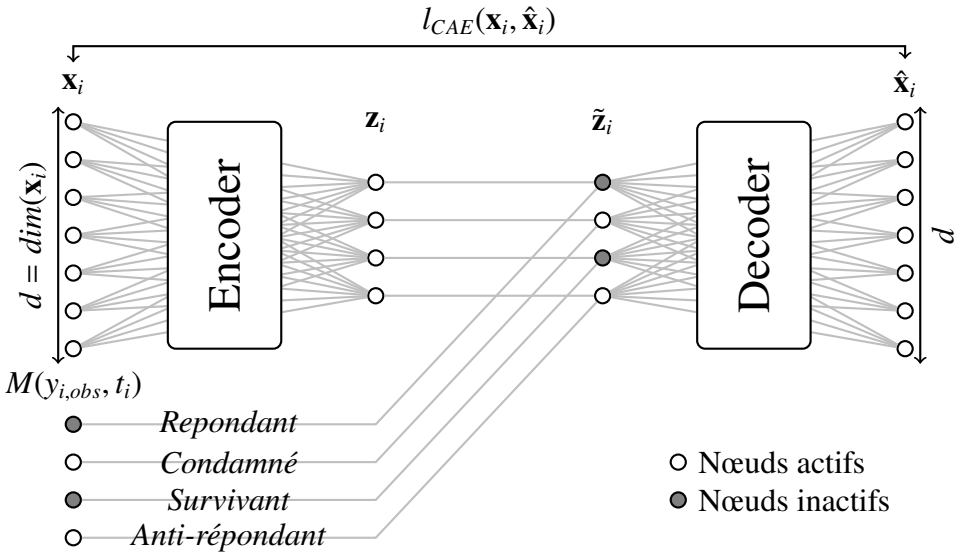


Figure D.1: Vue d'ensemble de l'architecture du Causal-Auto-Encoder (CAE). Les nœuds blancs (respectivement les nœuds gris) correspondent aux neurones actifs (respectivement aux neurones inactifs). $\mathbf{x}_i$ sont les covariables et $\hat{\mathbf{x}}_i$ sont leur reconstruction par l'Auto-Encoder. $\mathbf{z}_i$ est la variable latente construite par l'encodeur. Elle est ensuite contrainte par le masque $M(y_{i,obs}, t_i)$ (certains de ses neurones sont désactivés) à partir du résultat observé $y_{i,obs}$ et de l'attribution du traitement t_i.

D.5.3 Architecture de la couche latente

Pour coder les quatre populations causales, nous devons avoir au moins quatre nœuds, un pour chaque population. Cependant, plus qu'un seul nœud par population peut être considéré. Intuitivement, plus il y a de nœuds, plus l'information sur les covariables est préservée. La taille du masque dépend du nombre de nœuds l alloués pour encoder une population. Il contraint donc tous les nœuds associés à une population causale. La probabilité d'appartenir à une population donnée est calculée comme la somme des probabilités données par les nœuds affectés à cette population.

En plus des nœuds représentant les probabilités des populations causales, des nœuds non-contraints peuvent être ajoutés. Ces nœuds participent uniquement à l'encodage de l'information provenant des covariables, sans être perturbés par l'apriori causal. Leur nombre peut être choisi aussi grand que nécessaire, en fonction du nombre de caractéristiques. Cependant, le nombre de nœuds supplémentaires doit être inférieur à la taille des covariables, afin que l'information soit distribuée à la fois sur les nœuds supplémentaires et sur les nœuds de probabilité de distribution. Les valeurs des nœuds supplémentaires ne sont pas affectées par le masque et ne sont pas utilisées dans l'estimation des distributions de probabilité des populations causales.

D.5.4 Résultats et conclusion

La taille de l'espace latent est discutée, et l'efficacité de notre approche est démontrée dans des expérimentations sur des ensembles de données synthétiques et réelles. Le modèle proposé est efficace sur des données réelles et présente l'avantage significatif d'être non-paramétrique et applicable sur des ensembles de données à grande échelle. Comparé à l'ECM, il présente l'avantage significatif de ne nécessiter aucune hypothèse sur la distribution a priori des données. Cependant, il présente l'inconvénient de manquer de théorie pour justifier ses performances. De plus, les expérimentations sur des jeux de données synthétiques mettent en évidence les limites de notre modèle sur des distributions de données de faible complexité. Pour surmonter le problème de l'overfitting, trois architectures peuvent être considérées : les Auto-Encodeurs sparse, de débruitage et contractifs. D'autre part, contrairement à l'ECM, les modèles proposés dans cette section n'estiment pas la distribution des populations causales mais seulement la probabilité d'appartenance à chacune de ces populations. En conséquence, nous perdons des informations sur la distribution des individus et des comportements moyens. Cette information n'est pas directement obtenue avec l'approche CAE, cependant nous pouvons obtenir les caractéristiques d'un individu prototypique en utilisant le bloc décodeur entraîné de notre modèle. Si les vecteurs $\begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}$, $\begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix}$, $\begin{pmatrix}0 & 0 & 1 & 0\end{pmatrix}$ et $\begin{pmatrix}0 & 0 & 0 & 1\end{pmatrix}$ sont donnés en entrée du décodeur, en sortie on obtient respectivement les caractéristiques reconstruites d'un répondant, d'un condamné, d'un survivant et d'un anti-répondant.

D.6 Perspectives

Dans cette thèse, nous avons proposé une solution au défi que représente l'estimation de l'effet causal sur des données observationnelles, pour lesquelles la procédure d'attribution du traitement est inconnue. Nous espérons que ce travail sera utile à la communauté scientifique et qu'il contribuera aux avancées dans le domaine de l'inférence causale, avec une nouvelle perspective sur le problème de l'estimation des résultats contrefactuels. La principale force de cette approche est qu'elle peut être adaptée aux multi-traitements, qui seront les futurs enjeux dans ce domaine. De plus, les modèles proposés n'excluent pas la prise en compte de la non-conformité. Ces deux points sont donc succinctement développés dans ce qui suit.

D.6.1 Extension en multi-traitements

Bien que prise en considération par la plupart des méthodes de référence, l'hypothèse du traitement binaire reste forte. Dans le domaine médical par exemple, la question principale n'est pas de savoir s'il faut ou non donner un médicament, mais quel médicament choisir parmi ceux qui existent.

Un avantage majeur de l'approche proposée est son évolutivité vers des traitements multiples. Ceci peut être accompli en redéfinissant les groupes causaux. Considérons L le nombre de traitements potentiels. La variable de traitement peut prendre des valeurs dans $\{t^0, t^1, .., t^{L-1}\}$. Les populations causales 2^L sont donc définies par la valeur du résultat sur tous les groupes de traitement $(\mathbf{y}_0, \mathbf{y}_1, .., \mathbf{y}_{L-1})$ présentés dans le tableau D.3.

	$T = t^0$	$T = t^1$	$T = t^2$	...	$T = t^{L-2}$	$T = t^{L-1}$
Group 1	$y_{i,0} = 0$	$y_{i,1} = 0$	$y_{i,2} = 0$	...	$y_{i,L-2} = 0$	$y_{i,L-1} = 0$
Group 2	$y_{i,0} = 1$	$y_{i,1} = 0$	$y_{i,2} = 0$	...	$y_{i,L-2} = 0$	$y_{i,L-1} = 0$
Group 3	$y_{i,0} = 0$	$y_{i,1} = 1$	$y_{i,2} = 0$	...	$y_{i,L-2} = 0$	$y_{i,L-1} = 0$
Group 4	$y_{i,0} = 1$	$y_{i,1} = 1$	$y_{i,2} = 0$	...	$y_{i,L-2} = 0$	$y_{i,L-1} = 0$
...	...	...	...	...	...	...
Group 2^L	$y_{i,0} = 1$	$y_{i,1} = 1$	$y_{i,2} = 1$	...	$y_{i,L-2} = 1$	$y_{i,L-1} = 1$

Table D.3: Populations causales en multi-traitements

L'algorithme ECM et le CAE peuvent être utilisés de la même manière que la méthode décrite précédemment, avec un mélange de 2^L distributions distinctes.

D.6.2 Question ouverte de non-conformité

Tout au long de cette thèse, nous avons supposé que le traitement attribué est le traitement donné. Cependant, dans la pratique, la non-conformité est relativement fréquente. Par exemple, un individu peut accidentellement recevoir le mauvais traitement ou choisir de ne pas prendre le traitement qui lui a été attribué par désintérêt ou par crainte des effets secondaires potentiels.

Ainsi, le traitement pris D n'est pas toujours le même que le traitement assigné T. Dans le cas d'un traitement binaire et d'une conformité binaire (deux options de conformité : prendre ou ne pas prendre le traitement), quatre groupes peuvent être identifiés. Un individu assigné au traitement peut prendre ou ne pas prendre le traitement et un individu non assigné au traitement peut prendre ou ne pas prendre le traitement (voir le Tableau D.4). Ce problème est souvent appelé dans la littérature une *analyse de l'intention de traiter*.

	$t_i = 0$	$t_i = 1$
Compliant (c)	$d_i = 0$	$d_i = 1$
Jamais de prise (n)	$d_i = 0$	$d_i = 0$
Toujours un preneur (a)	$d_i = 1$	$d_i = 1$
Défieur (d)	$d_i = 1$	$d_i = 0$

Table D.4: Comportement de conformité

Les auteurs [Imbens and Rubin, 1997] présentent une méthode bayésienne d'estimation causale en présence de non-conformité où le traitement assigné et le traitement reçu sont tous deux observés. La distribution postérieure est estimée à l'aide d'algorithmes EM et d'augmentation des données. Nous proposons d'appliquer à notre modèle une méthode de prétraitement qui utilise ces travaux. L'idée est de modéliser un mélange de mélanges. Nous n'aurions pas 4 populations mais plutôt $4^2 = 16$, chacune définie par les résultats pour chaque traitement pris et pour chaque affectation de traitement.

If every cause is the consequence of an effect
and all causes are traced back to the original cause,
how could we not believe in the existence of God?

RÉSUMÉ

Dans un cadre contrefactuel, cette thèse formalise l'inférence causale comme un problème d'estimation de densité. L'objectif est d'estimer la distribution de probabilité d'un mélange de quatre populations distinctes, définies par les résultats avec et sans traitement. Le problème fondamental est que les deux résultats ne sont pas observables simultanément. Deux modèles, introduisant des contraintes de causalité à partir de l'information partielle des résultats observés, sont proposés. La première approche, paramétrique, est basée sur un algorithme d'Espérance-Maximisation. Les paramètres des distributions des populations causales sont estimés itérativement en maximisant la vraisemblance, tout en ajoutant un apriori sur les probabilités a posteriori dans une étape intermédiaire. La seconde approche non-paramétrique utilise une architecture d'Auto-Encodeur améliorée par un apriori. Ce dernier se présente sous la forme d'un masque dans la couche intermédiaire du réseau. Des expérimentations sont menées sur des ensembles de données synthétiques et réelles pour prouver l'efficacité de ces approches. Quelques extensions sont également proposées.

MOTS CLÉS

Inférence causale, Résultat contrefactuel, Effet de traitement individuel, Uplift, Contrainte de causalité, Espace latent, Espérance-Maximisation, Auto-Encodeur.

ABSTRACT

In a counterfactual framework, this thesis formalizes the causal inference as a density estimation problem. The aim is to estimate the probability distribution of a mixture of four separate populations, defined by the outcomes with and without treatment. The fundamental problem is that the two outcomes are not simultaneously observable. Two models, leveraging causal constraints built from the partial information of the observed outcomes, are proposed. The first parametric approach is based on an Expectation-Maximization algorithm. The parameters of the causal populations distributions are iteratively estimated by maximizing the likelihood, while adding a prior on the posterior probabilities in an intermediate step. The second non-parametric approach uses an Auto-Encoder architecture enhanced by a prior. The prior is materialized as a mask which is introduced in the intermediate layer of the network. Experiments are conduced on synthetic and real-life datasets to prove the efficiency of these approaches and some extensions are proposed.

KEYWORDS

Causal inference, Counterfactual outcome, Individual Treatment Effect, Uplift, Causal constraint, Latent space, Expectation-Maximization, Auto-Encoder.